



ARTICLE

Genetics and Genomics

Colon-specific eQTL analysis to inform on functional SNPs

Victor Moreno^{1,2,3,4}, M. Henar Alonso^{1,2,3}, Adrià Closa^{1,2,3}, Xavier Vallés^{1,2}, Anna Diez-Villanueva^{1,2}, Laura Valle^{2,5,6}, Sergi Castellví-Bel^{7,8,9}, Rebeca Sanz-Pamplona^{1,2,3}, Adriana Lopez-Doriga^{1,2,3}, David Cordero^{1,2,3} and Xavier Solé^{1,2,3}

BACKGROUND: Genome-wide association studies on colorectal cancer have identified more than 60 susceptibility loci, but for most of them there is no clear knowledge of functionality or the underlying gene responsible for the risk modification. Expression quantitative trait loci (eQTL) may provide functional information for such single nucleotide polymorphisms (SNPs).

METHODS: We have performed detailed eQTL analysis specific for colon tissue on a series of 97 colon tumours, their paired adjacent normal mucosa and 47 colon mucosa samples donated by healthy individuals. R package *MatrixEQTL* was used to search for genome-wide *cis*-eQTL and *trans*-eQTL fitting linear models adjusted for age, gender and tissue type to rank transformed expression data.

RESULTS: The *cis*-eQTL analyses has revealed 29,073 SNP-gene associations with permutation-adjusted *P*-values < 0.01. These correspond to 363 unique genes. The *trans*-eQTL analysis identified 10,665 significant SNP-gene associations, most of them in the same chromosome, further than 1 Mb of the gene. We provide a web tool to search for specific SNPs or genes. The tool calculates Pearson or Spearman correlation, and allows to select tissue type for analysis. Data and plots can be exported.

CONCLUSIONS: This resource should be useful to prioritise SNPs for further functional studies and to identify relevant genes behind identified loci.

British Journal of Cancer (2018) 119:971–977; <https://doi.org/10.1038/s41416-018-0018-9>

INTRODUCTION

Genome-wide association studies (GWAS) of common complex diseases have escalated during the past decade, and multiple susceptibility loci have been identified. For colorectal cancer (CRC), more than 60 single nucleotide polymorphisms (SNPs) have reached genome-wide significant results, but only few of them have functional studies that identify the gene involved.¹ Few SNPs associated to diseases are protein altering non-synonymous variants.² Most disease-associated variants fall into non-coding regions. Some are located in gene deserts far from putative functional genes,³ and only 5% of the currently validated disease associations are restricted to coding variants.⁴ Regarding CRC, only 3 (5%) of the GWAS SNPs known so far map to coding regions (Supplementary Table 1). Some SNPs are located in gene introns, but that does not guarantee that the real functional gene is the one where the SNP is located, as it is known that some enhancers are located in introns of nearby genes.⁵

There is a general agreement that the gene transcript abundance may act as an intermediate phenotype useful to assess the effect of genetic variation on a clinical phenotype.^{6–9} From this observation, rose the concept that disease-associated SNPs located in regulatory regions of close (*cis*) or distant (*trans*) genes act as quantitative trait loci through the modification of

gene expression levels (expression quantitative trait loci (eQTL)). Regulatory sequences of a given gene (promoter, enhancer, splice site), modified by genetic variation, may result in diverse RNA abundance and ultimately, diverse phenotype.

The emergence of microarray technologies, and next-generation sequencing more recently, has allowed the genome-wide exploration of both genetic variation and gene expression. Thus, agnostic search for eQTL is feasible and may be useful to explain SNP-disease associations, to provide prior weights for the statistical analysis of new SNPs discovery^{10–12} and to prioritise which SNPs should be followed-up with additional functional studies.

Data on eQTL-based association studies on CRC are still scarce. Most eQTL analyses have used lymphoblastoid cell lines.^{13–15} Although germ-line genetic variation is assumed constant for all cells, gene expression is tissue specific. The analysis of eQTL may vary across tissues, either because some genes may not be expressed in a specific tissue, thus being undetectable, or because other (epigenetic) regulatory mechanisms of gene expression may interact with the effect of genetic variation. The sharing of eQTL among tissues is of interest.¹⁶ The Genotype Tissue Expression (GTEx) project¹⁷ has been designed to address this issue, and preliminary analyses showed a high degree of sharing,¹⁸ though

¹Unit of Biomarkers and Susceptibility, Cancer Prevention and Control Program, Catalan Institute of Oncology (ICO), Barcelona 08908, Spain; ²Institut d'Investigacions Biomèdiques de Bellvitge (IDIBELL), Barcelona 08908, Spain; ³Centro de Investigación Biomédica en Red de Epidemiología y Salud Pública (CIBERESP), Madrid 28029, Spain; ⁴Department of Clinical Sciences, Faculty of Medicine, University of Barcelona, Barcelona 08907, Spain; ⁵Hereditary Cancer Program, Catalan Institute of Oncology (ICO), Barcelona 08908, Spain; ⁶Centro de Investigación Biomédica en Red de Oncología (CIBERONC), Madrid 28029, Spain; ⁷Department of Gastroenterology, Hospital Clínic de Barcelona, Barcelona 08036, Spain; ⁸Centro de Investigación Biomédica en Red de Enfermedades Hepáticas y Digestivas (CIBEREHD), Madrid 28029, Spain and ⁹Institut d'Investigacions Biomèdiques August Pi i Sunyer (IDIBAPS), Barcelona 08036, Spain
Correspondence: Victor Moreno (v.moreno@iconcologia.net)

