Gene expression variations: potentialities of master regulator polymorphisms in colorectal cancer risk

Barbara Pardini1,*, Alessio Naccarati1, Pavel Vodicka1,2 and Rajiv Kumar3

1Department of Molecular Biology of Cancer, Institute of Experimental Medicine, Academy of Sciences of the Czech Republic, Vidoňská 1083, 14220 Prague 4, Czech Republic, 2Institute of Biology and Medical Genetics, First Medical Faculty, Charles University, Albertov 4, 12800 Prague 2, Czech Republic and 3Division of Molecular Genetic Epidemiology, German Cancer Research Center (DKFZ), Im Neuenheimer Feld 580, 69120 Heidelberg, Germany.

To whom correspondence should be addressed. Institute of Experimental Medicine, Videnska 1083, 14200 Prague 4, Czech Republic. Tel: +420 2 41062782; Email: pardini@biomed.cas.cz

Received on July 13, 2011; revised on July 29, 2011; accepted on August 01, 2011

Colorectal cancer (CRC) is one of the most common cancers worldwide with a peak of incidence in industrialised countries. It is a complex disease related to environmental and genetic risk factors. Low-penetrance genetic variations contribute significantly to sporadic and familial form of CRC. Genome-wide association studies (GWAS) have uncovered numerous robust associations between common variants and CRC risk; only a few of those were protein altering non-synonymous polymorphisms. One of the hypotheses is that non-coding and intergenic variants may change the expression levels of one or several target genes and, thus, account for a fraction of phenotypic differences, including susceptibility to CRC. Such genetic variations have been detected as expression quantitative loci (eQTLs) that show linkage/association to a large number of genes and have been defined as “master regulators of transcription”. In the present work, we overview the potentialities to use results from GWAS and eQTL studies in the identification as well as investigation of master regulators in CRC susceptibility.

Colorectal cancer

General information

Colorectal cancer (CRC) is one of the most common malignancies in the world, with particularly high incidence rates in Western countries (1). Worldwide, there were estimated 1 230 000 new cases in the year 2008. CRC is also the third leading cause of cancer death, with an estimated 680 000 deaths reported in 2008 (2). One of the consistent etiological factors associated with risk of the disease has been the Western-style diet, a highly calorific and rich in animal fat, which in combination with sedentary lifestyle constitutes an extremely high risk factor. Those inferences have been reinforced by epidemiological studies on migrant populations. One of the possible explanations for the correlation between CRC and diet is perhaps direct mucosal contact, in colon and rectum, with food components and probable exposure to diet-induced metabolic and physiologic changes (2–6). However, CRC like other multifactorial diseases is fundamentally driven and modulated by individual genetic backgrounds that most likely interact with environmental risk factors and influence risk at population level.

Colorectal tumours progress through a series of clinical and histopathologic stages that range from single crypt lesions (aberrant crypt foci) through small benign tumours (adenomatous polyps) to malignant cancers (carcinomas). The transition of normal epithelia cells to carcinoma cells has been shown to involve a series of somatic genetic changes that include inactivation of tumour suppressor genes as well as oncogenic activation (7). The major somatic genetic alterations include activating mutations in K-RAS, inactivation of tumour suppressor genes on chromosomes 17p (TP53), 5q and 18q (8). Mutations in K-RAS occur in about 50% of CRC and in adenomas. Screening of K-RAS mutations constitutes an integral part of regular diagnosis prior to the treatment of patients with metastatised disease (9). Somatic mutations in adenomatous polyposis coli (APC) gene, located on chromosome 5q, are present in over 80% of sporadic CRC. Inherited mutations in the gene are responsible for familial adenomatous polyposis (FAP) coli, a genetic autosomal dominant disease. Somatic alterations in the APC gene that occur in sporadic CRC are similar to inherited truncating mutations in FAP. APC mutations are the earliest lesion and are, therefore, considered as gatekeeper event in colorectal tumorigenesis (10).

