2016 Colorectal Cancer: Global view

Integration of genome scale data for identifying new players in colorectal cancer

Viktorija Sokolova, Elisabetta Crippa, Manuela Gariboldi

Viktorija Sokolova, Elisabetta Crippa, Manuela Gariboldi, Department of Experimental Oncology and Molecular Medicine, Fondazione IRCCS Istituto Nazionale dei Tumori, 20133 Milano, Italy

Author contributions: Sokolova V and Gariboldi M wrote the manuscript; Crippa E contributed to writing the manuscript and collected the bibliography.

Supported by Associazione Italiana per la Ricerca sul Cancro, Grants No. 10529 and No. 12162; and by funds obtained through an Italian law that allows taxpayers to allocate 0.5% share of their income tax contribution to a research institution of their choice.

Conflict-of-interest statement: No potential conflicts of interest.

Open-Access: This article is an open-access article which was selected by an in-house editor and fully peer-reviewed by external reviewers. It is distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited and the use is non-commercial. See: http://creativecommons.org/licenses/by-nc/4.0/

Correspondence to: Manuela Gariboldi, PhD, Department of Experimental Oncology and Molecular Medicine, Fondazione IRCCS Istituto Nazionale dei Tumori, via G. Venezian 1, 20133 Milan, Italy. manuela.gariboldi@istitutotumori.mi.it

Telephone: +39-2-23902042
Fax: +39-2-23903141

Received: June 22, 2015
Peer-review started: June 26, 2015
First decision: September 11, 2015
Accepted: November 9, 2015

Abstract

Colorectal cancers (CRCs) display a wide variety of genomic aberrations that may be either causally linked to their development and progression, or might serve as biomarkers for their presence. Recent advances in rapid high-throughput genetic and genomic analysis have helped to identify a plethora of alterations that can potentially serve as new cancer biomarkers, and thus help to improve CRC diagnosis, prognosis, and treatment. Each distinct data type (copy number variations, gene and microRNAs expression, CpG island methylation) provides an investigator with a different, partially independent, and complementary view of the entire genome. However, elucidation of gene function will require more information than can be provided by analyzing a single type of data. The integration of knowledge obtained from different sources is becoming increasingly essential for obtaining an interdisciplinary view of large amounts of information, and also for cross-validating experimental results. The integration of numerous types of genetic and genomic data derived from public sources, and via the use of ad-hoc bioinformatics tools and statistical methods facilitates the discovery and validation of novel, informative biomarkers. This combinatory approach will also enable researchers to more accurately and comprehensively understand the associations between different biologic pathways, mechanisms, and phenomena, and gain new insights into the etiology of CRC.

Key words: Colorectal cancer; Copy number variations; Gene expression; miRNA expression; Methylome; Data integration

© The Author(s) 2016. Published by Baishideng Publishing Group Inc. All rights reserved.
Core tip: The development of colorectal cancer (CRC) is driven by the accumulation of various genetic and epigenetic alterations, which have been only partially identified. The increasing financial affordability of high-throughput genome-wide assays has enabled the comprehensive analysis of genomic, transcriptomic, and epigenetic data obtained by analyzing the same biologic samples, and thereby facilitated the identification of new molecular players in CRC. An integrative approach that considers all of these multiple factors provides better results when seeking to identify genes or microRNAs related to new interactions or biomarkers that might improve CRC diagnosis, prognosis, and treatment.


INTRODUCTION

Due to its incidence of one million new cases and mortality rate of > 500,000 deaths per year, colorectal cancer (CRC) is now the third most common type of cancer, and the third leading cause of cancer-related death worldwide[1]. CRC can be classified on the basis of its clinical, pathologic, and genetic characteristics, and is commonly described as a progressive malignant transformation of the normal colonic epithelium to invasive adenocarcinoma due to an accumulation of acquired genetic and epigenetic aberrations[2].

At least three different pathogenetic mechanisms have been proposed to explain the development of CRC. Chromosomal instability (CIN) is thought to account for 85% of all cases, while microsatellite instability (MSI) and the presence of a CpG island methylator phenotype (CIMP) may account for a majority of the remaining cases. The CIN pathway involves the sequential deregulation of tumor suppressors and oncogenes, and includes mutation of the APC gene and/or the loss of chromosome 5q where it maps. CIN can also refer to mutation of the KRAS oncogene, the loss of chromosomal arm 18q, and deletion of chromosome 17p, which harbors the tumor suppressor gene TP53[3].

MSI results from loss of the DNA mismatch repair system. This loss destabilizes repetitive units of DNA (DNA microsatellites), resulting in the generation of inactivating frameshift mutations in the coding sequences of tumor suppressor genes. Tumors that display MSI are divided into two subtypes: MSI-H (instability in > 30% of microsatellites examined) and MSS/MSI-L (instability in < 30% of microsatellites examined)[4]. The CIMP pathway is characterized by hypermethylation of DNA located in CpG islands (specific regulatory sites enriched in CpG motifs) found in the promoter regions of tumor suppressor genes. Such hypermethylation suppresses gene transcription[5]. CRCs with a CIMP display unique epigenetic phenotypes, and well-defined clinical, pathologic, and molecular profiles[6].