Received: 24 May 2017 Revised: 17 September 2017 Accepted: 29 September 2017
Published online: 4 October 2018

colon tissue was not available at the time of that publication. Some studies have analysed tumour tissue from The Cancer Genome Atlas project,¹⁹ but gene expression is strongly altered in the tumour tissue compared to normal tissue.²⁰ Other than GTEx data, to the best of our knowledge, there are no easy sources for eQTL analysis specific to colon tissue, though some researchers have published their analyses.^{21,22}

In this article, we present an eQTL analysis specific for colon tissue, and a web tool that allows the user to query specific SNPs or genes for analysis, with options to combine data from healthy patients, adjacent normal mucosa from patients with colon cancer and tumour tissue. This source of data has been useful to identify genes associated with CRC susceptibility SNPs,²³ and may also be of interest for studies of genetic determinants of inflammatory bowel disease (IBD).²⁴

MATERIALS AND METHODS

Subjects and samples

Colon tumour and paired adjacent normal mucosa tissue samples were selected from a series of cases with a new diagnosis of colon adenocarcinoma attending the University Hospital of Bellvitge in Barcelona between January 1996 and December 2000. Patients included were diagnosed of stage II, microsatellite stable colon cancer, were surgically treated and had not received adjuvant chemotherapy. Adjacent mucosa was obtained from the proximal surgical margins and was at least 10 cm distant from the tumour lesion. Healthy colon mucosa samples were obtained during colonoscopy between February and May 2010. These samples were donated by a series of unselected individuals who underwent a colonoscopy indicated by screening or suspicion of colonic pathology but no colonic lesions were observed. Biopsies were obtained from the left and right colon. For this study, we randomly selected approximately half from each tumour location. All subjects provided written informed consent to participate in the study and the ethics committee of the hospital cleared the protocol with reference PR074/11. Additional information about the study can be found at <https://www.colonomics.org/eqtl-browser>. The eQTL analysis was focused on expression data assessed in normal mucosa. Though we initially selected 100 patients and 50 healthy controls, the final sample size after quality control of the data was ($N = 144$): 97 adjacent normal mucosae from cancer patients and 47 from healthy donors. Gene expression in tumours ($n = 97$) was also analysed, and these data can be compared or combined with those of normal mucosa.

Gene expression and SNP analyses

DNA was extracted from colon mucosa specimens using a standard phenol–chloroform protocol. Total RNA was isolated from tissue samples using the miRCURY™ RNA Isolation Kit (Exiqon, Vedbæk, Denmark) according to manufacturer's protocol, quantified by NanoDrop® ND-1000 Spectrophotometer (Nanodrop technologies, Wilmington, DE) and stored at -80°C . The quality of these RNA samples was assessed with the RNA 6000 Nano Assay (Agilent Technologies, Santa Clara, CA). RNA integrity numbers showed good quality (mean = 8.1 for tumours, 7.5 for adjacent normal and 8.2 for healthy normal). RNA purity was measured with the ratio of absorbance at 260 nm and 280 nm (mean = 1.96, SD = 0.04), with no differences among tissue types.

Expression data were obtained with Affymetrix Human Genome U219 ArrayPlate platform (Affymetrix, Santa Clara, CA). Three 96-array plates were used with a block experimental design to avoid batch effects. Four arrays (two normal–tumour pairs) were excluded due to low quality. Therefore, a final data set of 246 arrays was used for subsequent analyses. Raw data were normalised using the Robust Multiarray Average algorithm implemented in the *affy* package of R/Bioconductor. Expression levels of a set of

genes on this microarray have been validated with quantitative PCR and showed excellent correlation coefficients.²⁵ Prior to the analysis of eQTL, expression probe sets were mapped to genes. For genes with more than one probe set in the array, a principal component analysis was used to capture the largest common variability extracting the first component. The expression array provided data on 20,070 genes, but those with very low variability (SD < 0.1 among all samples) and those on chromosome Y and mitochondrial were excluded for the analysis of eQTL, rendering 15,298.

Genotypes were obtained hybridising genomic DNA extracted from colonic mucosa in Affymetrix Genome-Wide Human SNP 6.0 array (Affymetrix), which includes nearly 1 million SNP markers. One cancer patient and three healthy subjects had to be excluded because the array quality was not good enough. Thus, the final sample size for eQTL analyses were 47 healthy colon mucosae and 97 paired tumour and adjacent normal tissues. Genotype calling was performed for samples of healthy mucosa and normal tissues with the Corrected Robust Linear Model with Maximum Likelihood Classification algorithm as implemented in R/Bioconductor package *crmm*. In addition, untested genotypes were imputed using IMPUTE2²⁶ after haplotyping with SHAPEIT.²⁷ The 1000 Genomes panel (March 2012 version) was used as reference (<http://www.1000genomes.org>). SNPs and *indels* with low imputation quality (info < 0.2 or minor allele frequency [MAF] concordance < 0.9) were excluded from the data set. Also, SNPs with MAF < 0.05 were ignored, and the eQTL analysis will be based on 6.76 million SNPs. No filters for redundant SNPs related to linkage disequilibrium were applied.

