IDENTIFICATION OF MICRORNA NETWORKS ASSOCIATED WITH TUMOR PROGRESSION IN COLORECTAL CANCER

PORNPIMOL CHAROENTONG

DOCTORAL THESIS

Institute of Genomics and Bioinformatics
Graz University of Technology
Petersgasse 14, 8010 Graz, Austria

Graz, November, 2010

EIDESSTATTLICHE ERKLÄRUNG

Ich erkläre an Eides statt, dass ich die vorliegende Arbeit selbstständig verfasst, andere als die angegebenen Quellen/Hilfsmittel nicht benutzt und die den benutzten Quellen wörtlich und inhaltlich entnommenen Stellen als solche kenntlich gemacht habe.

Graz, am …… ……..
(Unterschrift)

Englische Fassung:

STATUTORY DECLARATION

I declare that I have authored this thesis independently, that I have not used other than the declared sources / resources and that I have explicitly marked all material which has been quoted either literally or by content from the used sources.

…… ……..
date (signature)

“True success is not in the learning, but in its application to the benefit of mankind”

H.R.H. Prince Mahidol of Songkla

Abstract

Background: MicroRNAs (miRNAs) are a class of small non-coding RNAs that regulate gene expression at the post-transcriptional level. Recent studies suggest that miRNAs are involved in the initiation and progression of many cancers types accompanied by changes in the immune system. We performed a computational-experimental study to identify miRNA signatures and high confidence miRNA target genes driving tumor progression in colorectal cancer (CRC).

Results: We measured the expression of 365 microRNAs and selected genes for samples from 103 and 125 CRC patients, respectively and performed microarray analyses for samples from 105 patients. We could identify miRNAs and high confidence targets showing a strong association with a higher UICC stage compared to normal colon mucosa. Our results demonstrated in different tumor stages a significant negative correlation between 4 miRNAs and Fractalkine (CX3CL1) expression, an activator of T-cell function previously shown to be a good prognostic factor for CRC. Genomic deletion events detected by array-CGH data provide a mechanism by which that miRNA could be involved in the progression of the disease. Finally we selected high-confidence miRNA target genes to reconstruct miRNA-mRNA networks and pinpointed immune processes controlling tumor progression.

Conclusion: Our results suggest that miRNAs and their high-confidence targets may have a functional effect on tumor progression. Furthermore, some miRNAs with prognostic potential could provide the basis for an in-depth study of certain miRNAs as clinical prospective markers and as new pharmaceutical targets.

Keywords: MicroRNA, Colorectal Cancer, High Confidence Target

Zusammenfassung

Hintergrund: MicroRNAs gehören einer Klasse von kleinen nicht-kodierenden RNAs an, die die Genexpression auf post-transkriptionelle Weise regulieren können. Laut neuesten Studien sind MicroRNAs im Entstehen von verschiedenen Krebsarten und den dazugehörigen Änderungen im Immunsystem involviert. Wir verfolgten einen kombinierten Ansatz um MicroRNAs und deren Zielgene zu identifizieren, welche am Fortschreiten von Dickdarmkrebs beteiligt sind.

Ergebnisse: Wir haben für Tumorproben von 103 Patienten die Expression von 365 MicroRNAs und von 125 Patienten die Expression ausgewählter Gene mittels qPCR bestimmt sowie für Proben von 105 Patienten Microarray-Experimente durchgeführt. Wir konnten eine Reihe von MicroRNAs und Zielgenen identifizieren, die einen Zusammenhang in fortgeschrittenen Stadien von Dickdarmkrebs im Vergleich zu normaler Dickdarmschleimhaut aufwiesen. Die Ergebnisse zeigten für unterschiedliche Tumorstadien eine signifikante negative Korrelation zwischen 4 MicroRNAs und der Expression von Fraktalkine (CX3CL1), einem Aktivator der T-Zellfunktion und guten prognostischen Marker von Dickdarmkrebs. Mittels array-CGH detektierter genomischer Ereignisse (Deletionen) könnten erklären warum eine dieser MicroRNAs direkt im Entstehen von Dickdarmkrebs beteiligt ist. Schließlich wurden noch MicroRNA-Zielgene ausgewählt sowie MicroRNA-Gen-Netzwerke konstruiert, die es ermöglichen Immunprozesse die im Krebsverlauf eine Rolle spielen aufzuzeigen.

Schlussfolgerung: Unsere Ergebnisse zeigen, dass MicroRNAs und deren Zielgene einen direkten Einfluss auf den Krebsverlauf haben könnten und dass einige microRNAs sich als gute prognostische Marker und pharmazeutische Targets eignen würden und als Basis für weitere eingehendere Studien dienen.

Schlüsselwörter: MicroRNAs, Dickdarmkrebs, Zielgene, Prognose

Publications

This thesis is based on the following publications as well as upon unpublished work:

1. Hackl H, Stocker G, **Charoentong P**, Mlecnik B, Bindea G, Galon J, Trajanoski Z.Information technology solutions for integration of biomolecular and clinical data in the identification of new cancer biomarkers and targets for therapy. *Pharmacol Ther*, 2010, 128(3):488-98.
2. Mlecnik B, Sanchez-Cabo F, **Charoentong P**, Bindea G, Pagès F, Berger A,Galon J, Trajanoski Z. Data integration and exploration for the identification of molecular mechanisms in tumor-immune cells interaction. *BMC Genomics*, 2010, 10;11 Suppl 1:S7.
3. Mlecnik B, Tosolini M, **Charoentong P**, Kirilovsky A, Bindea G, Berger A, Camus M, Gillard M, Bruneval P, Fridman WH, Pagès F, Trajanoski Z, Galon J. Biomolecular network reconstruction identifies T cell homing factors associated with survival in colorectal cancer. *Gastroenterology*. 2009,138(4):1429-40.
4. Bindea G, Mlecnik B, Hackl H, **Charoentong P**, Tosolini M, Kirilovsky A, Fridman WH, Pagès F, Trajanoski Z, Galon J. ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. *Bioinformatics*. 2009, 15;25(8):1091-3.
5. Camus M, Tosolini M, Mlecnik B, Pagès F, Kirilovsky A, Berger A, Costes A, Bindea G, **Charoentong P**, Bruneval P, Trajanoski Z, Fridman WH, Galon J. Coordination of intratumoral immune reaction and human colorectal cancer recurrence. *Cancer Res*. 2009, 15;69(6):2685-93.
6. **Charoentong P**, Hackl H, Mlecnik B, Bindea G, Galon J, Trajanoski Z. Bioinformatics for cancer immunology (submitted).

Contents

Background

Globally, colorectal cancer (CRC) is the third most common form of cancer and the second cause of cancer-related death in the western countries, causing 655,000 deaths around the world per year [1]. This type of cancer is more common in developed than developing countries [2]. The life time risk of colorectal adenocarcinoma is approximately 6% and of colorectal adenoma approximately 50% with an increasing age depending risk, especially after 60 years [3].

CRC develops sporadically, in the setting of hereditary cancer syndromes, or on the basis of inflammatory bowel diseases [2]. Human CRC occurs when some of the cells that line the colon or the rectum become abnormal and grow out of control. Tumors of the colon and rectum are growths arising from the inner wall of the large intestine. This type of cancer can invade and damage adjacent tissues and organs. Surgery is the one choice of offering a potential cure. However, 30-40% of patients have loco regionally advanced or metastatic disease on a presentation which cannot be cured by surgery alone [4]. In addition, more than half of patients initially believed to be cured by surgery develop recurrence and die of the disease [5]. Adjuvant therapies have improved treatment and survival in patients with advance diseases but five years survival remains at approximately 50% both of colonic and rectal tumors [6].

The staging of CRC is determined by many systems but the two most common are the Dukes staging and TNM classification. Several different forms of the Dukes classification were developed and placed patients into four stages, A, B, C and D according to the degree of the extent of cancer spread [7-9]. More recently, the American Joint Committee on Cancer (AJCC) has introduced the TNM staging system [10-12], which relies on the depth of tumor invasion and the absence or presence of nodal and distance metastasis. The objectives of both classifications are to aid the clinician in the planning of treatment, give some indication of prognosis, assist in the evaluation of the results of treatment, and facilitate the exchange of information.

Cancer is a complex disease that involves the interaction of many cell types and appears at different scales from the subcellular to macroscopic one. It is well known that tumors induce immune response. The activation of the host immune system through tumor cells is complex cascade involving both the innate and adaptive immune systems. The immune system can respond to cancer cells in two ways: by reacting against tumor-specific antigens (molecules that are unique to cancer cells) or against tumor-associated antigens (molecules that are expressed differently by cancer cells and normal cells) [13]. The concept that the immune system can recognize and eliminate malignant tumors was originally embodied in the cancer immunosurveillance hypothesis of Burnet and Thomas [14]. Cancer immunosurveillance is considered to be an important host protection process to inhibit carcinogenesis and to maintain cellular homeostasis [15]. This hypothesis was abandoned shortly afterwards because of the absence of strong experimental evidence supporting the concept [16]. Extensive work in experimental systems has elucidated the mechanisms underlying spontaneous antitumor immunity, and has formed the basis for the cancer immunoediting hypothesis. This hypothesis divides the immune response to cancer into the “three E’s” which are elimination, equilibrium, and escape [16, 17]. Several publications reported that solid cancers (ovarian, colorectal, lung, head and neck, melanoma, so on) have immunogenic properties and evidence that host immune response can influence survival. The adaptive immune reaction within the tumor appeared to be the most important parameter predicting the outcome after surgical treatment with curative intent [18, 19].

Immunotherapy offers one such strategy. The demonstration that CRC has immunogenic properties and evidence that host immune response can influence survival [20-27]. Tumor infiltrating lymphocytes (TILs) have been isolated from the variety of solid human cancers. In a study using 959 specimens of resected CRC analyzed the correlation between tumor metastasis and T cell activation that tumor infiltrating memory and effector memory T cells (T_{EM}) are less likely to discriminate to lymphovascular, perineural structures and to the regional lymph nodes [28]. Taken together, this has been attributed to a beneficial outcome, and the enhancement of T cell activation through T cell receptor stimulation and co-stimulatory signal provides promising strategies for immunotherapy of CRC [29].

However, further studies are necessary to identify immune signatures. Recent biochemical and genetic studies have revealed that a class of small non-coding RNAs called microRNAs (miRNAs) supported a role in crucial physiological and biological processes such as cell proliferation [30], apoptosis [31], development [32], differentiation [33] and metabolism [34]. Given the global effect of miRNAs on gene expression, it is not surprising that miRNAs have been implicated in a common feature of various human diseases (developmental abnormality of muscular [35, 36], cardiovascular disorders [37, 38], cancers [39-48] and most recently inflammatory diseases [45, 49, 50].

Biological characteristics of miRNAs

MiRNAs is a class of non-coding small RNAs, typically about 21–23 nucleotides long [51]. Since the first discovered miRNA in 1993, there are almost 800 miRNAs identified in human beings and the discovery of miRNA has led to great progression in the understanding of human cancers [52]. Forming mature miRNA involves transcription from DNA and two cleavages by Drosha and Dicer, the two main enzymes in the procession. The mature miRNA forms RNA-induced silencing complex (RISC) with Argonaut proteins and the target mRNA. In animal cells, mature miRNAs work through binding to the 3'-UTR of their target mRNAs with imperfect or perfect complementarily (Fig.1). MiRNAs can repress the expression of target genes either through disruption the translation or decomposition the target mRNAs [53, 54]. Currently, there are several hundred miRNAs that have been identified [55]. Although miRNA-mediated mRNA degradation occurs in mammals, most miRNAs are thought to use a second mechanism of gene regulation via imperfect base paring to the 3´-untranslated regions (3´-UTRs) of their miRNA targets. This results in the repression of target gene expression post-transcriptional, likely that the translational level of gene expression [56, 57].

MicroRNA and the Immune Response

The discovery of miRNAs as regulators of developmental events in model organisms suggested to many investigators that miRNA might be involved in the immune system. In the past few years, widespread examination of this possibility has

produced notable results. The first indication that miRNAs might regulate the immune responses was a report in 2004 showing selective expression of miR-142a, miR-181a and miR-223 in immune cells [33]. Many results have shown that miRNAs affect mammalian immune cell; regulation of maturation, proliferation, differentiation and activation of immune (Fig.2) [33, 45, 58-63].

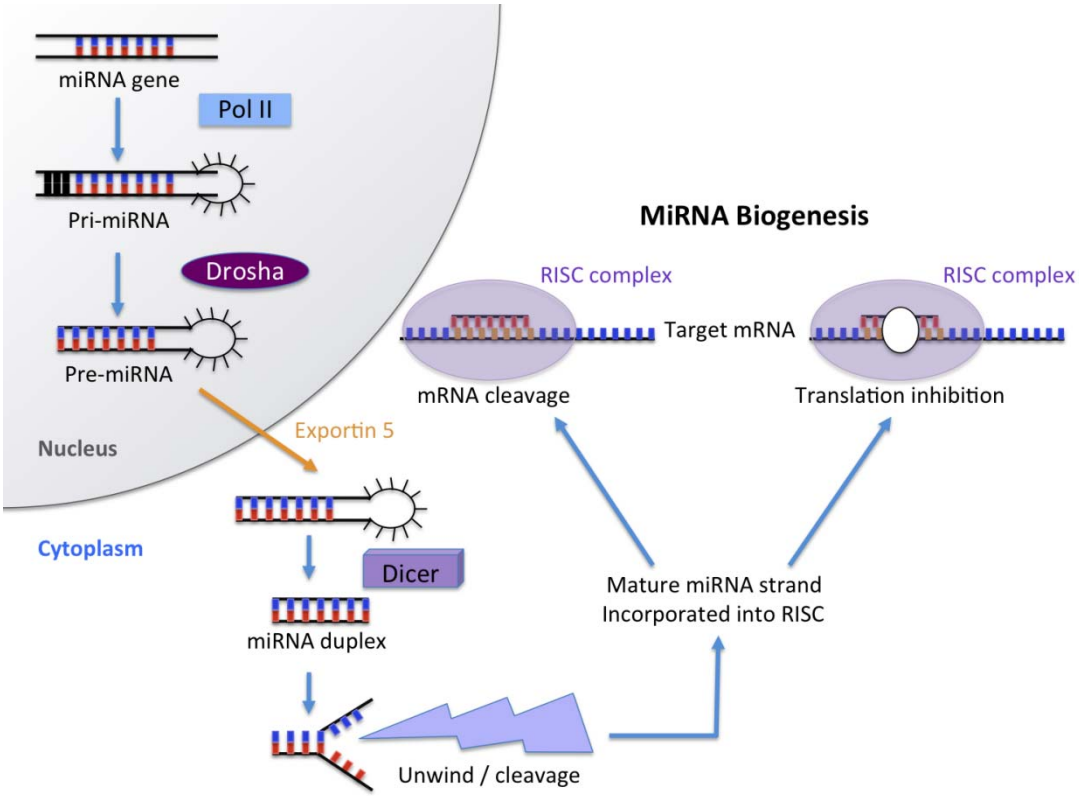


Figure 1: Pathway of microRNA biogenesis and action adopted from [64]. MiRNAs are non-coding, single-stranded RNAs of ~22 nucleotides and generally transcribed by RNA polymerase II (Pol II) into long primary miRNA transcripts of variable size (pri-miRNA), which are recognized and cleaved in the nucleus by the RNase III enzyme Drosha [65, 66] and its cofactor, DGCR8, to a pre-miRNA precursor product. The pre-miRNA is transported from the nucleus to the cytoplasm by exportin 5 [67]. Another RNase enzyme called Dicer [68] which produces a transient, 22 nucleotide duplex. Only one strand of the miRNA duplex (mature miRNA) is incorporated into a large protein complex called RISC (RNA-induced silencing complex). The mature miRNA leads RISC to cleave the mRNA or induce translational repression, depending on the degree of complementarily between the miRNA and its target.