Recent studies of these distinct pathways have highlighted the heterogeneity of CRC, and their results have suggested that all CRCs cannot be fully explained by the initial model proposed by Fearon et al[2]. The results of recent large-scale sequencing studies have shown that CRC development requires the participation of numerous critical “driver” genes and “passenger” genetic alterations, many of which remain to be identified[7-9]. Identification of these additional genetic alterations should allow physicians to better characterize and identify the various clinical stages of CRC, utilize such aberrations as biomarkers for the early diagnosis and prognosis of CRC, or to develop treatments for CRC patients.

In recent years, thousands of non-coding RNAs [e.g., microRNAs (miRNAs), long non-coding RNAs, and competitive endogenous RNAs] have been identified as key regulators of various cellular processes that control tumor initiation and progression[10,11]. As the most studied type of non-coding RNA, miRNAs are thought to regulate the post-transcriptional expression of tumor suppressor genes and oncogenes involved in CRC development. Moreover, due to their high chemical stability, miRNAs represent potential biomarkers for use in diagnosing and monitoring human cancers[12]. Additionally, certain miRNAs may be associated with a patient’s response to treatment[13].

The various genomic alterations described above can all contribute to helping investigators understand the vast landscape of CRC. Furthermore, the ability to simultaneously measure and integrate the effects of such alterations should assist in identifying previously unknown molecular changes in key genomic factors, and contribute to a better understanding of how such changes interact with each other to induce the development of CRC. A combination of these data might also be used to help predict disease risk or patient outcomes. The continuous improvement of high-throughput screening and sequencing technologies has made it possible to use the same biologic samples to gather data for several different projects. For instance, The Cancer Genome Atlas (TCGA)[14] was compiled with the goal of first profiling and then integrating the genomic changes that occur in cancers, including CRC[15]. Although several limitations can hamper such an analysis (e.g., patterns of missing data and noise across different data types), numerous bioinformatics tools and statistical frameworks are now available for integrating multiple genomic features found in the same sample, and then assist in investigating their related biologic pathways or gene sets[16].
Here, we reviewed the most relevant publications that have integrated various types of molecular data related to CRC, with the goal of identifying new biomarkers for CRC detection and progression, or targets that may assist in dissecting the mechanisms involved in CRC development.

INTEGRATION OF COPY NUMBER VARIATION AND GENE EXPRESSION DATA

Several studies have shown that a gain or loss in the number of copies of DNA segments affects the expression of genes as well as miRNAs positioned within it. Moreover, such changes are known to affect cancer-related biologic processes\(^{17-17}\). Although not all genes with increased copy number are overexpressed\(^ {18-20}\), those that display a strong positive correlation between expression and copy number may play important roles in cancer progression. However, copy number variations (CNVs) alone cannot fully explain the altered expression levels of all genes, because changes in gene expression are also determined by complex mechanisms that regulate gene transcription. Nevertheless, integrating the results of gene expression analyses with genomic profiling results in an efficient approach for discovering novel cancer-related genes.

Numerous studies on gene CNVs in samples of CRC tissue have yielded similar results regarding the focal genomic regions of gains and losses, and confirmed the high prevalence of gains on chromosomes 8q, 13, and 20q, and losses on chromosomes 8p, 17p, and 18q\(^ {15,16,21-23}\).

The most recent published studies that integrated data concerning gene expression and CNV are shown in Table 1. Ali Hassan et al\(^ {24}\) found that increased gene expression was correlated with significantly increased copy numbers of genes that were mostly located on chromosome 20q12, where eight highly expressed genes were identified. Moreover, three of those genes (TOP1, PLCG1, and PTPRT) were related to CRC. The highest number of copy losses was observed on chromosome 8p23.2 and was correlated with reduced expression of CSMD1 and DLC1. A mapping on KEGG pathways for genes showing an association between changes in their copy number and expression level highlighted their involvement in processes related to the cell cycle.

Loo et al\(^ {25}\) reported a correlation between gene expression and CNVs in 23% (356/1573) of the differentially expressed genes they analyzed. The most significant correlation between genomic alterations and changes in gene expression was found on chromosome 20q (20q11-13), where several overexpressed genes (AURKA, AHCY, POFUT1, RPN2, TH1L, and PRPF6) were amplified. However, an opposite pattern was detected on 8p, where the tumor suppressor genes MTUS1 and PPP2CB are located. Moreover, several of the identified genes have known involvement in the Wnt signaling pathway, which plays a role in CRC progression.

Yoshida et al\(^ {26}\) integrated the results of CNV and gene expression studies with the clinical characteristics of CRC cases they had analyzed. Their comparison of tumor and normal tissue samples revealed copy number gains on chromosomes 7, 8q, 13, and 20q, and losses on chromosomes 8p, 17p, and 18. A further analysis based on tumor stage revealed that UGT2B28 was downregulated and lost on chromosome 4q13.2 during the early stages of CRC (T1-2), while LOC440995 (3q29), CXCL6, and CXCL3 (both on chromosome 4q21), and SULTIB1 (4q13.3) distinguished T1-2-3 cases from T+ cases. Furthermore, copy numbers of RALBP1, TYMS, RAB12, and RNMT, (all mapping to chromosome 18p11) were higher in lymph node-negative cases, ARHGID1 (12p12) was absent in metastatic cases, while S1000A2 (1q21), ABHD2 (15q26), OIT3 (10q22), and ABHD12 (20p11) showed different copy numbers in cases of recurrent vs non-recurrent disease. Many of the genes identified were known to be associated with CRC, suggesting the validity of using data integration as a strategy for identifying new biomarkers for CRC or targets for its treatment.