Genetics of CRC

Colorectal carcinoma is generally classified into three categories. Familial CRC (around 30% of total CRC cases) refers to patients with at least one blood relative with CRC or an adenoma but with no specific germline mutations or clear pattern of inheritance. The majority, approximately 67% of CRC cases, are sporadic or show a pattern of familial aggregation, not fitting into models of Mendelian inheritance (10). The remaining 3% of CRC patients belongs to cases with inherited CRC syndromes (11,12). Those are carriers of highly penetrant germline mutations in cancer susceptibility genes. The most representative syndromes are the ones characterised by multiple benign colorectal polyps and others without polyposis. The first group includes several polyposis syndromes, such as FAP, Peutz-Jegher syndrome, familial juvenile polyposis. The hereditary nonpolyposis colorectal cancer, also known as Lynch syndrome, is characterised by germline mutations in different DNA mismatch repair genes (13).

The complex etiology of sporadic CRC involves a combinatorial impact of genetic and environmental risk factors (10,14). Epidemiological studies have identified several risk factors, including meat, tobacco and alcohol consumption, that modulate risk of CRC (reviewed in ref. 5). On the other hand, vegetable consumption, prolonged use of non-steroidal anti-inflammatory drugs, oestrogen replacement therapy and physical activity are known as protective factors (5). Other significant etiological factors are chronic inflammatory bowel
diseases and a positive family history of CRC. Population-based studies showed that ulcerative colitis results in a 20-fold increased incidence of CRC and a 4.4-fold increase in mortality, while Crohn’s disease causes 3-fold increase in the risk of colorectal malignancies (15,16). Family history is established to be one of the strongest risk factors for the development of the disease, as almost 30% of the individuals report a positive family history (17).

It had been long hypothesized and subsequently shown experimentally that low-penetrant genetic variants are probably causal for the inter-individual variation in susceptibility to CRC. CRC, as most of forms of human cancers, shows complex inheritance in the general population. Contributors to these complex patterns can be combination of genetic variants with common-to-rare frequency and low-to-intermediate penetrance (18). Although estimated inherited susceptibility accounts for about 35% of variance in CRC risk, high-penetrant germline mutations account for only 6% of cases (19,20). Much of the remaining variation in genetic risk is likely to be a consequence of the co-inheritance of multiple low-penetrance variants, which are common in general population. The “common-disease common-variant” model of CRC implied that association analyses based on scans of polymorphic variants should be a powerful strategy for identifying low-penetrance susceptibility alleles (14). While the association studies based on a candidate gene approach did identify some variants involved in CRC susceptibility, but success of such investigations remained limited and many times results were not reproducible by independent studies. Recent meta-analyses of various candidate-gene based case-control studies have provided limited support for the role of variants in the MTHFR, CCND1, GSTT1, XPC, NQO1 and NAT2 genes (21–27).

GWAS, genetic expression variation and CRC

The advent of so-called genome-wide association studies (GWAS) has resulted in robust identification of genetic variants associated with susceptibility of various complex diseases. Several GWAS have also been applied to investigate the impact of common genetic variations on CRC susceptibility (28–43). So far, various chromosomal loci have been shown to modestly increase CRC risk (Table I) (28–36,41,43). The identified variants, with minor allele frequencies between 7% and 90%, have been shown to increase risk of CRC with allelic odds ratios < 1.3 (44). Though, the individual associated variants identified through GWAS impart only a modest risk of CRC, the population attributable fractions have been shown to be significant because of substantial minor allele frequencies. Moreover, through the possibility of a concerted effect, the variants can potentially influence an individual’s risk of developing CRC significantly (14). Based on the location within the genome, most of the identified associated single-nucleotide polymorphisms (SNPs) cannot be per se causal but rather indicate the possibility of being in linkage with the “real” causal variants. Nevertheless, the role of some of the associated variants in transcription either through cis- or trans-regulation remains plausible as indicated by several investigators (45–48). In one such study, it has been shown that a variant at chromosome 8q23.3 may influence the transcriptional regulation of eukaryotic translation initiation factor 3H (EIF3H), providing support for the functional significance of the variant and observed association with CRC (46). This class