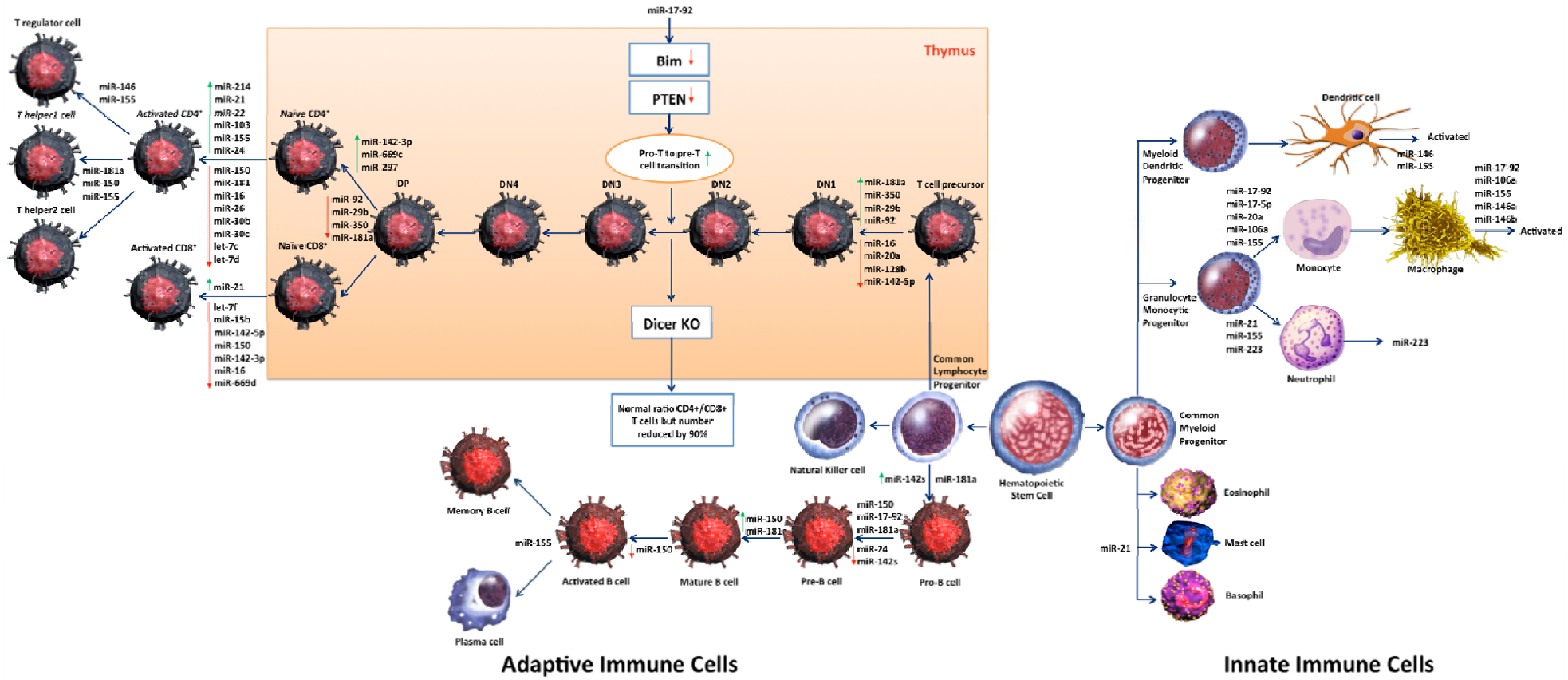


Figure 2: MiRNAs involved in the differentiation and maturation of innate and adaptive immune cells. Upregulated miRNAs are shown in green, downregulated miRNAs are shown in red.

The role of the immune system is to protect against infection and to eliminate disease from the host [69]. There are two type of the defense; innate and adaptive immunity. Most are detected and destroyed within minutes or hours by defense mechanisms that do not require a prolonged period of induction because they do not rely on the clonal expansion of antigen-specific lymphocytes: these are the mechanisms of innate immunity [70]. The specific pathogen destroying is known as an adaptive immune response. In the immune system, miRNA appear to have a key role in the early differentiation and effector differentiation of B cells [71-77]. In T cells, miRNAs have been shown to be key regulators of the lineage induction pathways, and to have a strong role in the induction, function and maintenance of the regulatory T-cell lineage [71-77]. MiRNAs are also important for regulating the differentiation of dendritic cells and macrophages via toll-like receptors, with responsibilities in suppressing effector function before activation and enhancing function after stimulation [69-81].

MicroRNA and Cancer

Human cancer studies are always the hotspots in the life science research. The possible involvement of miRNAs in cancers was based on following observations. Firstly, miRNAs play important roles in cell fate determination, proliferation, and cell death [74, 78-84]. Those are key mechanisms involved in cancer. Moreover, many genes encoding miRNAs are located in regions of the genome known to be frequently amplified or deleted in human cancers [78, 85, 79]. Finally, the expression profile of miRNAs in normal tissues and tumor tissues is different [86-90]. Significant progress in miRNAs and cancer has been made in the past few years. Two major categories of miRNAs, oncomiRs and tumor suppressor miRNAs (Fig.3 and Tab.1), have been described based on their effects in cancer [80, 91-94]. OncomiRs can act as oncogenes and have a negative impact on patient outcome [93, 94]. Their expression can be either positively influence oncogenes or inhibit tumor suppressor genes. Conversely, tumor suppressor miRNAs have a positive impact on patient outcome when high expressed [91-94]. Their expression may either inhibit oncogenes or activate tumor suppressor genes.

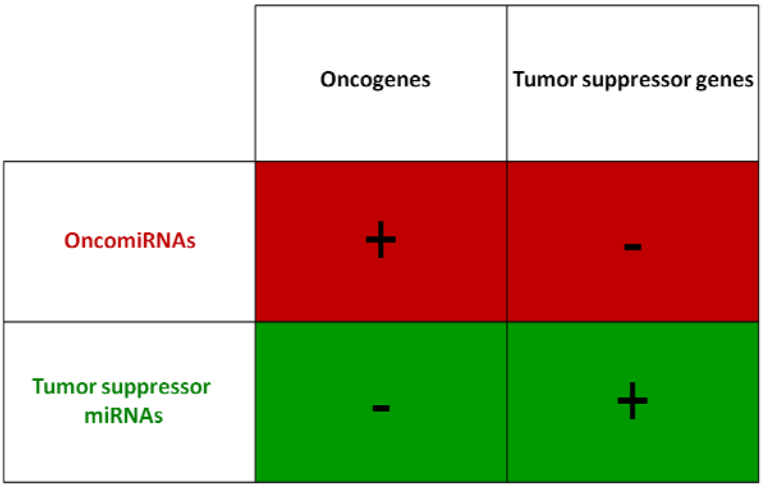


Figure 3: Effects of miRNAs on oncogenes and tumor suppressor genes. OncomiRs and tumor suppressor miRNAs impact is shown in red and green, respectively. + means promotion, - means inhibition.

Table 1: MiRNAs with experimental data supporting a tumor suppressor or oncogene function in cancer [64].

microRNA	Expression in patients	Functions
miR-15a, miR-16-1	Down-regulated in CLL	Tumor suppressor
let-7 (a,-b,-c,-d)	Down-regulated in lung and breast cancers	Tumor suppressor
miR-29 (a,-b,-c)	Down-regulated in CLL, AML (11q23), lung and breast cancers, and cholangiocarcinoma	Tumor suppressor
miR-34 (a,-b,-b)	Down-regulated in pancreatic, colon, and breast cancers	Tumor suppressor
miR-155	Up-regulated in CLL, DLBCL, FLT3-ITD AML, BL, and lung and breast cancers	Oncogene
miR-17-92 cluster	Up-regulated in lymphomas and in breast, lung, colon, stomach, and pancreas cancers	Oncogene
miR-21	Up-regulated in breast, colon, pancreas, lung, prostate, liver, and stomach cancers; AML(11q23); CLL; and glioblastoma	Oncogene
miR-372, miR373	Up-regulated in testicular tumors	Oncogene

Abbreviations: CLL, chronic lymphocytic leukemia; AML, acute myeloid leukemia; DLBCL, diffuse large B cell lymphoma; FLT3-ITD, FMS-like tyrosine kinase 3 in tandem duplication mutations; BL, Burkitt lymphoma.

Different tumors and tumor subtypes have specific miRNA signatures which may be useful as diagnostic and prognostic markers. Lu et al [95] used a new bead-based flow cytometric miRNA expression profiling method to present a systematic expression analysis of 217 mammalian miRNAs from 334 samples, including multiple human cancers. The miRNA profiles are surprisingly informative, reflecting the developmental lineage and differentiation state of the tumors. Furthermore, they were able to successfully classify poorly differentiated tumors using miRNA expression profiles, whereas messenger RNA profiles were highly inaccurate when applied to the same samples. These findings highlight the potential of miRNA profiling in cancer diagnosis. Bloomston et al reported 23 miRNAs that significantly distinguish pancreatic cancer from chronic pancreatitis, with several of these miRNAs being capable of predicting overall survival in the cancer patients [96]. In stage II colon cancer, miRNA expression profiles were capable of predicting recurrence rates with an accuracy of >80%, suggesting that miRNA profiling can also be used to determine a tumor's aggressiveness [97]. Similarly, miRNA profiling of hepatocellular carcinoma could accurately differentiate between the tumors and matched normal liver [98]. Here, we summarized miRNA profiling studies in human malignancies and examine the role of miRNAs in the pathogenesis of cancer in Tab.2 [99].

Table 2: Cancer-related miRNAs summarizing miRNA expression in the major cancer types [99].

Cancer	Up-regulated miRNAs	Down-regulated miRNAs
Breast cancer	miR-21, miR-155, miR-29b-2	miR-143, miR-145, miR-155, miR-200
Lung cancer	miR-21, miR-189, miR-200b, miR-17-92 cluster	let-7 family, miR-126, miR-30a, miR-143, miR-145, miR-188, miR-331, miR-34s
Colon cancer	miR-223, miR-21, miR-17, miR 106m, miR-34s	miR-143, miR-145, miR-195, miR-130a, miR-331
Prostate cancer	let-7d, miR-195, miR-203, miR-125b, miR-20a, miR-221, miR-22	miR-143, miR-145, miR-128a, miR-146a, miR-126
Brain cancer	miR-21, miR-221	miR-181
Hepatocellular carcinoma	miR-34s, miR-224, miR-18, miR-21	miR-17-19b cluster, miR-200a, miR-125a, miR199a, miR195
Chronic lymphocytic leukemia	miR-15, miR-16	
Ovarian cancer	miR-200a,b,c, miR-141	miR-199a, miR-140, miR-145, miR-125b
Pancreatic cancer	miR-221, miR-181a, miR-21,	miR148a,b
Papillary thyroid carcinoma	miR-221, miR-222, miR-146, miR-181	
Stomach cancer	miR-21, miR-103, miR223	miR-218

Since miRNAs are involved in multiple biological processes, metabolic regulation, including cell proliferation, differentiation, and programmed cell death, miRNAs can be viewed as major contributors to the pathogenesis of cancer, including initiation and progression of cancer [100].

Recently, Kataro et al [101, 86] reported miR-222 and miR-339 in cancer cells down-regulate the expression of ICAM-1, thereby regulating the susceptibility of cancer cell to cytotoxic T lymphocytes. This is among the first reports to demonstrate the role of miRNAs in cancer immunosurveillance. Hideho et [71] identified miR-17-

92 family, miR-155, and miR-181a are targets in T cells. In macrophages, miR-125b, miR-146, and miR-155 act as Pathogen Associated Molecular Pattern Molecule-associated microRNAs and miR-34C and miR-214 as Damage Associated Molecular Pattern Molecules-associated miRs. We have also demonstrated that the ability of tumors to serve as targets for cytolytic effectors is regulated by miR-222 and miR-339. In this discovery suggested that roles of miRNAs in immune-regulation will advance the field of cancer immunology and immunotherapy.

MicroRNA and Colorectal Cancer

In 2003, Michael et al [102] published the first study of miRNA in CRC, identifying novel dysregulated miRNAs, miR-143 and miR-145. Akao et al [103] examined the down regulation of let-7 miRNAs and DLD-1 in human colon cancer tumors and cell lines. These findings suggest the involvement of let-7 miRNA in the growth of colon cancer cells. Yang et al [104] reviewed the existing literature pertaining to the study specific expression patterns of miRNAs in CRC. Currently, two different approaches are applied to investigate the connection between miRNAs and CRC. On the one hand, miRNAs seem to regulate many known oncogenic and tumor suppressor signaling pathways involved in the pathogenesis of CRC. Their dissection in function studies is critical for better understanding of cancer biology, eventually aiming for identifications of novel pharmaceutical targets. On the other hand, expression profiles of hundreds of different miRNAs have been shown to bear a much higher potential as biomarkers than their mRNA counterparts. This allows a prediction of prognosis and distinctive stages of disease [105].

Many strong evidences showed that miRNAs have essential roles in the development malignancies and the regulation of immune system. We therefore initiated a study to identify the impact of miRNAs and their target genes on the tumor progression of CRC. The miRNA signatures identified from these clinical studies may serve as potential diagnostic and prognostic disease makers.

Objectives

The objective of this study was to reconstruct miRNA-mRNA networks. We therefore applied a computational-experimental approach to identify miRNA- target genes relationships. As a starting point we chose miRNA expression profiling since this class of small non-coding RNAs was implicated in both, the regulation of the immune system and contribution to the pathogenesis of cancer, including initiation and progression of cancer. Towards this end we performed miRNA and mRNA expression profiling in human colorectal cancer and used statistical methods to identify miRNAs associated with tumor progression. Additionally, bioinformatics methods were applied to select high-confidence miRNA target genes, reconstruct miRNA-mRNA networks and pinpoint immune processes controlling tumor progression.