Kikuchi et al\(^ {27}\) used a combination of copy number and mRNA expression data to demonstrate the clinical relevance of cancer-related gene protein expression. Those investigators searched for potential therapeutic targets or clinical biomarkers for advanced CRC, and then integrated their results with CNV and gene expression data obtained from a subgroup of patients with distant metastases. They found that 51 genes had both an elevated copy number and expression level. Among the three most highly expressed genes (NUCKS1, SYNPR, and TMEM63A), NUCKS1 (1q.32) is known to be overexpressed in several cancer types\(^ {28}\). The investigators found that NUCKS1 protein levels were upregulated in patients with distant metastasis, and associated with an invasive and metastatic tumor phenotype. Such findings suggest NUCKS1 as a potential biomarker for predicting CRC recurrence following colorectal surgery, and a novel target for CRC treatment.

Reid et al\(^ {28}\) identified 412 genes whose expression was correlated with CNVs, among which 80% and 20% mapped in gained and lost regions of chromosomes, respectively. Chromosomal arms 20q and 13q contained the highest numbers of genes whose expression correlated with copy number (182 and 118, respectively). Newly identified and possible CRC-related genes found in that study were PLCG1 on 20q, DBC1 on 8q21, and NDGR1 on 8p24.

Those investigators also analyzed combined data regarding the correlation between CNVs and gene expression, mutations of APC, KRAS, and TP53, 18q
Table 1 Relevant studies that integrated copy number variations and gene expression data when studying colorectal cancer

<table>
<thead>
<tr>
<th>Ref.</th>
<th>Case series</th>
<th>CNVs</th>
<th>mRNA</th>
<th>Chr 20q (gain)</th>
<th>Chr 8p (loss)</th>
<th>Additional genes/chromosomes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ali Hassan et al[26]</td>
<td>64 couples</td>
<td>15/64 couples</td>
<td>20q12</td>
<td>PTPRT1&lt;sup&gt;2&lt;/sup&gt;</td>
<td>8p22.2</td>
<td>Gains: 8q21-22, 8q24, 13q21, 13q34, 20p11-13, 20q11-13</td>
</tr>
<tr>
<td></td>
<td>(4 A, 33 B, 27 C)</td>
<td></td>
<td>CHD4</td>
<td></td>
<td>CSMD1</td>
<td>DLC1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>EMLIN3</td>
<td></td>
<td>EXT1</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>LPIN3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>PLCG1&lt;sup&gt;1&lt;/sup&gt;</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>TOP1&lt;sup&gt;1&lt;/sup&gt;</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>ZHAX3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>MAPB</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Loo et al[25]</td>
<td>40 couples</td>
<td>40 couples</td>
<td>20q11-13</td>
<td>AHCY</td>
<td>8p</td>
<td>Gains: 20pter-p12 (PCNA), 20p13 (CDC25B), 13q34 (CUL4A)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>POFG1</td>
<td></td>
<td>FBX025</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>TP1</td>
<td></td>
<td>MTUS1&lt;sup&gt;1&lt;/sup&gt;</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>THIL</td>
<td></td>
<td>EPHX2</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>PPRF16</td>
<td></td>
<td>KIF13B</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>AURKA</td>
<td></td>
<td>PPP2CB</td>
<td></td>
</tr>
<tr>
<td>Yoshida et al[24]</td>
<td>70 couples</td>
<td>70 couples</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kikuchi et al[27]</td>
<td>122 couples (18 1, 42 1, 37 2, 25 3)</td>
<td>115/122 couples (16 1, 41 1, 35 3, 23 4)</td>
<td>20q11-13</td>
<td>PLAG1</td>
<td>8p12-24</td>
<td>Gains: 1q21 (S100A2), 18p11 (RALBP1, TMY5), 15q26 (ABHD2), 20p11 (ABHD12)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>ADRM1</td>
<td></td>
<td>BIM</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>JAG1</td>
<td></td>
<td>DRC1</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>AURKA</td>
<td></td>
<td>TNFRSF10A</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>C20ORF20</td>
<td></td>
<td>TNFRSF10B</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>C20orf24</td>
<td></td>
<td>EXTL3</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>THIL</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>AHCY</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>TGF2</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<sup>1</sup>Duke’s; <sup>2</sup>CRC-related; <sup>3</sup>Stage. Chr: Chromosome; CNV: Copy number variations; CRC: Colorectal cancer; Couple: Tumor tissue matched with its normal colorectal mucosa. Genes identified in more than one study are shown in bold font.

loss of heterozygosity, and patient survival. Their results showed that chromosomal losses were frequent in wild-type TP53 patients, while TP53-mutant patients showed pronounced gains on chromosome 20q. Chromosomal alterations were rarely present in TP53-mutant cases without a 20q gain. These findings suggest that CRC can develop via two alternative routes: one mainly involving CIN, and the other involving the combined effects of having mutant TP53 plus a 20q gain. The simultaneous presence of TP53 mutations and a 20q gain might be sufficient to deregulate specific molecular pathways responsible for CRC progression. Finally, 34 genes located on chromosomes 7p, 8p, 13q, 18q, and 20 were found to be associated with overall survival.