The gene expression data set is available at the project website: <https://www.colonomics.org/data> and at Gene Expression Omnibus with GEO series accession number GSE44076. SNP data have been deposited at the European Genome-phenome Archive (EGA, <http://www.ebi.ac.uk/ega/>), which is hosted by the EBI, under accession number EGAS00001002453.

Statistical analysis

To reduce the number of tests performed, while maintaining high power to identify eQTL, only the additive genetic model was considered. Genotypes were coded as the number of variant alleles (0, 1, 2) and this variable treated as quantitative. For imputed genotypes, the posterior probabilities (dosage) were used to consider imputation uncertainty. Dosage was calculated as twice the posterior probability of BB genotype plus that of AB. The additive model is known to capture most of the dominant and recessive effects.²⁸

Analysis of eQTL were performed with the R package *MatrixEQTL*.²⁹ SNPs within 1 Mb of the gene transcription starting site were considered *cis*-eQTL. SNPs located further than 1 Mb or in other chromosomes were considered *trans*-eQTL. The search was limited to associations with P -values < $1e-6$. Linear models were calculated, adjusted for age, sex, tumour location and group (healthy/affected). To avoid spurious results derived from extreme expression values in some genes, a non-parametric analysis was performed, using a rank transformation of the expression values for each gene. This is equivalent to the Spearman correlation, but in a regression framework that allows adjusting for potential confounders.

To account for multiple comparisons, false discovery rates (FDR) were calculated for all findings,³⁰ taking into account all tests performed ignoring redundancy related to linkage disequilibrium (15,000 genes \times 5000 SNPs within 1 Mb per gene). The FDR for the least significant association ($P = 1e-6$) was 0.001. Also, for a more stringent analysis, a permutation procedure was used. The distribution of the most significant associations for the *cis*-eQTL analysis for 500 randomly permuted analyses were used to calculate the adjusted 0.01 significance level. This were nominal P -values of $2e-10$ for *cis*-eQTL and 4e

–13 for *trans*-eQTL. These significance levels are very similar to those based on Bonferroni correction ($1.3e-10$ and $1.1e-13$, respectively), indicating an overall low degree of correlation in the data.

RESULTS

Table 1 shows the distribution of the number of genes and SNPs per chromosome selected for the descriptive analysis of eQTL.

cis-eQTL

The distribution of SNPs near genes was fairly uniform. On average, 2255 SNPs with MAF > 0.05 were found within 500 Mb (SD = 978) and 4555 (SD = 1815) within 1 Mb. A total of 77,139 *cis*-eQTL associations were identified with $P < 1e-6$ (FDR < 0.001). Using the more stringent significance level based on permutations ($P < 2e-10$), 29,073 *cis*-eQTL were significant with adjusted P -value < 0.01. We provide in a supplementary file all identified eQTL, but will restrict further analyses to those significant according to the permuted P -value.

A total of 363 genes (2.4%) had at least one significant *cis*-eQTL within 1 Mb. We will refer to these as 'eGenes' (Supplementary Table 2). The median number of significant *cis*-eQTL per gene was 36 (range: 1–1,023). Approximately 27% of the significant *cis*-eQTL per gene were located intragenic, 44% when only the most significant eQTL per gene were considered. The distribution of eQTL was symmetric upstream and downstream around the gene transcription start site (TSS), irrespective of the gene orientation (Fig. 1). The median distance of the most significant eQTL to the

gene TSS was 14Kb (median absolute deviation = 72Kb). There was no association between the average gene expression and the significance of the eQTL or the proportion of variance explained. Regarding SNPs, 95% of the 29,067 significant eQTL were associated with one gene, 2% were associated with two genes and the other 3% ranged 3–6 genes.

To further analyse the relevance of the identified eQTL database, we analysed the impact of using tumour samples in the analysis of eQTL. When tumours were combined with the normal samples, the number of significant *cis*-eQTL increased to 37,099 significant gene–SNP pairs, 22,759 (78%) common to the analysis of normal mucosa only, but there were 14,337 new eQTL. Since part of this increment could be due to the power gain related to a larger sample size, we also compared the number of significant eQTL identified in tumour only ($n = 6063$), 45% less than those identified in the adjacent normal mucosa of patients ($n = 13,411$). Only 4858 of the eQTL were identified both in tumours and adjacent normal mucosa. A similar number intersected with our list identified in the combination of normal tissue from healthy donors and adjacent to tumour (18%). Figure 2 shows a Venn diagram with these numbers.