Results

The MicroRNome profile from a cohort of 105 CRC patients

The expression of 365 miRNAs was analyzed by qPCR using Low Density Arrays in 73 CRC patients (see Materials and Methods). For all the patients a long-term follow-up (>10 years) was available. The data and the corresponding clinical information were stored in a dedicated patient database, the Tumor MicroEnvironment (TME.db). The data was normalized in Genesis [106]. Unsupervised complete linkage clustering based on Pearson Uncentered algorithm was performed without taking into account the corresponding clinical information: i.e. cancer stage or patient outcome. High expressed miRNA profiles were represented in green and the low expressed miRNA in red. Three clusters of patients were revealed (Fig.4). The risk to relapse of those groups of patients was investigated using a Cox proportional hazards model and the significance was assessed by the logrank test. Interestingly, the groups of patients had a significant different risk to relapse (logrank P-value = 2.3e-03, HR = 1.5 [0.09-2.24]). Cluster2 patients showed a significant lower risk to relapse compared to the other patients. In contrast, Cluster3 patients showed a significantly higher risk to relapse. Cluter1 patients had a similar risk to relapse with the other patients. The overrepresentation of relapsing patients in the patient clusters was calculated using the two-sided Fisher's exact test (P-value = 0.0021). A significant predominance of relapsing patients was found in Cluster3 compared to Cluster1 (P-value = 0.0509). In the same time, Cluster2 had significant more non relapsing patients compared to Cluster3 (P-value = 6.0e-04). Cluster1 and Cluster2 had a similar distribution of relapsing patients. Kaplan Meier (KM) curves for three patient clusters based on miRNA expression at the tumor side. (A) Significant separation between the three groups of patients at the minimum p-value cutoff (logrank P-value = 2.3e-03, HR = 1.5 [0.09-2.24]). (B) Significant lower risk to relapse for patients in Cluster2 (P-value = 5.4e-03, HR = 2.7 [1.29-5.65]) and significant higher risk for (C) Cluster3 (P-value = 1.3e-03, HR = 2.63 [1.43-4.82]) compared to the rest of the cohort.

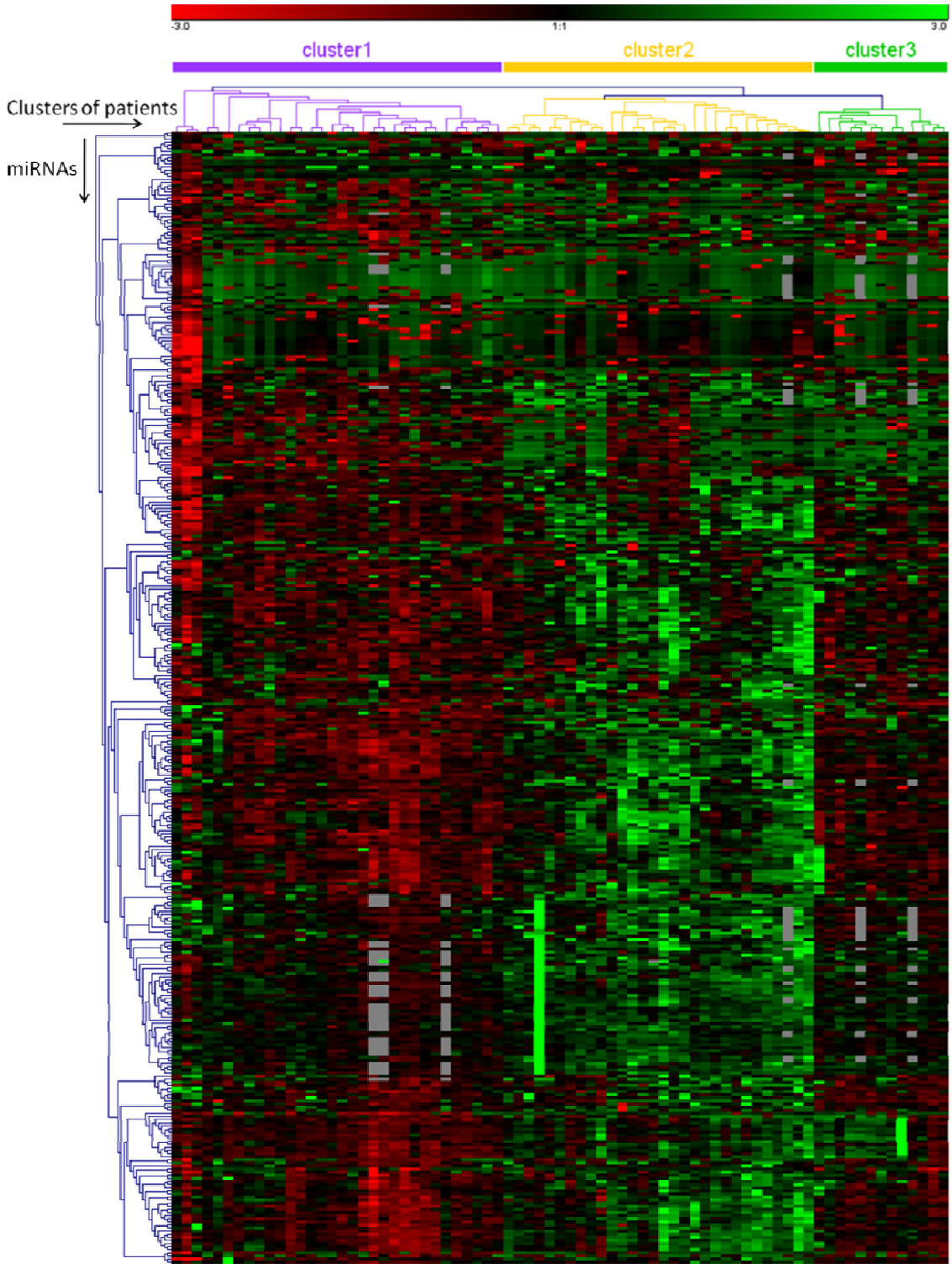


Figure 4: Hierarchical clustering of 365 miRNA expression from a cohort of 73 CRC patients. The data was normalized and hierarchical clustered in Genesis [107]. High expressed miRNA profiles were represented in red and the low expressed miRNA in green. Three patient clusters were defined.

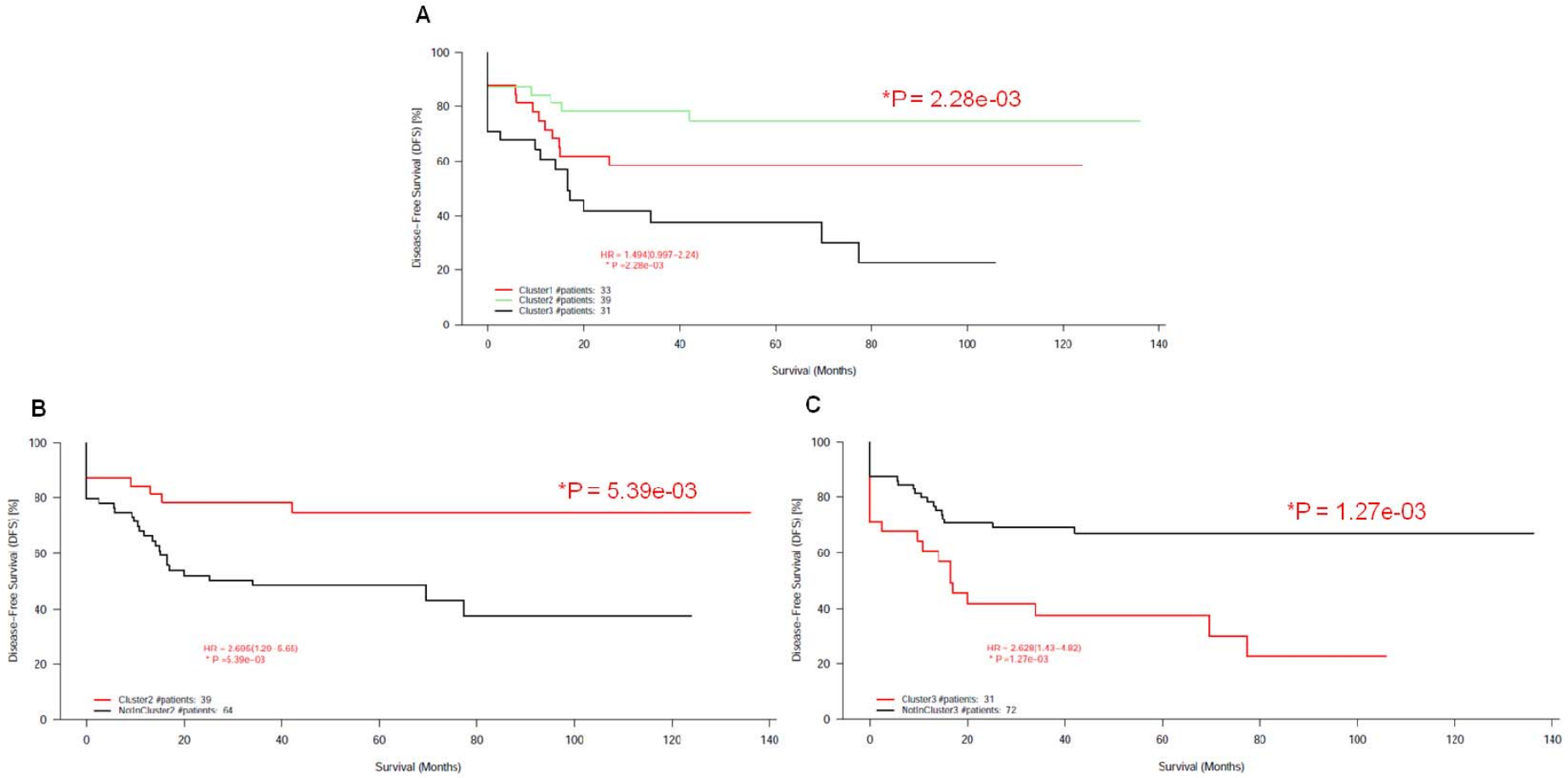


Figure 5: MiRNA expression impact on CRC patients DFS. A) Kaplan Meier (KM) curves for three patient clusters. Cluster1 in red, Cluster2 in green, Cluster3 in black. KM curves for Cluster2 (B) and Cluster3 (C) compared to the rest of the cohort. Cluster 2 and 3 in red, the rest of the cohort in black.

Univariate survival analysis on Low density array (LDA) Real-Time Taqman PCR and microRNome data using Log-Rank and Cox Regression

Log-Rank survival analysis could show the predictive strength of single miRNA marker predicting patient recurrence. Hazard Ratios (HR) were calculated for overall survival (OS) and disease free survival (DFS) using a univariate Cox regression model on the 381 genes for 125 patients and 365 miRNAs measured for 103 patients. Patients were separated into two groups (high, represented in red, and low, represented in black) depending on the level of expression of miRNAs at the median cutoff. Hazard ratios (HR) are shown in (Fig. 6). The miRNAs can be divided into two groups: good prognosis (HR<1), e.g. miR-626 and poor prognosis (HR>1), e.g. miR-519d.

Differential expression of miRNAs and tumor progression

Differential expression of a molecule in tumor and normal tissues often represents an important basis for cancer biomarker exploration and development. For diagnostic biomarkers, one would commonly focus on overexpressed targets, while for prognostic and predictive biomarkers, as well as for defining potential therapeutic targets, both over- and underexpression could be biologically interesting. To evaluate the changes in miRNA expression during tumor progression, we compared the miRNA expression in normal colon mucosa with the miRNA expression in different tumor stages. We found the expression of 12 miRNAs (miR-660, miR-657, miR-29a, miR-519d, miR-518c, miR-302a, miR-558, miR-603, miR-609, miR-376a, miR-130a, and miR-211) showing a strong association with a higher UICC stage compared to normal colon mucosa (Fig.7). The expression levels of miR-302a, miR-558, and miR-603 have been shown to be gradually decreased from normal mucosa to T4. In contrast, the expression of miR-660, miR-657, miR-29a and miR-519d were found to be differentially up-regulated in the different T-stage. The expression of a miRNAs may also decrease (or increase) at a particular stage of cancer (for example, miR-211 has shown down-regulated in T1, T2 and T4 but up-regulated in T3).

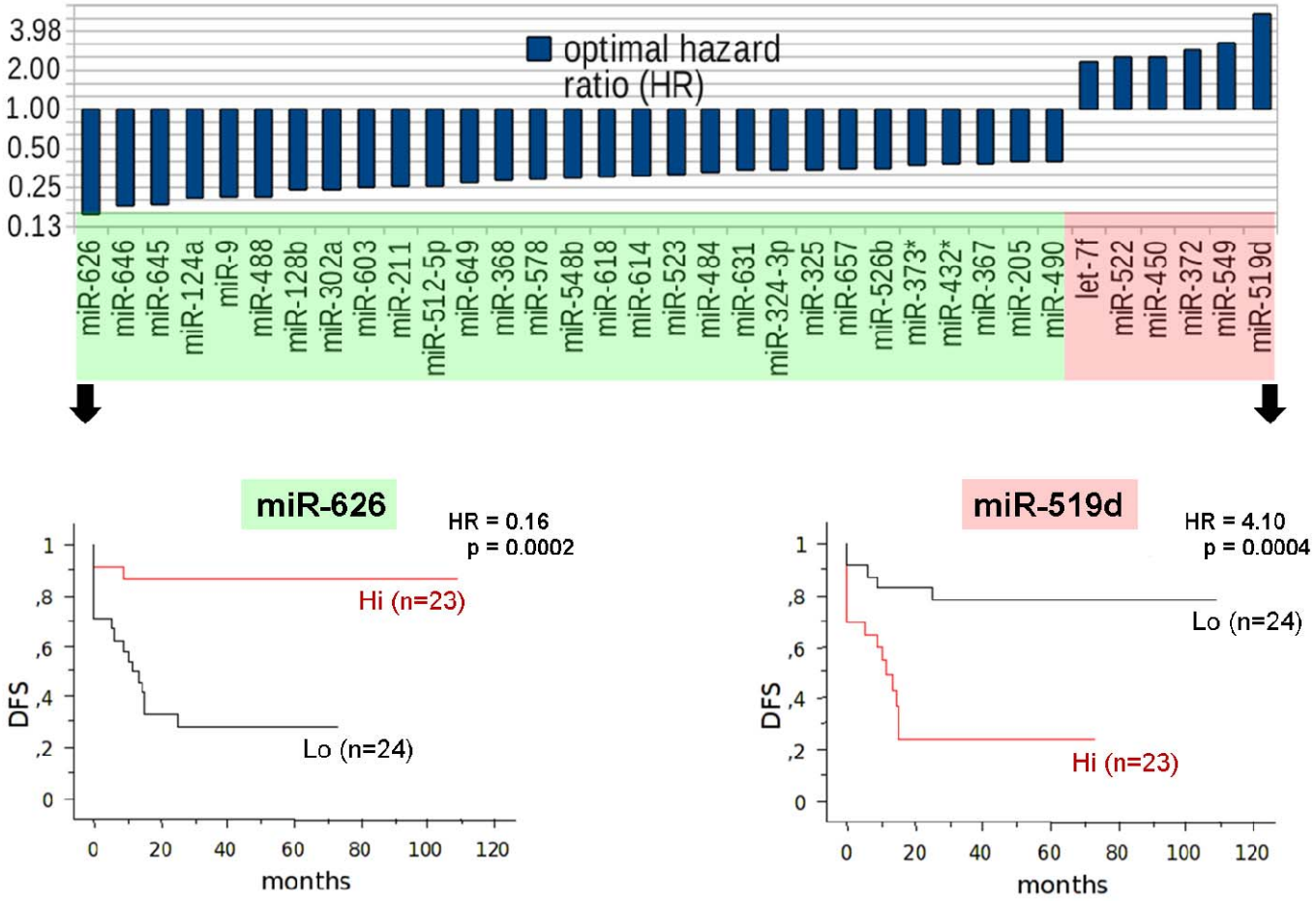


Figure 6: MiRNA impact on CRC patient DFS. (A) miRNAs with positive (marked in green) and negative impact (marked in red) on DFS. (B) Kaplan Meier curves for representative miRNAs with good (miR-626) and bad (miR-519d) impact on patient outcome, respectively. CRC patients with high miRNA expression are shown in red and with low expression in black.