Table I. SNPs associated with risk of CRC identified and validated through GWAS

<table>
<thead>
<tr>
<th>Polymorphism</th>
<th>Chromosome</th>
<th>Locus/genes (major/minor alleles)</th>
<th>References</th>
<th>Cases/controls</th>
<th>ORc</th>
<th>95% CI</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs16892766</td>
<td>8q23.3</td>
<td>EIF3H (A/C)</td>
<td>Tomlinson et al. (36)</td>
<td>922 (17872)/927 (17526)</td>
<td>1.27</td>
<td>1.20–1.34</td>
<td>3 × 10^-18</td>
</tr>
<tr>
<td>rs10795668</td>
<td>10p14</td>
<td>Intergenic (G/A)</td>
<td>Tomlinson et al. (36)</td>
<td>922 (17872)/927 (17526)</td>
<td>1.12</td>
<td>1.10–1.16</td>
<td>3 × 10^-13</td>
</tr>
<tr>
<td>rs69383267</td>
<td>8q24.21</td>
<td>c-MYC (G/T)</td>
<td>Tomlinson et al. (36)</td>
<td>922 (17872)/927 (17526)</td>
<td>1.24</td>
<td>1.17–1.33</td>
<td>7 × 10^-11</td>
</tr>
<tr>
<td>rs10936599</td>
<td>3q26.2</td>
<td>MYNN (C/A)</td>
<td>Tomlinson et al. (30)</td>
<td>930 (7334)/960 (5266)</td>
<td>1.27</td>
<td>1.16–1.39</td>
<td>1 × 10^-14</td>
</tr>
<tr>
<td>rs4939827</td>
<td>18q21.1</td>
<td>SMAD7 (T/C)</td>
<td>Zanke et al. (31)</td>
<td>1257 (6223)/1336 (6443)</td>
<td>1.17</td>
<td>1.12–1.23</td>
<td>3 × 10^-11</td>
</tr>
<tr>
<td>rs4779584</td>
<td>15q13.3</td>
<td>GREM1/SCG5 (C/T)</td>
<td>Cui et al. (41)</td>
<td>1583 (4809)/1898 (2973)</td>
<td>1.18</td>
<td>1.11–1.25</td>
<td>2 × 10^-8</td>
</tr>
<tr>
<td>rs3802842</td>
<td>11q23.1</td>
<td>CCND1 (A/C)</td>
<td>Broderick et al. (28)</td>
<td>940 (7473)/965 (5984)</td>
<td>1.16</td>
<td>1.09–1.27</td>
<td>1 × 10^-12</td>
</tr>
<tr>
<td>rs4444235</td>
<td>14q2.22</td>
<td>BMP4 (T/C)</td>
<td>Tenesa et al. (35)</td>
<td>981 (16476)/1002 (15351)</td>
<td>1.2</td>
<td>1.16–1.24</td>
<td>8 × 10^-10</td>
</tr>
<tr>
<td>rs9929218</td>
<td>16q22.1</td>
<td>CDH1 (G/A)</td>
<td>Tomlinson et al. (36)</td>
<td>922 (17872)/927 (17526)</td>
<td>1.23</td>
<td>1.14–1.34</td>
<td>5 × 10^-7</td>
</tr>
<tr>
<td>rs10412110</td>
<td>19q13.11</td>
<td>RHDN2 (C/T)</td>
<td>Tenesa et al. (35)</td>
<td>981 (16476)/1002 (15351)</td>
<td>1.11</td>
<td>1.08–1.15</td>
<td>6 × 10^-10</td>
</tr>
<tr>
<td>rs961253</td>
<td>20p12.3</td>
<td>Intergenic (C/A)</td>
<td>Houlston et al. (32)</td>
<td>902 (4878)/1929 (4914)</td>
<td>1.11</td>
<td>1.08–1.15</td>
<td>8 × 10^-10</td>
</tr>
<tr>
<td>rs10936599</td>
<td>3q26.2</td>
<td>MYNN (C/T)</td>
<td>Houlston et al. (38)</td>
<td>3334 (14851)/4628 (15569)</td>
<td>1.04</td>
<td>1.04–1.10</td>
<td>3 × 10^-8</td>
</tr>
<tr>
<td>rs667758</td>
<td>1q4</td>
<td>DUSP10 (A/G)</td>
<td>Houlston et al. (38)</td>
<td>3334 (14851)/4628 (15569)</td>
<td>1.09</td>
<td>1.06–1.12</td>
<td>2 × 10^-9</td>
</tr>
<tr>
<td>rs4925386</td>
<td>20q13.33</td>
<td>LAMAS (C/T)</td>
<td>Houlston et al. (38)</td>
<td>3334 (14851)/4628 (15569)</td>
<td>1.08</td>
<td>1.05–1.10</td>
<td>2 × 10^-10</td>
</tr>
<tr>
<td>rs11169552</td>
<td>12q13.12</td>
<td>LARP4/DIP2 (C/T)</td>
<td>Houlston et al. (38)</td>
<td>3334 (14851)/4628 (15569)</td>
<td>1.09</td>
<td>1.05–1.11</td>
<td>2 × 10^-10</td>
</tr>
<tr>
<td>rs6691170</td>
<td>1q4</td>
<td>DUSP10 (G/T)</td>
<td>Houlston et al. (38)</td>
<td>3334 (14851)/4628 (15569)</td>
<td>1.06</td>
<td>1.03–1.09</td>
<td>1 × 10^-9</td>
</tr>
<tr>
<td>rs7578229</td>
<td>6q25.3</td>
<td>SLC22A3 (G/T)</td>
<td>Cui et al. (41)</td>
<td>1583 (4809)/1898 (2973)</td>
<td>1.28</td>
<td>1.18–1.39</td>
<td>8 × 10^-9</td>
</tr>
<tr>
<td>rs7014346</td>
<td>18q24.21</td>
<td>POU5F1/JHG57825/DQS15897 (G/A)</td>
<td>Tenesa et al. (35)</td>
<td>981 (16476)/1002 (15351)</td>
<td>1.19</td>
<td>1.15–1.23</td>
<td>9 × 10^-11</td>
</tr>
</tbody>
</table>