Results from the above-mentioned studies support the hypothesis that genomic aberrations that result in CNVs lead to the deregulation of normal gene expression and directly affect critical cellular functions related to CRC tumorigenesis. Therefore, integrative approaches that utilize both gene CNV and expression data may enable the identification of markers for early detection of cancer and favor the development of new molecular agents for chemoprevention and chemotherapy.

INTEGRATED ANALYSIS OF miRNA AND mRNA EXPRESSION DATA

Numerous miRNA microarray profiling studies[29-31] and next-generation sequencing[32-34] analyses have identified miRNAs that are differentially expressed in samples of CRC tissue and adjacent non-cancerous tissue. Specialized bioinformatic tools that predict the complementarity between a miRNA seed sequence and the 3′ untranslated region of its target mRNA can only provide information concerning possible miRNA/mRNA interactions; thus identify miRNA/mRNA pairs with opposite expression patterns. The predicted interaction between a miRNA and its target mRNA can only be used to combine these data with those obtained from mRNA expression profiling studies, and thus identify miRNA/mRNA pairs with opposite expression patterns. The predicted interaction between a miRNA and its target mRNA can only be used to combine these data with those obtained from mRNA expression profiling studies, and thus identify miRNA/mRNA pairs with opposite expression patterns. The predicted interaction between a miRNA and its target mRNA can only be used to combine these data with those obtained from mRNA expression profiling studies, and thus identify miRNA/mRNA pairs with opposite expression patterns.
clinical samples. The regulatory miRNA/mRNA pairs vary in different diseases, and their expression profiles can vary with the stage of a disease. Currently, there is no efficient method for stratifying miRNA/mRNA interactions that reflects the clinical characteristics of a tissue specimen. However, the integration of molecular and bioinformatic tools represents a promising approach for fully understanding the miRNA regulatory mechanisms that underlie CRC development.

As summarized in Table 2, Fu et al.\cite{39} and Vishnubalaji et al.\cite{40} applied genome-wide mRNA and miRNA microarray expression profiling techniques to the same samples for purposes of indentifying CRC-specific miRNA/gene pairs with potential diagnostic, prognostic, or therapeutic roles. Pizzini et al.\cite{41} performed similar studies with samples of metastatic tissue to investigate how miRNA/mRNA changes affected CRC progression. Finally, Lanza et al.\cite{42} focused on base-pair differences between specific CRC subgroups (MSI and MSS).

Fu et al.\cite{39} identified 72 predicted miRNA/mRNA pairs and found that a large number of genes were participants in the Wnt signaling pathway, which is crucial for the initiation of CRC development. The most relevant pairs were validated in a study conducted using 40 additional, matched CRC tissue samples. The highest negative correlation was found between miR-224/SFRP2 and miR-29a/KLF4, which provided new information useful for elucidating CRC tumorigenesis. Moreover, Wang et al.\cite{43} analyzed those data using a multiple linear regression model and identified three additional pairs: miR-16/BCL2, miR-567/SMAD4, and miR-142-5p/MSH6.

In an attempt to find putative tumor suppressor miRNAs, Vishnubalaji et al.\cite{40} identified 794 pairs that showed opposite expression patterns between normal and tumor samples. Those studies focused on the interactions between EZH2, a gene frequently overexpressed in cancer, and its negatively correlated miRNAs (miR-26a-5p and let-7b-5p). Pharmacologic inhibition of EZH2 in CRC cell lines was found to markedly reduce both cell proliferation and migration, while in vitro silencing of EZH2 and overexpression of both miR-26a-5p and let-7b-5p decreased cell viability. Those study results suggest that miR-26a-5p and let-7b-5p play prominent roles in regulating EZH2 expression in CRC.

Pizzini et al.\cite{41} investigated changes in miRNA and mRNA expression in samples of normal colonic mucosa, primary CRC tissue, and liver tumor metastases, and found that 95% of miRNAs and 93% of genes that were deregulated in CRC samples when compared to samples of matched, normal tissue remained invariant after metastasis had occurred. Only five miRNAs (miR-146a, miR-15a, miR-15b, miR-196a, and miR-708) were deregulated during the "tumor-to-metastasis" transition period. The data regarding miRNAs and genes with opposite expression patterns were integrated to define putative post-transcriptional regulatory networks. The tumor vs normal network comparison included two components of six upregulated and 17 downregulated miRNAs, together with