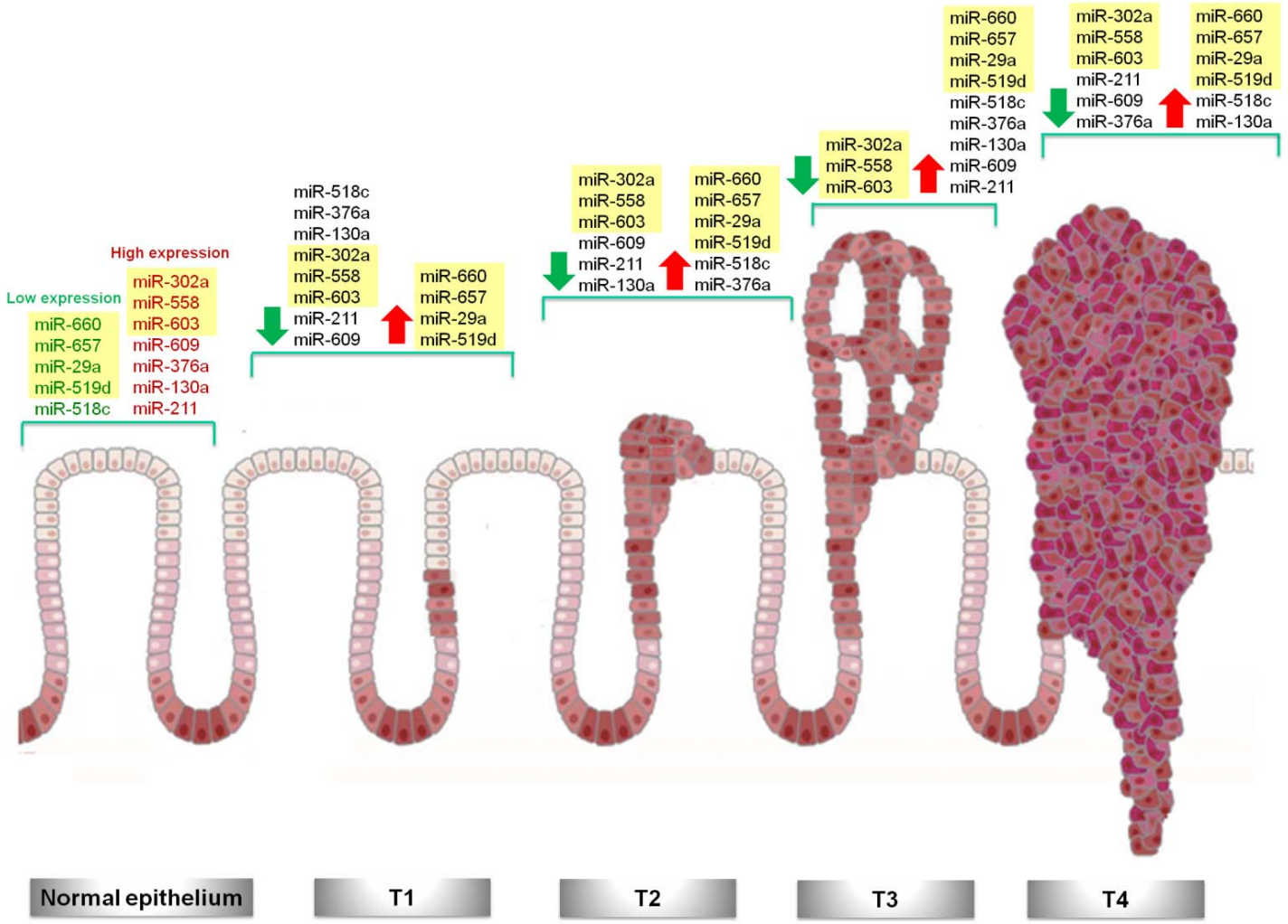


Figure 7: MiRNAs expression during tumor progression. Red and green arrows show miRNAs up- or downregulated in each tumor stage (T1, T2, T3, T4) compared to the normal epithelium. In yellow, miRNAs over- or underexpressed at different tumor stages.

Identification of miRNA-target genes for tumor progression

First we used 10 prediction tools (TargetScan (conserved), TargetScan (nonconserved), PicTar (4-way), PicTar (5-way) miRanda (miRBase), miRanda (microrna.org), PITA, EIMMo, RNA22, DIANA-microT) to obtain, based on public data, candidate miRNA-mRNA target interactions for 365 miRNAs. The predictions reflect miRNA:mRNA paring, site location, conservation, site accessibility, multiple sites and expression profile (see Materials and Methods). The resulting gene-miRNA pairs were used as input for GenMiR++.

In the next step we used a customized GenMiR++ code (see Materials and Methods) to integrate the *in silico* predicted pairs with miRNAs and mRNA expression profiles from colorectal tumors. 365 miRNAs and 19,806 mRNAs were investigated in 73 tumors from CRC patients in all tumor stages. The top 25% predicted pairs (365 miRNAs and 381 target genes) were selected as representative for CRC at a false detection rate of 0.05. This analysis allowed us to identify 12,065 high confidence miRNA-mRNA pairs. Since, the miRNAs will degrade their target genes, so miRNA expression should show a negative correlation with the respective target gene. We calculated Pearson correlation coefficients between miRNAs and their targets with a strong negative correlation ($R < 0$, P-value < 0.05). In total, we associated both analyses and obtained 788 specific pairs for CRC. Figure 8 shows the landmark of the 12 most differential expressed miRNA and their target genes specific for tumor progression.

Interestingly, our results have shown that miR-302a and miR-660 are significantly negative correlated in every stage of tumor (T1-T4) with their target gene, Fractalkine (CX3CL1). Additionally, CX3CL1 was one of the target genes of miR-29a and miR-519d. Significant negative correlation between CX3CL1 and those miRNAs was shown in tumor stages T2-T4.

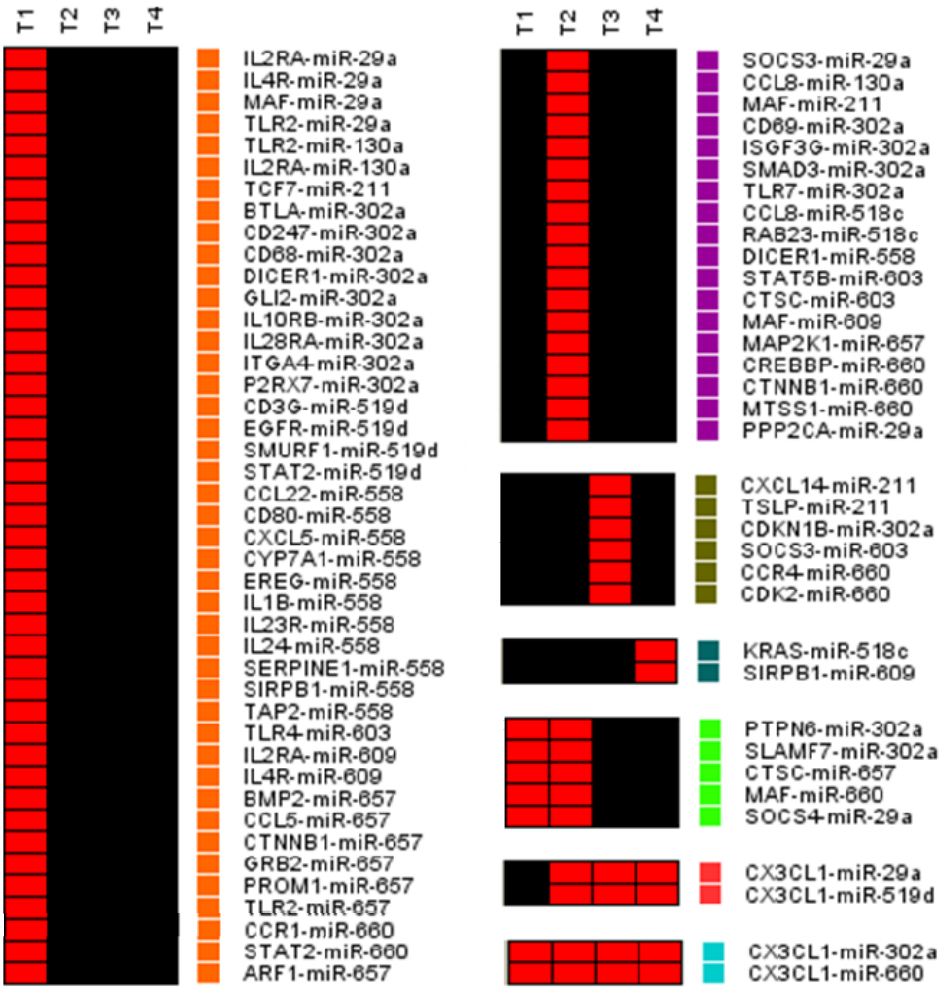


Figure 8: A landmark of specific miRNAs and their target genes in colon cancer tumor progression. 78 high confidence miRNA-gene pairs, negatively correlated, from 73 CRC patients were visualized in Genesis [107]. In red, pairs specific for different tumor stages: T1, T2, T3 and T4. In black, absence of correlation.

From the total of 78 high confidence miRNA-target genes selected 48 were specific for T1, 18 were specific for T2, 6 for T3 and 2 for T4. We could also show 5 miRNA-gene pairs specific for early stages (T1 and T2), miRNA-gene pairs present in more advanced stages (T2, T3 and T4) as well as miRNA-genes present in all tumor stages.

We hypothesized on predicted “direct” target of miRNA following the involvement of miRNAs in cancer [91, 92, 87, 108] and has significant negative correlation between miRNA and its target. To gain our confident on predicted direct

target, CX3CL1, of miR-29a, miR-519d, miR-302a and miR-660 based on expression profiles and regulation, we used the definition of “oncomiRNAs” as the over expression of miRNAs can downregulate the expression of tumor suppressor genes following [91-94, 108]. Similarly, miRNAs can downregulate oncogenes, we referred as “tumor suppressor miRNAs” [91-94, 108]. As a result of this study, miR-660, miR-29a and miR-519d can be oncomiRNAs and showed strong negative correlation with tumor suppressor gene, CX3CL1. We may suggest that CX3CL1 is the direct target of these miRNAs. In contrast, based on functional of miRNA defined miR-302a as tumor suppressor miRNA and showed significant negative correlation with CX3CL1, we may conclude that CX3CL1 is “not” the direct target of miR-302a because of its role in caner.

We used the median expression value of CX3CL1 over the 43 patients as cutoff to define a low and high gene expression patient group. For each of the two patient groups the expression of selected miRNAs was calculated and plotted (Fig.9). MiR-29a and miR-660 were significantly different (t-test; P-value < 0.05) but miR-302a and miR-519d were not significant difference between two patient groups. Thus, we concluded that CX3CL1 is the predicted direct target of miR-29a and miR-660 and be predicted “indirect” target of miR-302a and miR-519d.

In the next step we investigated the miR-302a, miR-660, miR-29a and miR-519d expression in relation to the expression CX3CL1. We used the median expression value of CX3CL1 over the 43 CRC patients as cutoff to define a low (Lo) and high (Hi) gene expression patient group. For each of the two patient groups the expression of selected miRNAs was calculated and plotted (Fig.9). In the CX3CL1 Lo group, miR-29a showed a significant higher expression compared to the CX3CL1 Hi group (Fig.9A, P-value = 0.037). A similar result was found for miR-660 (Fig.9D, P-value = 0.004). Those findings are in concordance with our previous results showing a significant negative correlation between miR-29a and miR-660 and their target gene, CX3CL1. In contrast, the expression of miR-302a and miR-519 was higher in the CX3CL1 Hi group compared to the CX3CL1 Lo group.