*The risk allele for each SNP is highlighted in bold. EIF3H, eukaryotic translation initiation factor 3, subunit B; c-MYC, v-myc avian myelocytomatosis viral oncogene homologue; SMAD7, SMAD family member 7; GREM1, gremlin 1; SCG5, secretogranin V; C110r93, chromosome 11 open reading frame 93; BMP4, bone morphogenetic protein 4; CDH1, E-cadherin; RHNP2, Rho GTPase binding protein 2; MYNN, myoneuron; DUSP10, dual specificity phosphatase 10; LAMAS, laminin alpha 5; LARP4, La ribonucleoprotein domain family, member 4; DIP2, DIP2 disco-interacting protein 2 homologue A (Drosophila); SLC22A3, solute carrier family 22 (extraneuronal monoamine transporter), member 3; POU5F1/JHG57825/DQS15897, hypothetical LOC727677.

OR, odds ratio.

95% CI, 95% confidence interval.
of susceptibility variants has been suggested to be of potential public health importance, allowing risk stratification within populations (14). The identification of new risk variants may also be helpful in characterising, more specifically, new cancer pathways that may lead to new prevention or treatment strategies for CRC (14).

Despite remarkable success of GWAS in identifying disease-associated loci, a gnawing gap remains in understanding the mechanism that leads to the associated increased risk. The genetic risk may be probably due to both highly prevalent loci with very modest effect sizes and rare loci with strong effect sizes (49). There have been, so far, only some rare examples in which the observed pattern of genetic association corresponded to a functional candidate variant with a precise localisation of the functional locus. In the majority of GWAS, regional linkage disequilibrium extends across multiple genes and the disease-associated variants serve as proxies for unrecognised non-coding variation, precluding claims of disease-specific gene identification (50). Another common feature characterising GWAS is that many disease-linked SNPs identified to date are located within introns or inter-genic regions of human genome, which have no direct relations to known protein-coding genes or microRNA genes (51). It has been correctly suggested that the biological interpretation of GWAS signal is daunting. Candidate loci fall either in gene deserts or in regions with equally plausible causative genes (52). Thus, non-canonical mechanisms of phenotype-altering effects of genetic variations cannot be ruled out. For example, Glinskii et al. proposed that inter-genic DNA sequence variations associated with human disorders may affect phenotypes through trans- action via non-protein-coding SNP sequence-bearing RNA intermediates (51,53). An alternative approach to solve these problems and to reduce costs associated with GWAS and to overcome limitations of multiple testing corrections and sample size is to select SNPs on basis of a functional effect, such as changes in gene expression levels (52,54–56).