Validation

The significant eQTL identified were compared to those reported by the GTEx consortium for transverse tissue. GTEx (version 6) had reported a total of 592,069 eQTL in their analysis. Of those, 149,773 could be considered significant at $P < 2e-10$. We could find 12,076 (42%) of our eQTL in GTEx significant results using the same significance level. A 6500-additional list of our significant

Table 1. Distribution of SNPs, genes and eQTL across chromosomes

CHR	Genotyped SNPs	1000G Imputed SNPs	SNPs MAF > 0.05	Genes	<i>cis</i> -eQTL SNPs ^a	eGenes ^b	<i>cis</i> -eQTL / gene	<i>trans</i> -eQTL SNPs ^c
1	63,788	2,355,554	500,136	1548	2226	34	65.5	
2	66,223	2,583,715	541,685	1019	2110	30	70.3	71
3	54,141	2,168,103	466,623	850	529	16	33.1	
4	49,864	2,168,950	483,706	606	573	14	40.9	57
5	50,522	1,990,693	420,406	683	1244	16	77.8	
6	50,532	1,926,303	450,052	814	3344	26	128.6	2791
7	42,042	1,753,429	387,420	737	3025	26	116.3	1
8	43,775	1,715,508	358,977	532	1270	12	105.8	
9	36,939	1,298,383	281,170	603	610	14	43.6	
10	43,414	1,486,816	335,916	582	763	18	42.4	
11	39,718	1,485,686	325,463	899	1085	19	57.1	
12	38,043	1,440,247	316,735	781	1414	18	78.6	66
13	30,842	1,086,129	244,462	268	460	8	57.5	
14	25,227	989,387	216,774	501	890	11	80.9	
15	23,462	883,939	187,482	489	576	5	115.2	
16	24,942	942,939	198,649	684	347	12	28.9	
17	18,379	818,215	177,606	924	2957	22	134.4	7474
18	23,775	855,576	187,532	211	296	4	74.0	
19	10,429	652,249	155,600	1034	1418	22	64.5	
20	20,617	668,157	145,784	417	923	7	131.9	
21	11,323	409,662	95,007	157	317	4	79.3	
22	10,169	394,297	91,478	379	1454	15	96.9	204
X	32,089	958,335	194,730	580	1236	10	123.6	
All	810,255	31,032,272	6,763,393	15,298	29,067	363	80.1	10,664

MAF minor allele frequency, SNPs single nucleotide polymorphisms, TSS transcription start site ^a *cis*-eQTL within 1 Mb of the TSS with nominal $p < 2e-10$ (adjusted $p < 0.01$) ^b eGenes: unique genes with significant eQTL ^c *trans*-eQTL with nominal $p < 4e-13$ (adjusted $p < 0.01$)

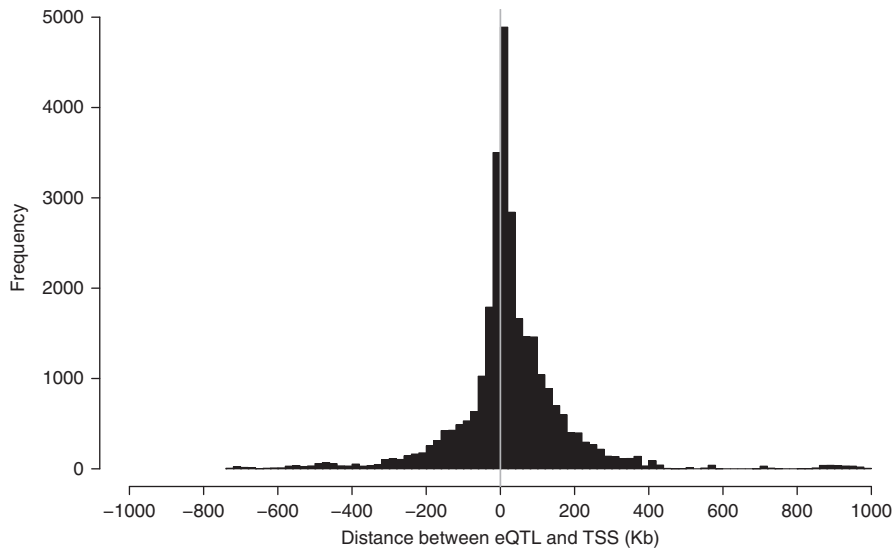


Fig. 1 The histogram shown the distribution of the distance between the significant *cis*-eQTL and the gene TSS in Kb. Most of the eQTL are within 200 Kb of the gene TSS