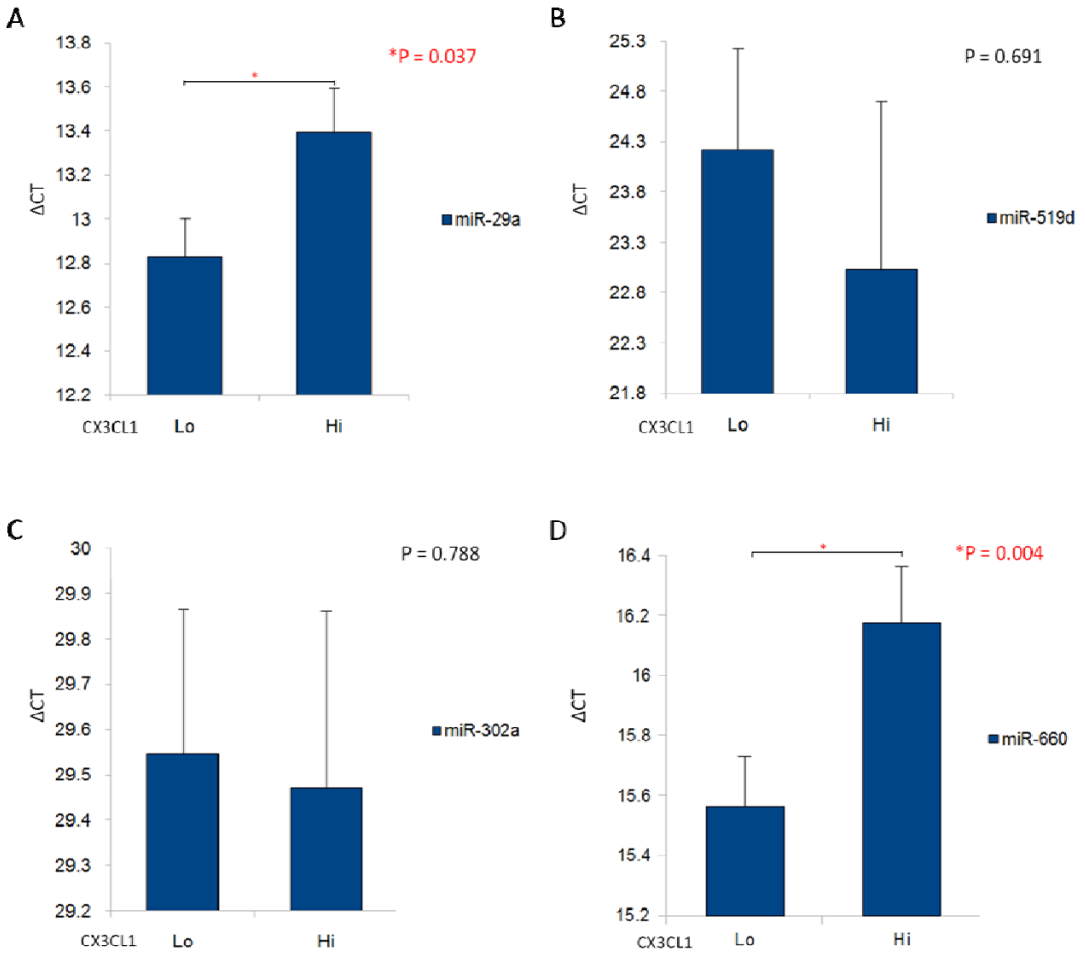


Figure 9: Expression of miR-29a, miR-302a, miR-519d and miR-660 in 43 CRC patients. The patients are split into a high (Hi) and a low (Lo) group based on the CX3CL1 expression level. The t-test significant different miRNA expression is marked in red.

Target gene networks and involvement in immunological processes of miR-29a, miR-519d, miR-302a and miR-660

Further we performed functional analysis for the four selected miRNAs (miR-29a, miR-519d, miR-302a, miR-660). For each tumor stage we investigated the immune roles of the target genes predicted for all those miRNAs. A Cytoscape [109] network presents the miRNA and its predicted target genes, selected based on high confidence scoring and significant negative correlation (upper panel of Fig.10-13). Triangles, circles and diamonds in the network are indicating miRNAs, target genes, and transcription factors, respectively. Survival analysis results were included in the

network. The impact of the miRNAs and target genes on the patients' DFS was shown by using two different node colors: green (good outcome, HR<1) and red (bad outcome, HR>1). The size of the nodes is based on log2 (HR) for disease free survival. Triangles, circles and dimonds in the network indicated miRNAs, target genes, and transcription fators, respectively. The impact of the miRNAs and target genes on the patients' DFS was shown by using two different node colors: green (good outcome, HR<1) and red (bad outcome, HR>1). The size of the nodes is based on log2 (HR) for disease free survival. Edge represents significant negative correlation between miRNAs and mRNAs.

We investigated the functions of miRNAs target genes and using Gene Ontology [110] annotations. Functional analysis was performed for the negative correlated high confidence pairs. Cytoscape [109], a public available software was used to visualize the miRNAs and their target genes. Further, the immune functions triggered by the miRNA through its target genes were investigated using ClueGO [111], a Cytoscape plugin. For miR-29a and miR-302a, GO annotations revealed specific terms like "T cell differentiation" in tumor stage T1. The common functionality of each selected miRNAs refers to characteristics of lymphocytes (chemotaxis) and leukocyte (adhesion). In the tumor stage T3-T4, the function of these miRNAs is involved in both the innate (macrophage) and adaptive immune response (T helper 1 cell).

The impact of the miRNAs and target genes in tumor recurrence

The expression level of miRNA and target genes have, for some cancers, been reported to associated with clinical diseased courses [39, 97, 98, 112-114]. Hence, we were interested in evaluating the prognostic potential of the four miRNAs (miR-29a, miR-519d, miR-302a and miR-660) of their target gene, CX3CL1. Survival analysis in a cohort of 63 CRC patients could show that patients with high expression of CX3CL1 have a significant lower risk to relapse compared to the patients with low expression of this gene (Fig.14A, P-value = 2.39e-03, HR = 1.84 [1.08-3.33]). Similar, a significantly lower risk to relapse was shown in patients with high expression of miR-302a (Fig.14D, P-value = 8.07e-03, HR = 2.21 [1.18-4.15]). In contrast, a significant higher risk to relapse was shown for the patients having high

expression of the miR-660 (Fig.14E, P-value = 5.68e-04, HR = 3 [1.54-5.85]). Patients having a high expression of miR-29a (Fig.14B) and miR-519d (Fig.14C) had a higher risk to relapse (HR = 1.66 [0.903-3.07], HR = 1.59 [0.852-2.99], respectively) but not significantly different compared to the patients with low expression of those miRNAs.

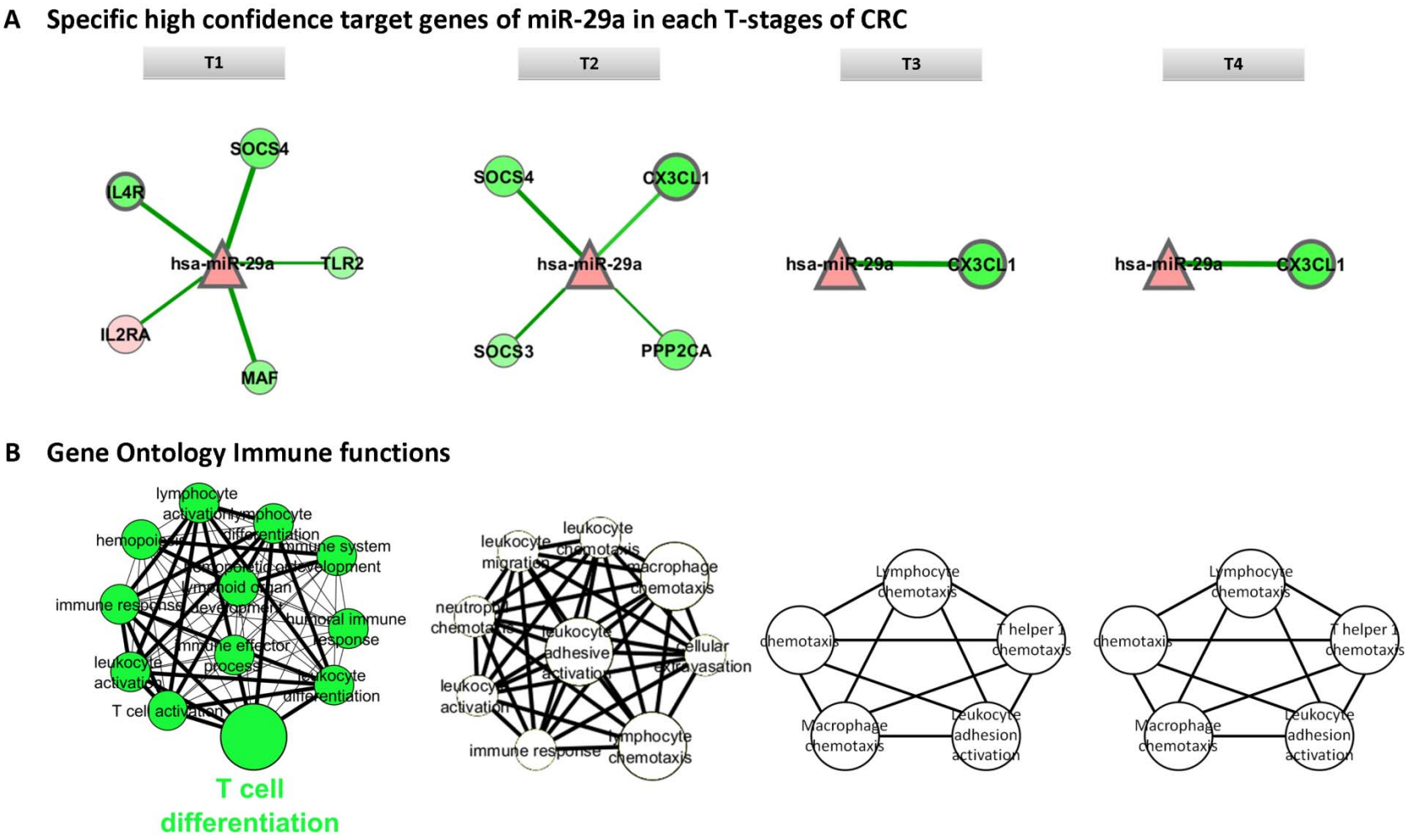


Figure 10: Functional analysis of miR-29a and of its target genes with tumor progression. Panel A miR-29a and its target genes at different tumor stages. Triangles, circles and diamonds are miRNAs, target genes, and transcription factors, respectively. The node color shows the good (green, HR<1) and bad (red, HR>1) outcome. The size of the nodes is based on log2 (HR) for DFS. Panel B Networks of Gene Ontology, KEGG and BioCarta terms were visualized in ClueGO. The size of the nodes shows the significance of the terms and the links are based on kappa statistics.

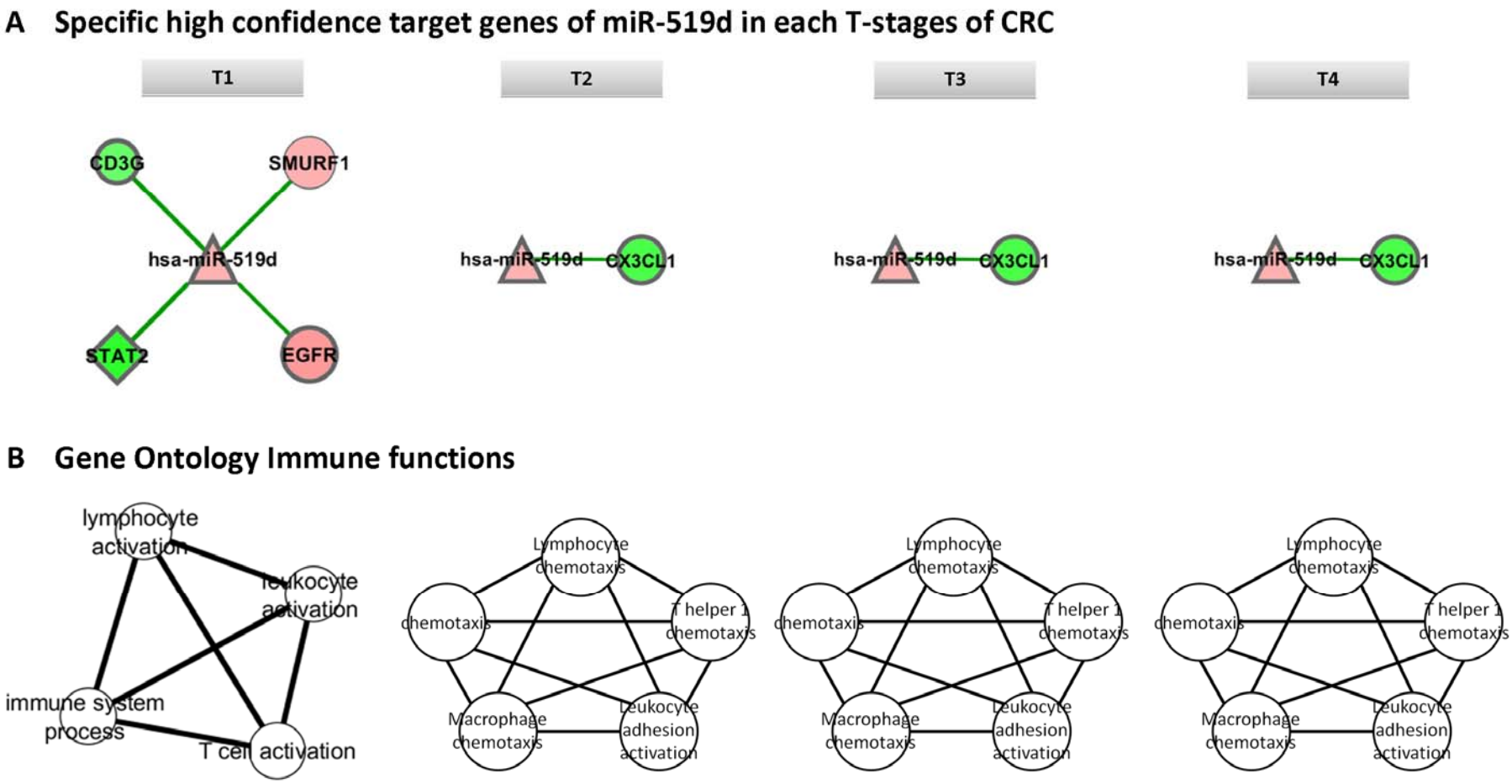


Figure 11: Functional analysis of miR-519d and of its target genes with tumor progression. Panel A miR-519d and its target genes at different tumor stages. Triangles, circles and diamonds are miRNAs, target genes, and transcription factors, respectively. The node color shows the good (green, HR<1) and bad (red, HR>1) outcome. The size of the nodes is based on log2 (HR) for DFS. Panel B Networks of Gene Ontology, KEGG and BioCarta terms were visualized in ClueGO. The size of the nodes shows the significance of the terms and the links are based on kappa statistics.