### Table 2. Studies that integrated gene and miRNA expression data when studying colorectal cancer

<table>
<thead>
<tr>
<th>Ref.</th>
<th>Case series</th>
<th>Modulated mRNA genes</th>
<th>Modulated miRNA</th>
<th>miRNA/mRNA pairs with opposite expression</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fu et al.\cite{39}</td>
<td>8 couples (1, 4, 8, 38)</td>
<td>8 couples (1, 4, 8, 38)</td>
<td>2196 N vs T</td>
<td>72 miRNA/mRNA</td>
</tr>
<tr>
<td>Vishnubalaji et al.\cite{40}</td>
<td>13 couples (5, 8, 38)</td>
<td>13 couples (5, 8, 38)</td>
<td>3175 N vs T</td>
<td>794 miRNA (downregulated)/miRNA</td>
</tr>
<tr>
<td>Pizzini et al.\cite{41}</td>
<td>80 samples (23 NOR, 30 CRC, 27 liver metastasis)</td>
<td>78 samples (23 NOR, 31 CRC, 24 liver metastasis)</td>
<td>12748 N vs T</td>
<td>3078 miRNA/mRNA</td>
</tr>
<tr>
<td>Lanza et al.\cite{42}</td>
<td>39 couples (23 MSS, 16 MSI-H)</td>
<td>39 couples (23 MSS, 16 MSI-H)</td>
<td>72 MSI-H vs MSS colon cancers</td>
<td>miR182/ENTPD5, miR-145/c-Myc</td>
</tr>
<tr>
<td>Gattolliat et al.\cite{43}</td>
<td>9 NOR, 37 CRA, 9 CRC</td>
<td>5 NOR, 28 CRA, 15 CRC</td>
<td>7629 N vs T</td>
<td>70 N vs T</td>
</tr>
<tr>
<td>Reid et al.\cite{44}</td>
<td>7 public gene expression datasets CRC vs N samples (10 A, 10 B, 10 C, 10 D)</td>
<td>40 couples (10 A, 10 B, 10 C, 10 D)</td>
<td>7629 N vs T</td>
<td>70 N vs T</td>
</tr>
<tr>
<td>Ling et al.\cite{45}</td>
<td>SW480 cells after miR-224 overexpression (84 metastasis related genes analyzed)</td>
<td>4 CRC with metastasis, 8 CRC without metastasis, SW480, SW620</td>
<td>13 metastasis-related genes down-regulated</td>
<td>miR-224/SMAD4, miR-224/CDH1</td>
</tr>
</tbody>
</table>

<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

1Stage. Couple: Tumor tissue matched with its normal colorectal mucosa; CRA: Colorectal adenoma; CRC: Colorectal cancer; NOR: Normal colorectal mucosa.
their putative target genes. In comparison, a network constructed using the five miRNAs identified in the metastasis vs tumor comparison plus their target genes was smaller, and contained five unrelated components. The results of that study suggest that significant changes in the transcriptome mostly occur during the early stages of CRC progression. Moreover, opposite expression patterns for miR-145 and its target gene MYC were confirmed, and a proposed interaction between miR-182 and ENTPDS, a gene involved in energy metabolism, was functionally validated. Finally, an early survival analysis indicated that miR-10b expression (modulated between tumor and metastatic tissues) was inversely correlated with survival in stage IV CRC patients.

Lanza et al.[42] added miRNA and mRNA profiling data to molecular signature data that could distinguish between MSI-H and MSS CRCs. Those investigators then used prediction algorithms that combined information regarding the 14 differentially expressed miRNAs and 72 deregulated genes identified in MSI-H and MSS-type tumors to construct a combinatoric predictor containing 27 elements (miRNAs and genes) that was more sensitive than the predictor lists that contained only one type of element. The addition of miRNAs to the molecular predictor improved their categorization and represented a novel approach for elucidating the respective molecular features of the two CRC subgroups. The combinatoric predictor includes various members of the miR-17-92 family, which is a class of miRNAs with proven oncogenic characteristics that are probably involved in the molecular mechanisms that distinguish MSS from MSI colon cancers.

Several studies have combined global mRNA and miRNA microarray expression data derived from two independent cohorts of patients, and in some cases also integrated publically available datasets (Table 2). Gattolliat et al.[43] combined miRNA and gene expression results obtained from two independent sets of normal mucosa (NOR), colorectal adenoma (CRA), and CRC tissue samples. They found different levels of miR-320b expression in CRC tissue when compared to NOR tissue, as well as differences in miR-15b and miR-16 expression in CRA tissue vs NOR tissue. Their data also showed that miR-21, miR-24, miR-145, miR-150, and miR-378 were deregulated in samples of both CRA and CRC tissue when compared to samples of NOR tissue. When the expression of these miRNAs were compared with expression of the genes predicted as their putative targets, 30 pairs with opposite expression patterns were identified, including PDCC4 and MARCKS for miR-21, ZEB2 for miR-200b, and BCL2 for miR-15b, miR-16, and miR-21.