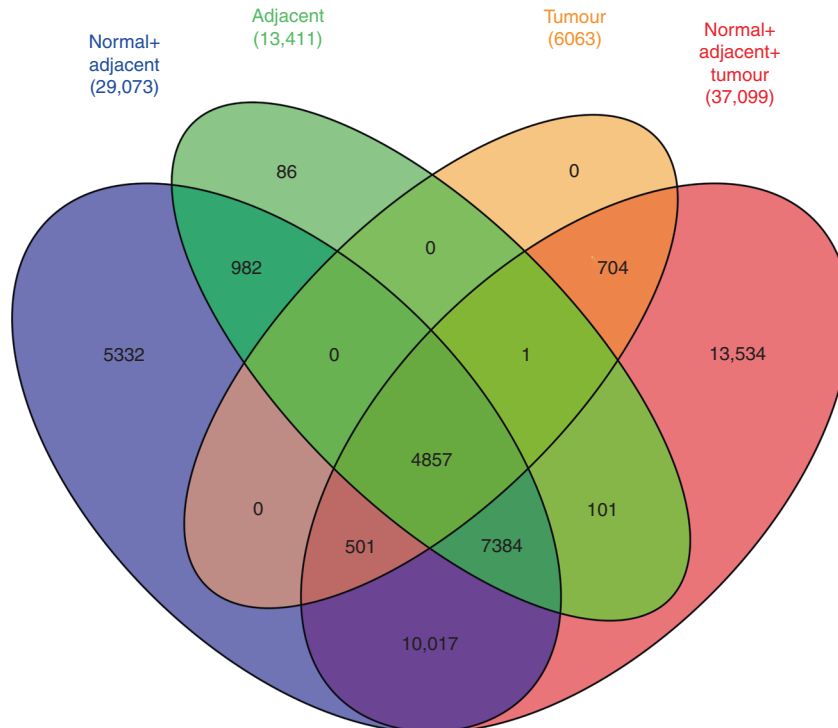


Fig. 2 Venn diagram, of *cis*-eQTL intersections according to the tissues used in the analysis

eQTL were in the GTEx file with lower significance level, increasing the percentage of validated eQTL to 64%. The analysis of sigma samples was similar and the combination slightly increased the number of validated eQTL. The concordance of genes with at least one eQTL was higher, 263 genes in common out of 362 (73%). We also compared our eQTL with that published by Ongen et al. These authors report 1693 genes with the most significant SNP for each gene. Our equivalent list included 363 genes, and 210 of them (58%) were in the list by Ongen et al., but only 15 were perfect match with gene and SNP. However, 157 of Ongen's eQTL were in our extended list containing all significant SNPs for each gene, not only the most significant one.

Mapping to other functional chromatin elements

We explored whether eQTL mapped to regions of the genome with marks of functional activity. We used the Haploreg database³¹ to search eQTL and could retrieve chromatin states from 24,697 (85%) of them. To interpret the information, we also searched 100,000 random SNPs within 1 Mb of TSS and retrieved data from 90,558 of them. In general, eQTL were more often located in chromatin marks (89% vs 84%), DNase (30% vs 25%), promoter histone marks (17% vs 11%) and enhancer histone marks (48% vs 44%). We also observed that eQTL were more often GWAS SNPs (0.5%) than expected (0.2%).

trans-eQTL

Long-distance effects were frequently identified in our analysis. These have not been implemented in the eQTL-browser, but the files can be downloaded from the site. As each SNP is tested against all genes, a more stringent P -value should be used to consider an association significant. We report here significant associations after permutation correction ($P < 4e-13$), but the supplementary file includes all the associations with nominal $P < 1e-6$ (FDR < 0.001).

There were 209,144 *trans*-eQTL identified at nominal $P < 1e-6$, but only 10,665 statistically significant when the permutation-derived significance level ($P < 4e-13$) was applied. Interestingly, they were distributed predominantly in chromosomes 6 and 17, with a minor frequency in chromosomes 2, 4 and 12 (Table 1). These *trans*-eQTL were essentially among SNPs in the same chromosome, but at a larger distance than 1 Mb, with three exceptions: one block of 66 SNPs in chromosome 12 that are eQTL for pseudogene RPS26P11 in chromosome X. The SNPs are near RPS26, the active ribosomal protein 26, located in chromosome 12, but less expressed than the pseudogene in the colon. The second block of 51 SNPs in chromosome 22 that are eQTL for gene GGTL1 (gamma-glutamyltransferase light chain 1) in chromosome 20. The same SNPs are significant *cis*-eQTL for GGT1, GGT2, GGT3P and GGTL2, genes of the gamma-glutamyltransferase family, located in chromosome 22. Finally, there was a bloc of SNPs mapping to HLA-A in chromosome 6, that were eQTL of troponin T3 (TNNT3), a gene located in chromosome 11 related to muscle contraction.

eQTL browser

A web tool was designed to explore gene expression in colon mucosa and analyse eQTL. The R package *shiny* (<https://shiny.rstudio.com>) was used to develop the application, which can be accessed at <https://www.colonomics.org/data-browser>. Some screenshots are available as Supplementary Figs. 1 and 2. The eQTL browser allows searching either one gene by its gene symbol to explore nearby SNPs as candidate eQTL, or one specific SNP, either by rsID or chromosome/position to explore whether its genotypes are associated to the expression of nearby genes. In both types of searches, the output includes location plots and tables with the statistical analyses. The application by default selects samples both from healthy mucosa and adjacent normal tissue, but the used can also exclude some of these or include tumours. Also, samples can be selected according to sex and tumour location (left or right colon). The initial search includes SNPs within 100 Kb upstream and downstream of the selected gene, but the window can be modified up to ± 2 Mb. SNPs can be pruned by allele frequency (MAF > 0.01 by default).

If tumours are also included in the analysis, the pairing is ignored. Thus, the P -values for analyses that include both tumours and adjacent normal tissue should be interpreted with caution. By default, partial Pearson correlation coefficients are calculated, adjusted for age, sex, tumour location and group (healthy/affected). Non-parametric partial Spearman correlation can also be used when expression values for some genes show gross asymmetric distribution (expression plots are provided).