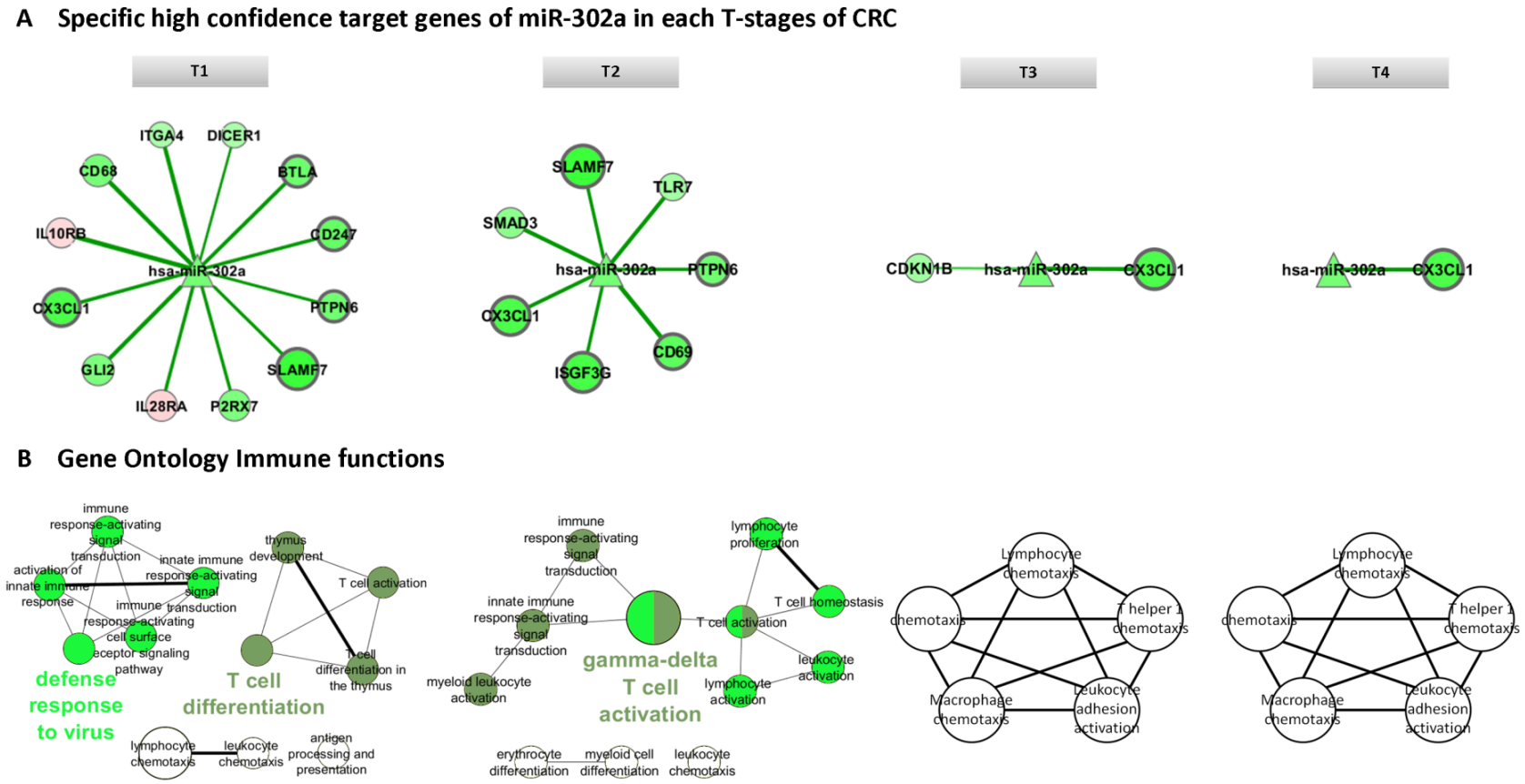


Figure 12: Functional analysis of miR-302a and of its target genes with tumor progression. Panel A miR-302a and its target genes at different tumor stages. Triangles, circles and diamonds are miRNAs, target genes, and transcription factors, respectively. The node color shows the good (green, HR<1) and bad (red, HR>1) outcome. The size of the nodes is based on log2 (HR) for DFS. Panel B Networks of Gene Ontology, KEGG and BioCarta terms were visualized in ClueGO. The size of the nodes shows the significance of the terms and the links are based on kappa statistics.

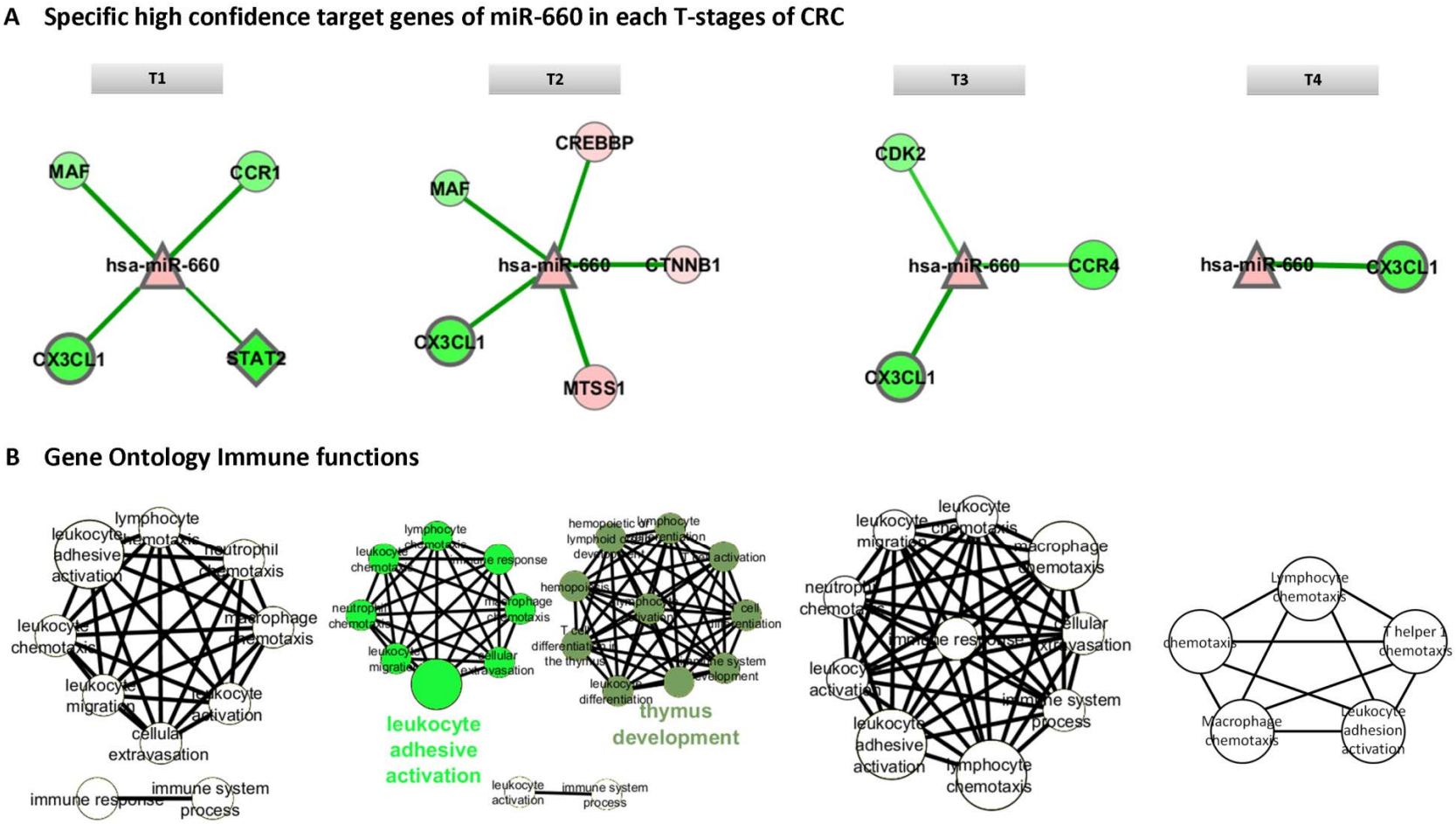


Figure 13: Functional analysis of miR-660 and of its target genes with tumor progression. Panel A miR-660 and its target genes at different tumor stages. Triangles, circles and diamonds are miRNAs, target genes, and transcription factors, respectively. The node color shows the good (green, HR<1) and bad (red, HR>1) outcome. The size of the nodes is based on log2 (HR) for DFS. Panel B Networks of Gene Ontology, KEGG and BioCarta terms were visualized in ClueGO. The size of the nodes shows the significance of the terms and the links are based on kappa statistics.

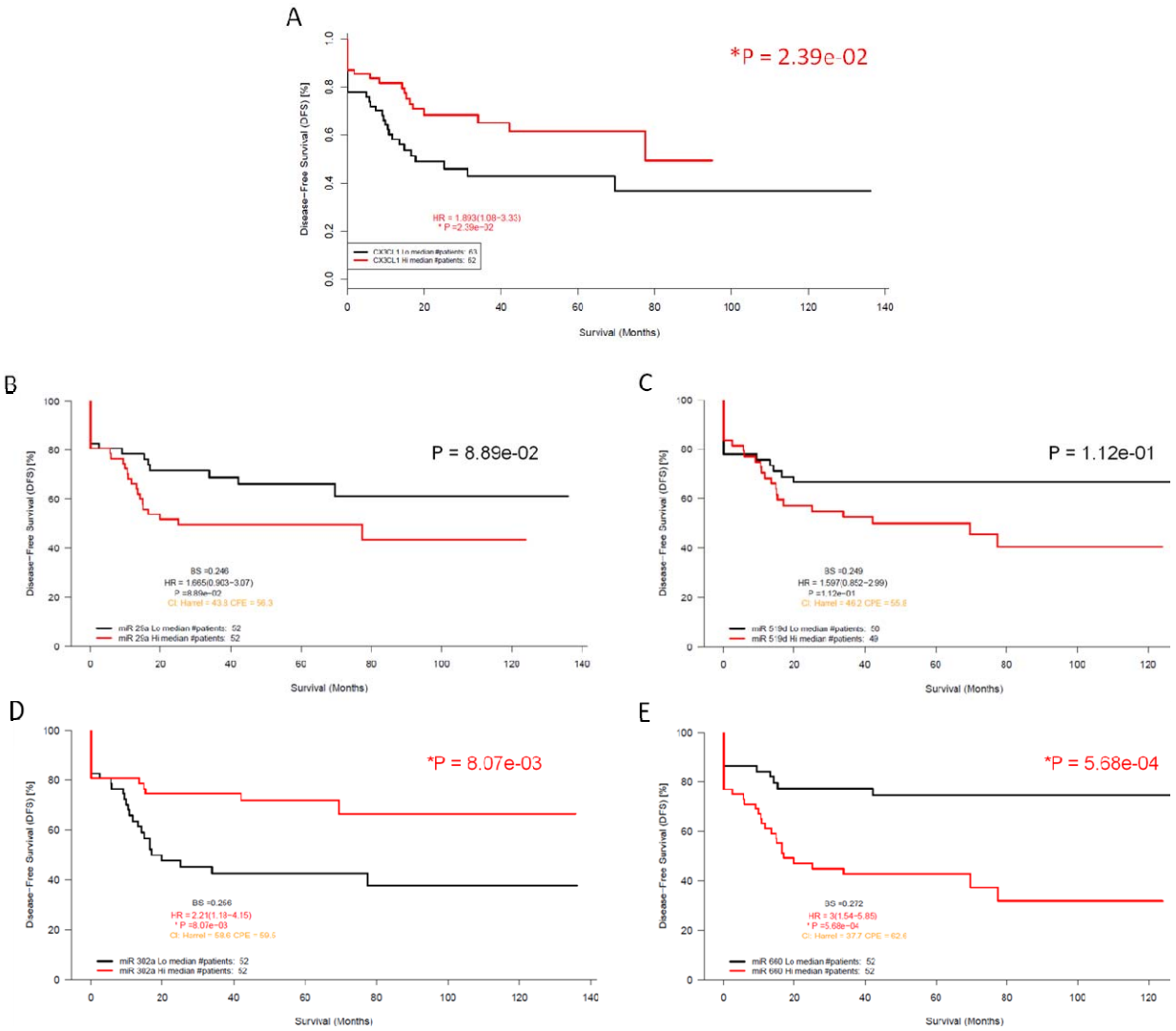


Figure 14: Kaplan-Meier survival curves for miR-29a, miR-302a, miR-519d, miR-660 and CX3CL1 at the median cutoff. High and low expression is shown in red and black, respectively.

Array comparative genomic hybridization (aCGH) analysis for selected miRNAs correlated with tumor progression

At the cellular level, cancer is a genetic disease; genetic changes in somatic cells are essential events in neoplasia. The importance of DNA copy number aberrations has been demonstrated in many tumors [115]. Detecting these aberrations by array comparative genomic hybridization (array-CGH) provides information on the locations of important cancer genes and can have clinical use in diagnosis, cancer classification and prognostic. For example, Geigl et al [116] proposed a new protocol for single-cell isolation and whole genome amplification by array-CGH which is crucial and suitable for genomic instability patterns within primary tumors. Other technical considerations related to array-CGH analysis of tumor cells have been proposed and reviewed [117, 118].

In this study, we used array-CGH from 216 CRC patients to investigate the amplification and deletion status of the 11 miRNAs identified as specific for CRC tumor progression. The patients were split in groups based on their tumor stage. The amplification score (Fig. 15A) was calculated for each miRNA considering the mean amplification of the miRNA and its frequency in a group. The deletion score (Fig.15B) was calculated in the same way.

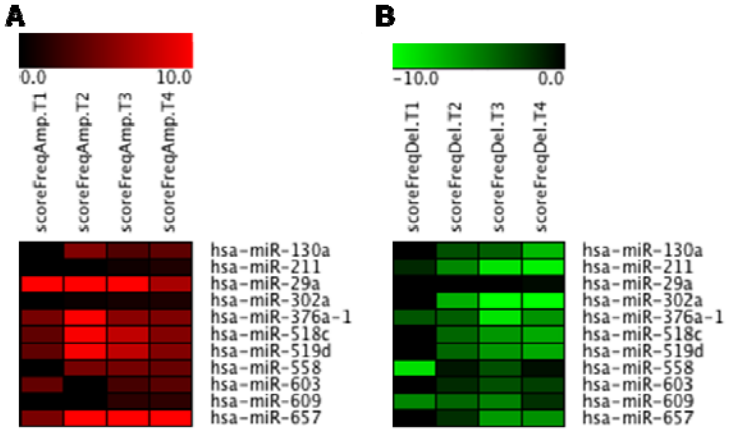


Figure 15: Array-CGH data for 11 miRNAs involved in CRC tumor progression from 216 CRC patients. Mean amplification (A) and deletion (B) score for each tumor stage visualized in Genesis [107].

The amplification score of miR-29a was gradually increased in tumor stages T1-T3 and decreased in T4. Conversely, the amplification score of miR-657 were rapidly increased in from T1 to T2 and slightly decreased in tumor stages T2 to T4. As a result of the deletion scores of miR-302a which were rapidly increased in all tumor stages (T1-T4), suggested that miR-302a could be involved in the progression of the disease.

In the next step we investigated the impact of the amplifications/deletions of the selected miRNAs on patient survival. Interestingly, the patients having deleted miR-302a on chromosome 4 showed a significantly higher risk to relapse compared to the patients that had no aberration (Fig.16A, logrank P-value = 0.007, HR= 1.945). Furthermore, patients with low expression of miR-302a showed a higher risk to relapse compared to patients having high miR-302a expression (Fig.16B, logrank P-value = 8.07e-03, HR = 2.21 [1.18-4.15]). Thus, the decreased expression of miR-302a is likely due to deletion of chromosome region and correlate with bad outcome.