Reid et al.[45] identified 23 miRNAs that were differentially expressed in matched samples of CRC and normal tissue, and then searched seven public gene expression datasets, plus an independent cohort of patients from the same group, to identify putative target genes of these miRNAs. The selected pairs were mapped on the KEGG pathway, and many were found to be included in CRC-related molecular pathways. One candidate pair (miR-1 and the MET oncogene) was functionally validated in studies that used in vitro models of CRC. miR-1 overexpression was found to diminish MET levels and result in reduced cell proliferation, migration, and invasion, suggesting a prominent role for miR-1 in CRC progression. As miRNA expression can also be affected by genomic alterations in regions in which they are positioned, the miRNA profiles were integrated with data obtained from a whole-genome copy-number analysis[18]. The results showed that chromosomal regions that frequently gained copy numbers contained upregulated miRNAs, whereas regions which lost copy numbers contained both up- and downregulated miRNAs. The role of miR-20a, which was overexpressed and localized in an amplified region, was functionally investigated in CRC cell lines, where studies showed that it interfered with transforming growth factor-β-induced growth arrest[46].

Ling et al.[47] conducted whole genome miRNA expression studies using samples of primary CRCs with or without metastasis, and also two cell lines: one derived from a primary CRC lesion (SW480), and the other derived from a metastasis located in a lymph node (SW620). The results were validated in a large international patient cohort, and also by using data obtained from TCGA.[47] The investigators identified four miRNAs (miR-141, miR-181b, miR-221, and miR-224) that were upregulated in primary CRCs with metastatic dissemination at the early stages. The highest levels of expression were observed in the metastasis-related cell line SW620. Only miR-224 overexpression induced the migration of CRC cells, and expression of metastasis-related genes was analyzed following their insertion into SW480 cells. Two target genes (CDH1 and SMAD4) showed reduced expression, and a functional analysis demonstrated that SMAD4 mediated the miR-224-induced prometastatic effect. Furthermore, elevated miR-224 expression was shown to correlate with survival in CRC patients, indicating a prominent role for miR-224 as a specific diagnostic marker for CRC.

While all of the above-mentioned studies identified several CRC-related miRNAs with potential value as diagnostic or prognostic markers, and/or therapeutic targets, only a few of the identified miRNA/gene target pairs have been functionally analyzed. This limits our understanding of their role in CRC, as the effects of miRNA deregulation are complex and impact the modulation of entire pathways rather than just a single gene.
INTEGRATION OF METHYLATION AND GENE EXPRESSION PROFILES

Hypermethylation of DNA segments located in promoter CpG islands is the most studied epigenetic alteration involved in the transcriptional repression commonly observed in various cancer types\[^{49-51}\]. In CRC, aberrant DNA methylation in CpG islands occurs during the early stages of an oncogenic transformation process, and can be detected in aberrant crypt foci, which are the earliest detectable oncogenic changes in colonic mucosa\[^{49-51}\]. The methylation of DNA in several genes, including APC, p16INK4a, and TIMP3, and its significance in CRC have previously been reported\[^{52,53}\], and represents a potential biomarker for use in the early detection of CRC\[^{54}\]. The development of array and sequencing-based high-throughput assay techniques now permits the profiling of genome-scale DNA methylation (methylomes), and within the same tumor type, has enabled the characterization of different subgroups that display heterogeneous DNA methylation\[^{55}\]. Integration of these data with analyses of gene expression can assist in identifying new candidate diagnostic biomarkers that become methylated during the early stages of oncogenesis.

Three recent studies (Table 3) have provided new insights into the role of the CpG island hypermethylation in the regulation of gene expression. The first study integrated gene expression and methylation data obtained from studies that used tissue samples from the same CRC cases, even if only a small number of matched normal and tumor tissue samples were available\[^{56}\]. Szmidza et al\[^{57}\] and Wang et al\[^{58}\] used these data to implement their methylation analyses, with the purpose of validating the results they obtained by integrating methylation and gene expression data retrieved from the TCGA database. All of their studies identified genes whose expression was affected by DNA hypermethylation. Furthermore, two of the studies identified genes involved in the same pathway (ErbB-signaling pathway), and thus highlighted the pivotal role played by DNA methylation in CRC development.

Hinoue et al\[^{56}\] performed a model-based cluster analysis to identify four distinct methylation-based subgroups of CRC patients with specific genetic and clinical features. They found that the CIMP was correlated with a high frequency of cancer-specific DNA hypermethylation (CIMP-H), a high incidence of the BRAF\[^{V600E}\] mutation, and high rates of KRAS mutation in a CIMP-low subgroup. Next, the non-CIMP tumors were separated into two subgroups: one with a high frequency of TP53 mutations in the distal colon, and the other with low incidences of gene mutation and cancer-specific DNA hypermethylation, but significantly enriched for rectal tumors. Those investigators also identified a panel of five genes that could specifically distinguish CIMP from non-CIMP tumors, and also another panel that was highly specific for CIMP-H tumors. Gene expression profiling studies revealed that about 7% of the promoter DNA methylations observed in CIMP-H tumors were linked to the downregulation of 112 genes. Those downregulated genes represented 25% of the genes with lower expression in CIMP-H tumors compared with their expression in adjacent normal tissue. Intriguingly, twelve genes were also downregulated and hypermethylated in non-CIMP tumors. This result is interesting because SFRP1 and SFRP2 are negative regulators of Wnt signaling.