The initial analysis for one gene includes a plot of r^2 , as a measure of the proportion of the gene expression variance explained by each SNP. The SNP with highest r^2 is highlighted in the plot. The user can click one SNP in the plot to show a new plot with the association of gene expression and genotypes (or dosage for imputed SNPs).

To reduce false positive results, a threshold line is shown at significance level of 0.01 by default. Also, to account for multiple comparisons, a line with Bonferroni correction threshold is shown, considering the number of SNPs analysed in the selected region. Bonferroni correction may be too conservative since it ignores

linkage disequilibrium (LD) among SNPs. Additional threshold lines can be added by the user.

When one SNP is selected, the plot shows r^2 for the expression of nearby genes (within 500 Kb by default). Options like the ones explained above can be used to select gene ranges, tissues, statistical test or covariates for adjustment.

Analysis of *trans*-eQTL, searching for genes associated to SNPs outside the ± 2 Mb region is not implemented in the web tool, but has been performed and can be downloaded for additional analysis (Supplementary file).

DISCUSSION

Our web tool for eQTL analysis is useful to inform a potential functional roles of SNPs identified in GWAS of CRC¹ or IBD.³² This data source can complement others available specific for colon tissue like the GTEx portal,¹⁷ and contribute to study the genetic architecture of colorectal diseases.

Our whole-genome analysis of eQTL has revealed that genetic variability may explain a substantial fraction of gene expression in colon tissue. Nearly 6% ($n = 869$) of the expressed genes had at least one *cis*-eQTL within 1 Mb. This number of *eGenes* is similar to the reported for other tissues in the report of the pilot GTEx project,¹⁸ for a similar sample size, or that reported by.²²

In an attempt to avoid false positive findings, we have used a non-parametric analysis method, and have restricted to SNP with MAF > 5%. Regarding protection against multiple testing, we have used a significance level of $1e-6$ to search for eQTLs. The reported results at this level had a theoretical FDR of 0.001. We have also performed a permutation test to define the significant threshold for 1% family wise false positive results, and found that we should consider significant only findings with $P < 2e-10$ for *cis*-eQTL and $P < 4e-13$ for *trans*-eQTL. These strict significance levels have reduced the number of significant *eGenes* to 363. As this might be too conservative, for the analysis of specific genes of candidate SNPs, the web tool only reports nominal significance levels for each association and provides several options to threshold (fixed significance levels or Bonferroni correction according to the number of SNPs/genes considered).

We have compared our eQTL findings with those of GTEx and Ongen et al.,²² and found a reasonable good agreement, taking into account that expression data had been generated in diverse platforms and also the populations analysed probably had different environmental exposures that may affect gene expression. If we relax the significance level, more than 64% of our eQTL were reported in GTEx and up to 78% of the *eGenes* were concordant. The agreement with Ongen's results was lower, but probably related to the fact that these authors only reported the most significant eQTL for each gene.

Our expression data is based on 3' microarrays that do not allow to estimate allele-specific expression at individual level or splicing variants. This is a limitation over studies based on RNA-seq for gene expression estimation, which may also better assess low abundance transcripts and have a larger dynamic range.³³ Our microarray data, however, has shown good reproducibility with qPCR in the validation studies performed.²⁵

Only few SNPs identified in GWAS are strong eQTL for candidate genes.²³ Many eQTL, though significant, have low correlation values, that indicate that the proportion of gene expression variance explained by the SNP is small. This may not rule out a causal effect, as lifetime exposures to small alterations in gene expression may be sufficient to increase cancer susceptibility. In fact, the associations observed for most common SNPs with cancer are also of small magnitude.

As the specific SNPs identified in GWAS studies are dependent on the genotyping platform used, SNP imputation is a useful option to increase the likelihood of narrowing the window where the functional SNP lies. We have also used SNP imputation in our

web tool to increase the likelihood of finding stronger eQTL. In fact, we have observed that often the genetic polymorphisms showing highest association with gene expression is an imputed *indel* in the region. *Indels* probably have higher effect disrupting the genome sequence required for enhancers or promoters. SNP imputation may introduce errors in the data and lead to false results. Though recent imputation panels are improving quality for rare variants, we have filtered imputed SNPs with low imputation quality, that generally correspond to low allele frequency. This filter also was decided due to the limited sample size of our sample (144 independent subjects).

The recommended use of our web tool, when the search starts with a candidate SNP identified in a GWAS, is to first search for nearby genes which expression may be related to the SNP. Then, if one or more genes are identified, specific searches by gene may reveal other SNPs in LD with the candidate SNP that are also eQTL but show stronger association. These SNPs are candidate to causal variants and may then be searched in the UCSC genome browser³⁴ to identify whether they lie in regions of open chromatin or other genomic marks suggesting a functional role. In fact, we have compared chromatin marks for our eQTL with those of random SNPs within 1 Mb of the genes' TSS and found that eQTL are significantly enriched in DNase, promoter and enhancer chromatin marks, which is in agreement with previous reports that eQTLs more often fall within regulatory elements.^{22,35} Also, we observed that eQTL more often correspond to GWAS SNPs of diverse traits than random SNPs near genes, as reported by the Haploreg database.³¹

Our study only included colon tissue, but the eQTL analysis should also be useful for rectal cancer. Previous studies have shown that colon and rectal tissue have very similar expression profiles,^{19,36} although these cancer sites may have important differences regarding clinical outcomes.