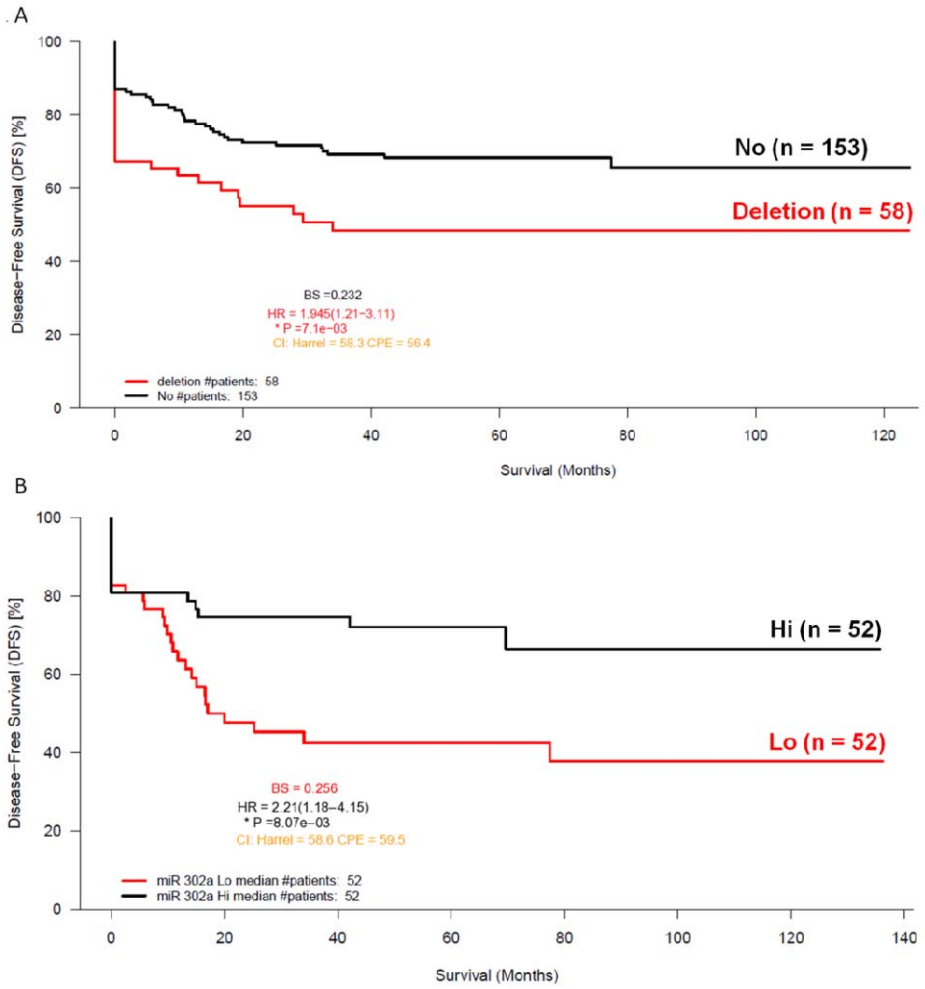


Figure 16: Disease-free survival of colorectal cancer patients according to expression of miR-302a and deletions in genomic area. Kaplan Meier curves for mirR-302a deletion (A) and expression (B). Hi expressed miRNA and the absence of aberration is shown in black. Lo expressed miRNA and deletion is shown in red.

Discussion

In this study we propose a landmark of miRNAs and high confidence targets in human colorectal cancer showing a strong association with a higher UICC stage compared to normal colon mucosa. We used expression profiles of miRNA and mRNA and analyzed the data using customized prediction tool (GenMiR++) as well as 10 public prediction tools to score and then to select high confidence miRNA-mRNA pairs. We hypothesized that predicted "direct" target gene has significant negative correlation with miRNA. Interestingly, our results demonstrated that miR-302a and miR-660 demonstrate significant negative correlations with Fractalkine (CX3CL1) which is their target genes in every stage of tumor development (T1-T4). Furthermore, miR-29a and miR-519d also show significant negative correlations with CX3CL1 which is their target gene as well in tumor stages T2-T4. On our previous study by Mlecnik et al [119] we proposed that the high expression of this chemokine showed a good prognostic factor in CRC. This observation was confirmed by Xin et al [120] showing strong evidence that CX3CL1 can be a suitable candidate for immunogene therapy of cancer because CX3CL1 induces both innate and adaptive immunity, and can act as tumor suppressor gene. Thus, the data suggested that these four miRNAs may suppress the function of CX3CL1 which may also be involved in tumor progression.

However, correlation analysis alone is not adequate to conclude that CX3CL1 may be direct target gene of these miRNAs. We therefore compared significant difference of expression of these miRNAs between two patient groups (based on mean cutoff value of CX3CL1 expression profiles). Our results demonstrated that miR-29 and miR-660 were significantly different (t-test; P-value < 0.05) but miR-302a and miR-519d were not significantly different between two patient groups. Thus, we conclude that CX3CL1 is a direct predicted target of miR-29a and miR-660 and can be an indirect predicted target of miR-302a and miR-519d.

Differential expression of gene/miRNA in tumor and normal tissues often represents an important basis for cancer biomarker exploration and development [121]. In general, miRNA expression levels have been shown to be decreased in

cancerous compared to corresponding normal tissue [122]. However, contrasting these findings are numerous studies on clinical samples showing that specific miRNAs can be differentially up- and down-regulated in different cancer types [123, 88, 93, 80]. Thus, for the development of diagnostic biomarkers, both over- and underexpression of candidate genes/miRNAs could be interesting. We found that the expression levels of miR-302a, miR-558, and miR-603 are decreased in cancerous compared to corresponding normal tissue. In contrast, the expression of miR-660, miR-657, miR-29a and miR-519d were found to be differentially up-regulated in the different T-stage. These 7 miRNAs may be involved in cancer relevant processes such as proliferation and differentiation.

In cancer, miRNAs function as regulatory molecules, acting as oncogenes which downregulate expression of tumor suppressor genes, so-called “oncomiRNAs”. Similarly, miRNAs can function as “tumor suppressor miRNAs” by downregulated expression of oncogenes. Then following a result of a landmark of miRNA suggest that miR-29a, miR-519d and miR-660 act as oncogenes and miR-302a may act as tumor suppressor gene. It is interesting to note that some miRNAs may have “dual” functions depending on the context. For example, miR-29a has been to shown to function as tumor suppressor in lung cancer [124, 125] and chronic lymphocytic leukemia [80, 126, 127]. In our in this study we found that miR-29a may act as oncogene in CRC. Hence, in appears that in different cancer types, some miRNAs may exhibit this type of dual function. Since misregulation of miRNA has been associated with various cancers, the identification of specific regulators of miRNAs will be helpful in developing new therapeutic agents.

In order to elucidate the mechanisms by which the identified miRNAs are driving tumor progression we used array-CGH to pinpoint genomic deletion events. Array-CGH is one of a growing number of “top-down” approaches that able to provide comprehensive information about aspects of biological status and functions [118]. Here, we found that patients with deletion of miR-302a (on Chromosome 4) have more risk to relapse and the decreased expression of miR-302a also showed a significant difference in DFS compare with the high expression. Following these strong evidences, the decreased expression of miR-302a is likely due to deletion of

chromosome region and correlate with poor outcome. Thus, miR-302a may be proposed as novel therapeutic marker of CRC.

Finally, we investigated the functions of miRNAs target genes and using Gene Ontology [110] annotations. Functional analysis was performed for the negative correlated high confidence pairs. Cytoscape [109], a public available software was used to visualize the miRNAs and their target genes. Further, the immune functions triggered by the miRNA through its target genes were investigated using ClueGO [111], a Cytoscape plugin. For miR-29a and miR-302a, GO annotations revealed specific terms like “T cell differentiation” in tumor stage T1. The common functionality of each selected miRNAs refers to characteristics of lymphocytes (chemotaxis) and leukocyte (adhesion). In the tumor stage T3-T4, the function of these miRNAs is involved in both the innate (macrophage) and adaptive immune response (T helper 1 cell).

In summary, the identified miRNA candidates, the analysis of their target genes, genome wide screening by array-CGH together with the analysis of clinical parameters show promising potential of selective miRNAs as a novel therapeutic means to treatment of CRC. The clinical validation of suggested miRNAs and their target genes is required to confirm these results. Generating a more complete picture of miRNA expression and clinical relevance in CRC and gaining knowledge of targets and cellular effects is a major task, but acknowledging the potential utility of mRNA as biomarkers it is worth the effort.

Conclusions

The discovery of miRNAs as regulators of developmental events in model organisms suggested that miRNA might be involved in the regulation of immune system and the tumor progression. Our results suggest that miRNAs and their high-confidence targets may have a functional effect on tumor progression. Furthermore, some miRNAs with prognostic potential could provide the basis for an in-depth studies as clinical prospective markers and as new pharmaceutical targets.

Materials and Methods

Patient cohort

The records of 415 colorectal cancer (CRC) patients who underwent a primary resection of their tumor at the Laennec-HEGP Hospitals between 1990 and 2003 were reviewed and previously described in (Pagès et al. 2005). The observation time in the cohort was the interval between diagnosis and last contact (death or last follow-up). Data were censored at the last follow-up for patients without relapse, or death. The mean duration of follow-up was 45.3 months. The min:max values until progression/death or last follow-up were (0:166) months, respectively. Time to recurrence or disease-free time was defined as the interval from the date of surgery to confirmed tumor relapse date for relapsed patients and from the date of surgery to the date of last follow-up for disease-free patients.

Histopathological and clinical findings were scored according to the UICC-TNM staging system (Dukes, Weitz). Post-surgical patient surveillance was performed at Laennec-HEGP Hospitals for all patients according to general practice for CRC patients. Adjuvant chemotherapy was administered to patients with stage III CRCs, to high-risk stage II CRCs, and palliative chemotherapy to patients with advanced colorectal cancers (stage IV) and to patients without complete resection of the tumor. Adjuvant chemotherapy was fluorouracil (FU)-based. Follow-up data were collected prospectively and updated.

Low density array (LDA) Real-Time Taqman PCR analysis

This study was based on tissue sample material collected at the Laennec-HEGP Hospitals (Hôpital Européen Georges Pompidou) which was snap-frozen within 15 minutes after surgery and stored in liquid nitrogen. From this material 154 frozen tumor specimens were randomly selected for RNA extraction. The total RNA was isolated by homogenization with the RNeasy isolation kit (Qiagen, Valencia, CA). A bioanalyzer (Agilent Technologies, Palo Alto, CA) was used to evaluate the integrity and the quantity of the RNA. The 125 analyzed RNA samples were all from

different patients. 381 genes were selected for real-time TaqMan analysis. This gene selection covers the representative cell subpopulations according to the Immunome selection. The RT qPCR experiments were all performed according to the manufacturer's instructions (Applied- Biosystems, Foster City, CA). The quantitative real-time TaqMan qPCR analysis was performed using Low Density Arrays and the 7900 robotic real-time PCR-system (Applied Biosystems). As internal control 18S ribosomal RNA primers and probes were used. The data was analyzed using the SDS Software v2.2 (Applied Biosystems) and TME.db statistical module.

Affymetrix gene chip analysis

The tissue sample material was collected at the Laennec-HEGP Hospitals (Hôpital Européen Georges Pompidou) which was snap-frozen within 15 minutes after surgery and stored in liquid nitrogen. 105 frozen tumor specimens and 5 normal specimens from distant tissue were randomly selected for RNA extraction. The total RNA was isolated by homogenization with the RNeasy isolation kit (Qiagen, Valencia, CA). A bioanalyzer (Agilent Technologies, Palo Alto, CA) was used to evaluate the integrity and the quantity of the RNA. From this RNA 110 Affymetrix gene chips were done on the same platform (HG-U133A plus) than the Immunome using the HG-U133A GeneChip 3' IVT Express Kit. The raw data was normalized with CARMAweb [22], using the GCRMA-algorithm. Finally, the log2 intensities of the gene expression data were used for further analysis. For correlation analysis the spots which were not significant with MAS5Calls and had a log2 intensity lower than 3 were excluded from the analysis due to insufficient sensitivity.

MicroRNAs expression analysis

Tissue samples were snap-frozen within 15 minutes after surgery and stored in liquid nitrogen. Randomly selected frozen tumor specimens from Laennec-HEGP Hospitals were extracted for RNA. Total RNA was isolated by homogenization with RNAnow (Biogentex, Seabrook, TX). The integrity and the quantity of the total RNA were evaluated on a bioanalyzer-2100 (Agilent Technologies, Palo Alto, CA). 103 samples from different patients were analyzed for 365 mature miRNA expressions. Low Density Arrays for multiplex miR expression analysis were performed according

to the manufacturer's instructions (Applied-Biosystems, Foster City, CA). RT-PCR experiments and quantitative real-time TaqMan-PCR was performed using the 7900 robotic real-time PCR-system (Applied-Biosystems). Mean Ct values obtained with small ribonucleotide primers (8 replicates of RNU44 and 8 replicates of RNU48) were used as internal control and for calculation of dCt values. Data were analyzed using the SDS Software v2.2 (Applied-Biosystems) and TME statistical module.

Array comparative genomic hybridization (array-CGH)

This study was based on tissue sample material collected at the Laennec-HEGP Hospitals (Hôpital Européen Georges Pompidou) which was snap-frozen within 15 minutes after surgery and stored in liquid nitrogen. From this material 216 frozen tumor specimens were randomly selected for DNA extraction. Samples were homogenized (ceramic beads and FastPrep-24, MP biomedical) in 430 ul of a lysis buffer (Tris 1M – EDTA 0,5M pH8; SDS 20%; proteinase K), and incubated overnight at 37°C. Genomic DNA was extracted by phenol–chloroform extraction and ethanol precipitation. Genomic DNA was re-suspended in 200 ul of highly pure water. Concentrations were evaluated by Optic Density measurement. Samples were labeled using a Bioprime Array CGH Genomic Labeling Kit according to the manufacturer's instructions (Invitrogen, Carlsberg, CA). 500 ng test DNA and reference DNA (Promega, Madison, WI) were differentially labeled with dCTP-Cy5 and dCTP-Cy3, respectively (GE Healthcare, Piscataway, NJ). Genome-wide analysis of DNA copy number changes was conducted using an oligonucleotide array containing 44,000 probes with a spatial resolution of 35 kb according to the manufacturer's protocol version 6.0 (Agilent, Santa Clara, CA). Slides were scanned using Agilent's microarray scanner G2505B and analyzed using Agilent DNA Analytics software 4.0.76 (statistical algorithm: ADM-2; sensitivity threshold: 6.0; consecutive clone filter: 10).