Szmidza et al\[^{57}\] integrated their previously published genome-wide methylation data\[^{59}\] with results from the gene expression and methylation profiling studies conducted by Hinoue et al\[^{56}\], and identified four ErbB-associated genes (PIK3CD, PKCB, ERBB4, and PAK7) that were differentially methylated in CRC. In particular, PKCB hypermethylation was correlated with the presence of a KRAS mutation, and hypermethylation of ERBB4 was linked with highly methylated epigenotypes HME and MSI with the presence of mutated BRAF. Methylation appeared to only impact the modulation of PKCB expression that was significantly downregulated in CRCs following methylation of its promoter. PKCB is a component of the vascular endothelial growth factor signaling pathway and regulates cell proliferation and survival processes that promote tumor angiogenesis. Indeed, therapies that target the vascular endothelial growth factor pathway are currently in clinical studies for treatment of late-stage CRC\[^{60}\].

Wang et al\[^{58}\] combined the gene methylation and expression profiles typical of CRC as retrieved from the

### Table 3  Relevant studies that integrated methylation and gene expression profiles

<table>
<thead>
<tr>
<th>Ref.</th>
<th>Methylation</th>
<th>mRNA</th>
<th>Case series</th>
<th>Genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hinoue et al[^{56}]</td>
<td>29 NOR, 125 CRC</td>
<td>19 couples</td>
<td>464 genes downregulated CIMP-H in normal samples, 112 of them (24%) exhibit promoter hypermethylation</td>
<td></td>
</tr>
<tr>
<td>Szmidza et al[^{57}]</td>
<td>12 couples</td>
<td>19 couples (from Hinoue et al[^{56}], 2012)</td>
<td>12 genes downregulated and hypermethylated in non-CIMP tumors</td>
<td></td>
</tr>
<tr>
<td>Wang et al[^{58}]</td>
<td>42 NOR, 231 CRC</td>
<td>26 NOR, 231 CRC</td>
<td>4 ErbB-associated genes (PIK3CD, PKCB, ERBB4, PAK7) differentially methylated in CRC</td>
<td></td>
</tr>
</tbody>
</table>

Couple: Tumor tissue matched with its normal colorectal mucosa; CRC: Colorectal cancer; NOR: Normal colorectal mucosa.
TCGA database and identified highly variable DNA methylation sites, as well as genes whose expression was affected by a tumor’s highly variable DNA methylation status, that were named methylation-perturbed genes (MP). Those results showed that the number of MP genes was significantly lower in samples of CRC tissue as compared to normal tissue. The genes were then clustered based on connectivity between their expression levels and subgrouped according to their MP status. The number of coexpressed partners of MP genes was significantly lower in samples of CRC tissue when compared to samples of normal tissue, and also when compared to the number of non-MP genes in both CRC and normal samples. Interestingly, the lost coexpression partners were often members of cancer pathways, such as the ErbB and mitogen-activated protein kinase signaling pathway. The loss of coexpression connectivity mediated by methylation heterogeneity as described in this study might play an important role in CRC development.

Similar to the observations regarding protein-coding genes, miRNAs can also be silenced by hypermethylation of a CpG island promoter region. Sometimes this can occur indirectly when miRNAs are transcribed from intronic regions of coding genes that are controlled by hypermethylation. While no comprehensive analysis and integration of miRNA and methylation data obtained from studies with CRC tissue have yet been performed, several methylation-sensitive miRNAs including miR-9, miR-129, and miR-137, miR-34b/c, miR-200, miR-342, and miR-345 have been identified.

DATA INTEGRATION BY COMBINING PUBLIC DATASETS: THE TCGA NETWORK

The information most valuable for better understanding CRC development and progression can be derived from studies in which tissue samples are characterized for various molecular changes that define the disease. However, most research groups who have combined genetic and genomic data conducted a maximum of two whole genome analyses on the same samples. Furthermore, the experiments were frequently implemented based on data retrieved from publically available datasets containing information regarding CNVs, gene and miRNA expression, and methylation (e.g., GEO and ArrayExpress). The TCGA research program has been of great benefit to investigators seeking to combine multiple data types in hopes of better understanding disease processes. TCGA was launched by the National Institutes of Health in 2006 for the purpose of comprehensively characterizing the genomic and molecular features of cancer by using high-throughput genome, transcriptome, and epigenome analysis techniques. These techniques include gene and miRNA expression profiling, CNV and genome-wide DNA methylation profiling, single-nucleotide polymorphism genotyping, and exon sequencing performed on thousands of samples obtained from about 20 different tumor types. Numerous robust studies have been conducted using different types of data retrieved from TCGA, and several projects have also been conducted by the TCGA research program itself on the different cancer types analyzed. As for CRC, an analysis of 224 paired tumor and normal tissue samples performed using different platforms (exome sequencing, DNA copy number, promoter methylation, miRNA and microRNA expression, and whole-genome sequencing for 97 samples) identified several specific characteristics of this tumor type. The CRCs were divided into categories of hypermutated (16%) and non-hypermutated (84%), and 24 genes in each category were found to be significantly mutated. It was interesting that, although colon and rectal cancers have different characteristics, they showed similar patterns of genomic alterations. An analysis of CNVs revealed amplifications of ERBB2 and IGF2 and fusion of NAV2 with TCF7L1, which is a member of the Wnt pathway. An integration of CNV, expression, and methylation data revealed that all cancers showed changes in genes known to be involved in MYC transcription, suggesting an important role for MYC in CRC. Numerous molecular signatures were linked to tumor aggressiveness, and in particular, two chromosomal regions (20q13.12 and 22q12.3) that showed amplifications of genes linked to tumor aggression. To better interpret the effects of genomic abnormalities in terms of cancer biology, genomic data have also been integrated with proteomic data obtained from the same tumors analyzed by TCGA, but generated by the Clinical Proteomic Tumor Analysis Consortium. This combined proteomic and genomic data can be used to link genotypes with phenotypes and assist in prioritizing genes that merit further examination. Results from examinations of CRC tissue revealed a weak correlation between levels of mRNA and proteins produced by genes in CNV regions, and only the 20q amplification was found to be associated with the largest global changes in both mRNA and protein levels. The identified candidate genes were HINFA4A, which codes for a transcription factor that plays a key role in normal gastrointestinal development, TOMM34, which is involved in the growth of CRC cells, and SRC, which encodes a non-receptor tyrosine kinase implicated in several human cancers, including CRC.