Our study combines expression data from normal colon tissue obtained from volunteers at colonoscopy, with adjacent macroscopically normal colon tissue obtained from patients. Other studies on eQTL in CRC have analysed only tumour tissue or paired tumour and adjacent normal, without healthy tissue as reference.^{21,22} The inclusion of normal tissue adjacent to tumour may not be completely optimal, as we know that gene expression is altered in adjacent macroscopically normal tissue.³⁷

The inclusion of tumour tissue in the analysis is a matter of debate. Tumour gene expression is very different from normal,²⁰ with genes changing both towards over and under-expression. Although tumours may reveal some new eQTL that are only evident when the gene is overexpressed, diverse genomic alteration occur simultaneously in tumours like copy number aberrations or epigenetic changes, that may generate false positive and false negative results. In our analysis, the list of eQTL identified in tumours had low overlap with that of normal tissue, similar results were obtained by Ongen et al., who observed that 36% of the detected eQTL were tumour specific, and more prone to carry somatic mutations and epigenetic modifications, like methylation changes, compared to matched samples of normal colon mucosa.²² Our web tool allows the user to select which tissues are incorporated into the analysis among healthy colon, adjacent normal colon and tumour. The user should be aware that, if tumour and adjacent tissue are combined, the fact that these samples belong to the same subject may show correlated expression, which is not considered in the statistical analysis.

Even though the potential for a real functional role under a *trans*-eQTL is low, since some long-distant effect mechanism must mediate the association, it is interesting to note that a large number of *trans*-eQTL effects were identified at $P < 1e-6$, but very few remained significant when proper multiple comparisons were accounted for. Most of them were in the same chromosome, probably related to long-distance linkage disequilibrium. The limited sample size of this study does not provide enough power

for a proper *trans*-eQTL analysis, and a deeper analysis would require validation in other datasets and functional studies like knocking-down the candidate *trans*-eQTL to understand the mediation pathway that leads to the association.¹⁴

In conclusion, eQTL analysis is an approach to study functional SNPs identified in GWAS. The web tool that we provide at <https://www.colonomics.org/eQTL-browser> allows easy analysis of expression and eQTL specific for colon tissue and will help researchers in the area to identify which SNPs deserve further functional research.

ACKNOWLEDGEMENTS

We would like to thank Josipa Bilic, Pilar Medina, Carmen Atencia and Isabel Padrol for their technical assistance. The "Xarxa de Bancs de Tumors de Catalunya" sponsored by "Pla Director d'Oncologia de Catalunya (XBTC)", the ICOBiobanc and PLATAFORMA BIOBANCOS PT13/0010/0013 helped with sample collection. This article is based upon work from COST Action BM1206, supported by COST (European Cooperation in Science and Technology) www.cost.eu. Also by the Catalan Institute of Oncology, the Instituto de Salud Carlos III and the Spanish Ministry of Economy and Competitiveness, co-funded by FEDER funds—a way to build Europe—(grants PI08-1635, PS09-1037, PI11-1439, PI14-613, PI14-173 and SAF2016-80888-R), CIBERESP, CIBERONC, CIBEREHD and the Catalan Government DURSI (grants 2014SGR647, 2014SGR255). Samples were obtained with the support of "Plataforma Biobancos" PT13/0010/0013.

AUTHOR CONTRIBUTIONS

V.M. conceived the study and wrote the final version of the manuscript. M.H.A. and A.L.-D. performed statistical analyses; A.C., R.S.-P., D.C. and X.S. performed bioinformatics analyses. A.D. contributed to the web tool implementation. X.V., L.V. and S.C.-B. substantially contributed to define the manuscript focus and discussion. All authors revised and contributed to the final version of the manuscript.

ADDITIONAL INFORMATION

Supplementary information is available for this paper at <https://doi.org/10.1038/s41416-018-0018-9>.

Competing interests: The authors declare no competing interests.