**Expression quantitative trait loci studies: a way to mapping DNA variants that influence gene expression**

The levels of expression of different genes vary among individuals and contribute to phenotypic diversity and differences in disease susceptibility (57). It has been shown that several chromosomal regions contain germline determinants that regulate gene expression phenotypes (58). The quantitative differences in gene expression might be responsible for a large source of natural variation in populations, as well as crucial in accounting a large fraction of phenotypic variability (59,60). Variable gene expression levels among individuals can be analysed like other quantitative phenotypes (61). Functional genomics, in particular expression quantitative trait loci (eQTL) studies, are useful for determining relationship between sequence variation in human population and phenotypic diversity (62). The main goal of eQTL studies in humans has been the identification of DNA variants (polymorphisms) that influence the expression levels of genes (also defined as gene expression phenotype). The basic idea is that sequence variation may be crucial in affecting the steady-state level of mRNA molecules of a particular gene in a specific cell type or tissue. eQTLs studies have led to interesting results for the identification of regulatory regions and DNA sequence variants that may modulate levels of expression of genes in a range of organisms. Interestingly, eQTL studies could fill in the gaps in understanding the causal reasons for the association of loci with the disease identified through GWAS (63). eQTL studies connect variation at DNA level with the quantitative variation in RNA transcription (61). Over 15 million SNPs have been identified so far and while some of those are functional, the majority is most probably neutral (64). eQTL studies simplify the search for functional variants that regulate gene expression through scan of genome for regulators of gene expression without the need for a priori knowledge of regulatory mechanisms (61). Gene expression levels are controlled by a combination of cis- and trans-acting regulators. Gene expression variation may also be considered a possible candidate to establish a link between DNA sequence variability and clinical phenotypes (65). This represents a paradigm shift from the traditional point of view about the role of genes in the susceptibility to human diseases as an effect of their variability in the sequence.

Several disease-linked associations with significant eQTLs have emerged and such studies have also been extended to cancer susceptibility (63). eQTLs and gene expression differences have been evaluated to elucidate the causal relation between validated SNPs identified through GWAS and the risk of lung cancer in never smokers (66). Out of 44 candidate SNPs identified that might alter lung cancer risk in the latter, only one was replicated in four independent studies. From the eQTL analysis, the investigators observed a strong correlation between genotypes of the replicated SNP and the transcription levels of GPCS gene in normal lung tissues, showing that genetic variants at a specific locus (13q31.3), associated with susceptibility to lung cancer in never smokers, alter the expression of GPCS (66). Out of 14 loci associated with susceptibility to CRC, discovered through various GWAS, identification of the causal variants and putative functionalities for some have been attempted. In one such study, investigators fine mapped the risk-associated region, identified earlier through GWAS, for identification of functional variants (47). In the three fine-mapped regions, a cluster of SNPs within the narrow areas showed stronger signals for association with the CRC risk than the surrounding GWAS identified SNPs. Using an eQTL browser, some of the top associated variants in the regions analysed have shown an association with different transcript levels (47). Interestingly, some of the genes previously proposed as best candidates in those blocks (EIF3, CDH1 and RHPN2) showed no association with different eQTLs. Those genes had been originally selected because of their putative involvement in CRC tumorigenesis and vicinity to the risk-associated tagSNPs. On the other hand, a SNP in an intron of ZFP90 was found to be associated with CRC risk and with different transcript levels of the gene. Similarly, UTP23 appeared to be a possible target of the genetic variation associated with CRC in the 8q23.3 region (47). However, it is still difficult to draw too many conclusions for now, since eQTL data sets are not available specifically for colorectal cells and the observed associations have been detected only in blood cells. It is possible that implementation of expression studies in colorectal mu cosa could clarify the functionality of the SNPs associated with the CRC risk. In fact, most eQTL studies, currently, are conducted on lymphoblastoid cell lines, which do have some utility for the interpretation of human disease associations, particularly with immunity-related phenotypes (52). At the same time potentiality of artefacts associated with
immortalisation, subsequent passage, and growth conditions prior to harvest cannot be precluded (67).