TCGA and Clinical Proteomic Tumor Analysis Consortium data are available to the scientific community, and several research groups are currently using TCGA data for their studies. Additionally, some investigators have expanded their own datasets or combined them with datasets for different categories of genetic aberrations. Other investigators have

Sokolova V et al. CRC biomarkers combining genome and transcriptome

WJG | www.wjgnet.com 541 January 14, 2016 | Volume 22 | Issue 2 |
utilized specific analytical methods to obtain more in-depth analyses that have enabled the identification of additional genes associated with specific features of CRCs. For example, Ashktorab et al[54] and Zhu et al[50] used the elastic-net regression method to perform a supervised analysis that integrated multiple types of genomic data, and then compared the results with the clinical stage of CRC to identify genes associated with advanced CRC. They found that the tumor suppressor gene WRN exhibited the highest number of genomic variations (CNVs, expression changes, and methylation) that could be used to delineate advanced CRC.

More simple and immediate uses of these data include the development of portals that collect all CRC-related -omics data, and tools that facilitate data visualization or allow an investigator to perform queries and integrate different types of information[59].

Finally, RNA-Seq and miRNA-Seq profiles obtained from TCGA project have been used to identify numerous non-coding RNAs (miRNAs, long non-coding RNAs, and competitive endogenous RNAs) in multiple types of cancer. These RNAs are capable of binding to each other, to mRNA, or even to proteins, and regulate their expression. These capabilities make them promising sources of possible new cancer biomarkers or targets that can be used to study cancer development[72]. Similar to what has been done using genetic and genomic data, several groups have recently began to integrate non-coding RNA, RNA, and protein data obtained from TCGA for the purpose of constructing experimentally supported networks of RNA-RNA and protein-RNA interactions that may become deregulated in different types of cancers[72]. Although these data are derived from 14 different cancer types, they represent an additional source of information that can be accessed to elucidate disease processes and identify new targets for the treatment of CRC.

CONCLUSION
The development of new high-throughput screening and sequencing technologies has significantly increased the amount of information available concerning the cancer genome and transcriptome. A combinative approach that integrates genetic and genomic data obtained from the same samples provides a more realistic view of the biologic system being analyzed and assists in identifying new therapeutic targets and disease biomarkers. In fact, it can compensate for missing or unreliable information regarding any single data type and better explain the genetic complexities and basic biologic pathways involved in cancer. Additionally, identification of the same gene or biologic pathway in various datasets, and also in different studies, supports its involvement in a disease. This was observed for the four genes that were identified after integrating CNV and gene expression data obtained from three different studies, PLCG1[18,24], AHCY, TH1L, and AURKA[18,24,29] all mapping to 20q11-13. Amplification of the 20q11-13 region has been linked to processes that facilitate the progression of adenoma to carcinoma[73], AURKA, which has previously been associated with gains in 20q11-13, is known to affect cell migration[74] and may synergistically act with the other three genes (PLCG1, AHCY, and TH1L) to promote CRC progression.

As for miRNA and epigenetic profiles, their integration with gene expression data provides a more comprehensive picture of the regulatory networks involved in cancer, and has confirmed that the Wnt and ErbB pathways play major roles in CRC development.

In summary, the integrative approaches described in this review will ultimately provide investigators and physicians with a more accurate and detailed picture of the complex molecular characteristics of cancers. Furthermore, they should provide new insights that will allow us to better predict cancer, develop a prognosis for cancer patients, and identify new treatments.

REFERENCES


Target Scan Human. Available from: URL: http://www.targetscan.org/


de Wiel MA, van Criekinge W, Thas O, Matthäi A, Cuesta MA, Terhaar Sive Droste JS, Craanen M, Schröck E, Ylstra B, Meijer GA. Multiple putative oncogenes at the chromosome 20q amplicon contribute to colorectal adenoma to carcinoma progression. Gut 2009; 58: 79-89 [PMID: 18829976 DOI: 10.1136/gut.2007.143065]


P- Reviewer: Huang ZH, Meshikhes AW
S- Editor: Qi Y
L- Editor: A
E- Editor: Liu XM