REFERENCES

1. Peters, U., Bien, S. & Zubair, N. Genetic architecture of colorectal cancer. *Gut* **64**, 1623–1636 (2015).
2. Pardini, B., Naccarati, A., Vodicka, P. & Kumar, R. Gene expression variations: potentialities of master regulator polymorphisms in colorectal cancer risk. *Mutagenesis* **27**, 161–167 (2012).
3. Hindorf, L. A. et al. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl Acad. Sci. USA* **106**, 9362–9367 (2009).
4. Peltekova, V. D. et al. Identification of genes expressed by immune cells of the colon that are regulated by colorectal cancer-associated variants. *Int. J. Cancer* **134**, 2330–2341 (2014).
5. Pennacchio, L. A., Bickmore, W., Dean, A., Nobrega, M. A. & Bejerano, G. Enhancers: five essential questions. *Nat. Rev. Genet.* **14**, 288–295 (2013).
6. Nica, A. C. & Dermitzakis, E. T. Using gene expression to investigate the genetic basis of complex disorders. *Hum. Mol. Genet.* **17**, R129–R134 (2008).
7. Cheung, V. G. et al. Natural variation in human gene expression assessed in lymphoblastoid cells. *Nat. Genet.* **33**, 422–425 (2003).
8. Zhu, J. et al. Increasing the power to detect causal associations by combining genotypic and expression data in segregating populations. *PLoS Comput. Biol.* **3**, e69 (2007).
9. Gilad, Y., Rifkin, S. A. & Pritchard, J. K. Revealing the architecture of gene regulation: the promise of eQTL studies. *Trends Genet.* **24**, 408–415 (2008).
10. Hou, L. & Zhao, H. A review of post-GWAS prioritization approaches. *Front Genet* **4**, 280 (2013).
11. Joehanes, R. et al. Integrated genome-wide analysis of expression quantitative trait loci aids interpretation of genomic association studies. *Genome Biol.* **18**, 16 (2017).
12. Li, L. et al. Using eQTL weights to improve power for genome-wide association studies: a genetic study of childhood asthma. *Front. Genet.* **4**, 103 (2013).
13. Hulse, A. M. & Cai, J. J. Genetic variants contribute to gene expression variability in humans. *Genetics* **193**, 95–108 (2013).

14. Cheung, V. G. et al. Polymorphic cis- and trans-regulation of human gene expression. *PLoS Biol.* **8**, e1000480 (2010).
15. Stranger, B. E. et al. Population genomics of human gene expression. *Nat. Genet.* **39**, 1217–1224 (2007).
16. Nica, A. C. et al. The architecture of gene regulatory variation across multiple human tissues: the MuTHER study. *PLoS. Genet.* **7**, e1002003 (2011).
17. The GTEx Consortium. The genotype-tissue expression (GTEx) project. *Nat. Genet.* **45**, 580–585 (2013).
18. The GTEx Consortium. Human genomics. The genotype-tissue expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* **348**, 648–660 (2015).
19. The Cancer Genome Atlas Network. Comprehensive molecular characterization of human colon and rectal cancer. *Nature* **487**, 330–337 (2012).
20. Cordero, D. et al. Large differences in global transcriptional regulatory programs of normal and tumor colon cells. *BMC Cancer* **14**, 708 (2014).
21. Loo, L. W. et al. cis-Expression QTL analysis of established colorectal cancer risk variants in colon tumors and adjacent normal tissue. *PLoS ONE* **7**, e30477 (2012).
22. Ongen, H. et al. Putative cis-regulatory drivers in colorectal cancer. *Nature* **512**, 87–90 (2014).
23. Closa, A. et al. Identification of candidate susceptibility genes for colorectal cancer through eQTL analysis. *Carcinogenesis* **35**, 2039–2046 (2014).
24. Singh, T. et al. Characterization of expression quantitative trait loci in the human colon. *Inflamm. Bowel Dis.* **21**, 251–256 (2015).
25. Sole, X. et al. Discovery and validation of new potential biomarkers for early detection of colon cancer. *PLoS ONE* **9**, e106748 (2014).
26. 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.* **39**, 906–913 (2007).
27. Delaneau, O., Zagury, J. F. & Marchini, J. Improved whole-chromosome phasing for disease and population genetic studies. *Nat. Methods* **10**, 5–6 (2013).
28. Freidlin, B., Zheng, G., Li, Z. & Gastwirth, J. L. Trend tests for case-control studies of genetic markers: power, sample size and robustness. *Hum. Hered.* **53**, 146–152 (2002).
29. Shabalin, A. A. Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics* **28**, 1353–1358 (2012).
30. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B* **57**, 289–300 (1995).
31. Ward, L. D. & Kellis, M. HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic Acids Res.* **40**, D930–D934 (2012).
32. Cleynen, I. & Vermeire, S. The genetic architecture of inflammatory bowel disease: past, present and future. *Curr. Opin. Gastroenterol.* **31**, 456–463 (2015).
33. Zhao, S., Fung-Leung, W. P., Bittner, A., Ngo, K. & Liu, X. Comparison of RNA-Seq and microarray in transcriptome profiling of activated T cells. *PLoS ONE* **9**, e78644 (2014).
34. Kent, W. J. et al. The human genome browser at UCSC. *Genome Res.* **12**, 996–1006 (2002).
35. Grubert, F. et al. Genetic control of chromatin states in humans involves local and distal chromosomal interactions. *Cell* **162**, 1051–1065 (2015).
36. Sanz-Pamplona, R. et al. Gene expression differences between colon and rectum tumors. *Clin. Cancer Res.* **17**, 7303–7312 (2011).
37. Sanz-Pamplona, R. et al. Aberrant gene expression in mucosa adjacent to tumor reveals a molecular crosstalk in colon cancer. *Mol. Cancer* **13**, 46 (2014).



Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License, which permits any non-commercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. If you remix, transform, or build upon this article or a part thereof, you must distribute your contributions under the same license as the original. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by-nc-sa/4.0/>.