Hotspots or master regulators

It is likely that genetic variation within a gene, encoding a transcription factor with multiple targets, can potentially influence expression of most of its target genes. Such variations are of particular biological interest, which can be, in principle, used to identify groups of genes that are controlled by a common transcription factor (68). The term “master regulators” has been applied to those genetic variants that have been detected as eQTL and which show linkage/association to a large number of genes (58). One of the common features observed in studies of transcript expression is the presence of hotspots, that is the individual loci that affect a large number of transcripts (69). Hotspots or “master regulators of transcription” are defined as the loci for which the number of associated transcripts exceeds the expected, based on assumption of random distribution along the genetic map (58,70). The description of master regulators is currently not uniform. One approach takes into consideration the number of expression traits that best map to a genetic marker regardless of statistical significance at the whole-genome level (71). Another study divided the autosomal genome of lymphoblastoid cell lines into 491 windows of 5Mb and determined the number of regulators mapping to each window (71). Assuming random distribution across the genome of phenotype regulators, the probability of finding six or more hits per window would be very low. On the contrary, in that study, investigators found two hotspots with six or more hits. In others studies, the number of whole-genome significant linkages at each marker have been counted to identify transcriptional master regulators (70,75,76). In one study, two different linkage strategies to localise master regulators of transcription were applied. One strategy involved classical-variance components with use of information from grandparents and the other based on a linkage method by regression with floating variable selection that allowed testing linkage models with combinations of genetic locations. The former method identified 11 hotspots and the latter 17 with an overlap of five hotspots (71).

Studies in yeast and other organisms have identified hotspots that contain genetic variants that influence multiple expression phenotypes (reviewed in refs. 61 and 77). Human studies, thus far, have shown less uniform results. The presence of master regulatory regions has been either reported (58,78) or not (54,79–82). Surprisingly, only very few verified hotspots have been identified. The majority of the results indicate a large abundance of local eQTLs that coincide with the position of the gene and are presumably cis-acting polymorphisms in the promoter region. Genetic variants in hotspots, due to their characteristics, are expected to act more in a trans manner than observed. It is likely that the differences among studies are partly because of differences in power to detect trans-acting variants. Distal eQTLs are more difficult to be identified. It is probable that indirect regulation mechanism results in reduced statistical power that limits the reliable detection of hotspots (77,83).

The mechanism of gene expression regulation by master regulators, identified thus far, currently can be at the best speculated. The target genes with phenotypes that map to the same hotspots often share similar functions or reside close to each other. As genes that share functions are often co-regulated, their polymorphic regulators would appear as hotspots in eQTL studies. The expression levels of co-

---

**Fig. 1.** Integration between GWAS and eQTL mapping in cancer susceptibility studies (modified from ref. 63).
regulated genes frequently show significant correlations that may be biologically important. But sometimes the number of phenotypes. Phenotypic buffering of protein-coding polymorphisms has been already documented and it represents a general property of complex gene-regulatory networks (90–92). Most likely, common alleles are predominantly buffered by the robust properties of the system and then largely “neutral” for the rest of the molecules in the system. This scenario may have profound consequences for the design and interpretation of eQTL studies. eQTLs and their hotspots, from the reported studies, seem to be scarce and difficult to identify; those already reported need to be interpreted with caution. The rarity of eQTL master regulatory loci needs to be elucidated in future studies.

Conclusions

The complexity of CRC like other cancers has made understanding of underlying genetic factors difficult. The success of GWAS in identifying various genetic loci associated with the disease has opened new avenues from the perspective of understanding the genesis of the disease and its prevention and treatment. However, a gap in understanding the role of identified variants through GWAS remains to be fulfilled before any practical utility. Those identified variants being associated with expression regulation through cis- and trans-action remains a probability. Scarcity of information available about master regulators of expression and cancer risk association is mainly due to the difficulties in performing eQTL studies on cell lines other than the lymphoblastoid ones. eQTL studies may be a powerful tool for clarifying the role of candidate loci identified in GWAS. Master regulators can be more pertinent to provide information about the loci widely involved in the disease risk and in an extension also about the target genes and therefore, legitimate targets for future investigations.

Acknowledgements

Conflict of interest statement: None declared.

Funding


References