Computational and Bioinformatics Methods

The data resulting from those diverse high-throughput technologies were integrated using in house developed tools: TME.db, ClueGO [111] and Genesis [106] and public available tools: Cytoscape [109], TargetScan [128], PicTar [129] ,

miRanda [130], PITA [129], ElMMo [131], RNA22 [132] , DIANA-microT [133] and GenMiR++ [134],

The Tumoral Microenvironment Database (TME.db)

TME.db is a Web based application built on a 3-tier architecture which is implemented using the Java2 Enterprise Edition (J2EE) technology and is accessible via a standard web browser. The underlying relational database model is designed as a cancer patient oriented database which takes all the patients anamnesis and clinical and medical history information into account whereby all patients are linked to a specific hospital. The patient information additionally includes personal problems, surgery and detailed cancer information. Additionally the model allows storing a variety of different high-throughput experiments: Flow Cytometric (FACs) phenotyping, proliferation analysis data, Real Time TaqMan qPCR gene expression assay data and Immunohistochemical Tissue Micro Array (TMA) data, Microsatelite Instability (MSI), Single Nucleotide Polymorphism (SNP). Most of these experiments were performed on the available dissected cancer patient tissue samples. TME.db joins and integrates all different types of data analyses and stores them in a common place where all the determined analysis parameters are linked in a clear way dependent on the sample material and the experiment type. For accessing all the stored information again sophisticated query methods were developed in order to retrieve the data in a pre-modified way, already prepared for statistical analysis. The web interface to TME.db also provides a statistical module that connects to customized R services which allow for the automatic testing of normality, calculation of logrank tests and Cox-Regression hazard ratios by using R and Bioconductor packages. Hypergeometric test is used for over-significance calculations. If multiple hypotheses are tested, P-value correction methods are applied. The analysis result includes all raw data files, a description of the methods used and comprehensive tables and graphs created during the analysis. The analysis is fast, reliable and transparent.

Computational analysis to find predicted target genes of miRNAs

Identifying miRNA targets in animals has been very challenging. Many biological features of miRNA targeting have been revealed experimentally and computationally. We divided the miRNA target features into 6 categories; miRNA:mRNA paring, site location, conservation, site accessibility, multiple sites and expression profile [135]. Many target prediction tools have been developed (Tab.4)

Table 4: List of miRNA target prediction tools

Tool	Pair	Site	Consv	Access	Multi	Expr	Link
TargetScan	•	•	•	•	•		http://www.targetscan.org/
PicTar	•		•	•	•		http://pictar.mdc-berlin.de/
miRanda	•		•	•	•		http://www.mirbase.org/ and http://www.microrna.org/
PITA	•		•	•	•		http://genie.weizmann.ac.il/
EIMMo	•		•		•		http://www.mirz.unibas.ch/ElMMo2/
RNA22	•		•	•	•		http://cbcsrv.watson.ibm.com/rna22.html
DIANA-microT	•		•	•			http://diana.cslab.ece.ntua.gr/microT/
GenMiR++	•			•		•	http://www.psi.toronto.edu/genmir/

Correlation of miRNAs and their prediction targets

After we validated the miRNA-mRNA pairs for both the global and stage analysis, we investigated the linear relationship between miRNAs and their high confidence target genes by using Pearson correlation coefficients, $P_{x,y}$, which is defined as:

$$P_{x,y} = \frac{E((X-\mu_x)(Y-\mu_y))}{\sigma_x \sigma_y}$$

where X and Y are miRNA and gene expression patterns, with mean expression value, μ_x and μ_y, respectively, and σ_x and σ_y are the standard deviations.

Statistical analysis

For pairwise comparisons of parametric and non-parametric data the Student's t-test and Wilcoxon rank-sum test were used, respectively. Kaplan Meier estimators of survival were used to visualize the survival curves. Hazard ratio (Cox proportional hazards model) and the logrank test were used to compare disease-free and overall survival between patients in different groups. All through the text a p-value < 0.05 was considered statistically significant. All analyzes were done with the statistical software R (survival package) and Statview.

Network and Landscape Visualization

Cytoscape was used for the creation and visualization of correlation analysis based networks of different datasets. A Cytoscape filtering of the edges with low confidence and positive correlation $R > 0$ was performed. No positive correlation remained after filtering. The organic algorithm that determines the node positions based on their connectivity was used for laying out the network. The color of the nodes is based on different node attributes available for the analyzed dataset: i.e. mean cell count, log2 HR, log2 Maximal HR.

The ClueGO cytoscape plug-in

For an improved biological interpretation of large lists of genes, ClueGO, a Cytoscape plug-in, was developed. ClueGO integrates Gene Ontology (GO) terms as well as KEGG/BioCarta pathways and uses kappa statistics to create a functionally organized GO/pathway term network. A variety of flexible restriction criteria allow for visualizations in different levels of specificity. It can analyze one or compare two lists of genes and comprehensively visualizes functionally grouped terms. A one-click update option allows ClueGO to automatically download the most recent GO/KEGG release at any time. New organisms and ID types can be easily included in a transparent, plug-in like manner. ClueGO provides an intuitive representation of the analysis results and can be optionally used in conjunction with the GOlorize plug-in. In the attempt to define phenoclusters of CRC patients, ClueGO was used to calculate the enrichment in certain cell types of gene clusters of interest. A left sided

(Enrichment) test based on hypergeometric distribution was performed. The enrichment was calculated as reported to the Immunome Ontology that includes the analyzed immune cell types and the corresponding preferentially expressed genes in a flat hierarchy.

Acknowledgements

This thesis is supported in part by the Austrian Federal Ministry for Science and Research (Technology grants), Project Bioinformatics Integration Networks (BIN II-III) and, the COMET Center ONCOTYROL.

I would first of all to thank my major advisor, Prof. Zlatko Trajanoski, gave me the great opportunity to study in Austria and furthermore in France and his guidance during my PhD study. I am forever indebted to my co-advisers, Dr. Jérôme Galon and Asst. Prof. Dr. Hubert Hackl for their advise, suggestions, encouragement, inspiring discussions and in particular, patience. I am gratefully thank my best friends, Dr. Gabriela Bindea and Dr. Bernhard Mlecnik for their suggestions for improving this dissertation, patience, encouragement and especially they always take care of me during living and working and in France.

Special thank from me would highly place on Integrative Cancer Immunology Team (INSERM U872) under supervised by Dr. Jérôme Galon for all kindness in sharing themselves collecting data. The privacy of all subjects has been fully respected throughout the research. I would like to give my warmly thanks to all my colleges and staffs from Institute for Genomics and Bioinformatics, Graz University of Technology and Section for Bioinformatics, Biocenter, Innsbruck Medical University for all the support and company. Above all, I would like to dedicate my work to my beloved family.

References

1. Ricci-Vitiani L, Fabrizi E, Palio E, De Maria R: **Colon cancer stem cells**. *J. Mol. Med* 2009, **87**:1097-1104.

2. Weitz J, Koch M, Debus J, Höhler T, Galle PR, Büchler MW: **Colorectal cancer**. *Lancet* 2005, **365**:153-165.

3. Davies RJ, Miller R, Coleman N: **Colorectal cancer s creening: p rospects fo r molecular stool analysis**. *Nat. Rev. Cancer* 2005, **5**:199-209.

4. Penland SK, Goldberg RM: **Current strategies in previously untreated advanced colorectal cancer**. *Oncology (Williston Park, N.Y.)* 2004, **18**:715-722, 727; discussion 727-729.

5. Midgley RS, Kerr DJ: **ABC of colorectal ca ncer: adjuvant therapy**. *BMJ* 2000, **321**:1208-1211.

6. Tepper JE, O'Connell M, Niedzwiecki D, Hollis DR, Benson AB, Cummings B, Gunderson LL, Macdonald JS, Martenson JA, Mayer RJ: **Adjuvant therapy in rectal cancer: analysis of stage, sex, and local contro l--final report of intergroup 0114**. *J. Clin. Oncol* 2002, **20**:1744-1750.

7. Dukes C: **The classification of cancer of the rectum**. *J Pathol* 1932:35-323.

8. Astler VB CF: **The prognostic significance of di rect extension of carcinoma of the colon and rectum**. *Ann Surg* 1954, **139**:846.

9. Kyriakos M: **The President ca ncer, the D ukes class ification, and confusio n**. *Arch Pathol Lab Med* 1985, **109**:1063.

10. L.H. S, Wittekind C: *TNM classification of malignant tumors*. 6th edition. New York: Wiley-Liss; 2002.

11. Wittekind C, Compton CC, Greene FL, Sobin LH: **TNM residual tumo r classification revisited**. *Cancer* 2002, **94**:2511-2516.

12. Sobin LH: **TNM: evolution and relation to other prognostic factors**. *Semin Surg Oncol* 2003, **21**:3-7.

13. Finn OJ: **Cancer immunology**. *N. Engl. J. Med* 2008, **358**:2704-2715.

14. BURNET M: **Cancer: a biological appro ach. III. Viruses associated w ith neoplastic conditions. IV. Practical applications**. *Br Med J* 1957, **1**:841-847.

15. Kim R, Emi M, Tanabe K: **Cancer immunoediting fro m immune surveillance to immune escape**. *Immunology* 2007, **121**:1-14.

16. Dunn GP, Bruce AT, Ikeda H, Old LJ, Schreiber RD: **Cancer immunoediting: from immunosurveillance to tumor escape**. *Nat. Immunol* 2002, **3**:991-998.

17. Smyth MJ, Dunn GP, Schreiber RD: **Cancer immunosurveillance a nd immunoediting: the ro les o f immunity in suppressing tumor development and shaping tumor immunogenicity**. *Adv. Immunol* 2006, **90**:1-50.

18. Galon J, Costes A, Sanchez-Cabo F, Kirilovsky A, Mlecnik B, Lagorce-Pagès C, Tosolini M, Camus M, Berger A, Wind P, Zinzindohoué F, Bruneval P, Cugnenc P, Trajanoski Z, Fridman W, Pagès F: **Type, density, and location of immune cells within human colorectal tumors predict clinical outcome**. *Science* 2006, **313**:1960-1964.

19. Bindea G, Mlecnik B, Fridman W, Pagès F, Galon J: **Natural immunity to cancer in humans**. *Curr. Opin. Immunol* 2010, **22**:215-222.

20. Campi G, Crosti M, Consogno G, Facchinetti V, Conti-Fine BM, Longhi R, Casorati G, Dellabona P, Protti MP: **CD4(+) T cells fro m healthy subjects an d colon can cer patients recog nize a ca rcinoembryonic antigen-specific immunodominant epitope**. *Cancer Res* 2003, **63**:8481-8486.

21. Schmitt E, Parcellier A, Ghiringhelli F, Casares N, Gurbuxani S, Droin N, Hamai A, Pequignot M, Hammann A, Moutet M, Fromentin A, Kroemer G, Solary E, Garrido C: **Increased immunogenicity of co lon ca ncer cells by selective depletion of cytochrome C**. *Cancer Res* 2004, **64**:2705-2711.

22. Banerjea A, Bustin SA, Dorudi S: **The immunogenicity of colorectal cancers with high-degree microsatellite instability**. *World J Surg Oncol* 2005, **3**:26.

23. Shunyakov L, Ryan CK, Sahasrabudhe DM, Khorana AA: **The influence of host response on colorectal cancer prognosis**. *Clin Colorectal Cancer* 2004, **4**:38-45.

24. Jass JR, Love SB, Northover JM: **A new prognostic classificatio n of recta l cancer**. *Lancet* 1987, **1**:1303-1306.

25. Naito Y, Saito K, Shiiba K, Ohuchi A, Saigenji K, Nagura H, Ohtani H: **CD8+ T cells infiltrated w ithin cancer cell nests as a prognostic fa ctor in human colorectal cancer**. *Cancer Res* 1998, **58**:3491-3494.

26. Diederichsen ACP, Hjelmborg JVB, Christensen PB, Zeuthen J, Fenger C: **Prognostic value of the CD4+/ CD8+ ratio of tumour inf iltrating lymphocytes i n colorectal cancer and HLA-DR expression o n tumour cells** . *Cancer Immunol. Immunother* 2003, **52**:423-428.

27. Ropponen KM, Eskelinen MJ, Lipponen PK, Alhava E, Kosma VM: **Prognostic value of tu mour-infiltrating lympho cytes (TILs) in co lorectal ca ncer**. *J. Pathol* 1997, **182**:318-324.

28. Pagès F, Berger A, Camus M, Sanchez-Cabo F, Costes A, Molidor R, Mlecnik B, Kirilovsky A, Nilsson M, Damotte D, Meatchi T, Bruneval P, Cugnenc P, Trajanoski Z, Fridman W, Galon J: **Effector memory T cells, early metastasis, and survival in colorectal cancer**. *N. Engl. J. Med* 2005, **353**:2654-2666.

29. Waldner M, Schimanski C, Neurath M: **Colon cancer and the immune system: the role of tumor invading T cells**. *World J. Gastroenterol* 2006, **12**:7233-7238.

30. Cheng AM, Byrom MW, Shelton J, Ford LP: **Antisense inhibition of human miRNAs and indications for an involvement of miRNA in cell growth and apoptosis**. *Nucleic Acids Res* 2005, **33**:1290-1297.

31. Subramanian S, Steer CJ: **MicroRNAs as gatekeepers of apoptosis**. *J. Cell. Physiol* 2010, **223**:289-298.

32. Karp X, Ambros V: **Developmental biology. Encountering microRNAs in cell fate signaling**. *Science* 2005, **310**:1288-1289.

33. Chen C, Li L, Lodish HF, Bartel DP: **MicroRNAs modulate hematopoietic lineage differentiation**. *Science* 2004, **303**:83-86.

34. Poy MN, Eliasson L, Krutzfeldt J, Kuwajima S, Ma X, Macdonald PE, Pfeffer S, Tuschl T, Rajewsky N, Rorsman P, Stoffel M: **A pancreatic islet-specific microRNA regulates insulin secretion**. *Nature* 2004, **432**:226-230.

35. Ikeda S, Kong SW, Lu J, Bisping E, Zhang H, Allen PD, Golub TR, Pieske B, Pu WT: **Altered microRNA expression in human heart disease**. *Physiol. Genomics* 2007, **31**:367-373.

36. Eisenberg I, Eran A, Nishino I, Moggio M, Lamperti C, Amato AA, Lidov HG, Kang PB, North KN, Mitrani-Rosenbaum S, Flanigan KM, Neely LA, Whitney D, Beggs AH, Kohane IS, Kunkel LM: **Distinctive patterns of microRNA expression in primary muscular disorders**. *Proc. Natl. Acad. Sci. U.S.A* 2007, **104**:17016-17021.

37. Cheng Y, Ji R, Yue J, Yang J, Liu X, Chen H, Dean DB, Zhang C: **MicroRNAs are aberrantly expressed in hypertrophic heart: do they play a role in cardiac hypertrophy?** *Am. J. Pathol* 2007, **170**:1831-1840.

38. Ikeda S, Pu WT: **Expression and function of microRNAs in heart disease**. *Curr Drug Targets* 2010, **11**:913-925.

39. Schaefer A, Stephan C, Busch J, Yousef GM, Jung K: **Diagnostic, prognostic and therapeutic implications of microRNAs in urologic tumors**. *Nat Rev Urol* 2010, **7**:286-297.

40. Calin GA, Croce CM: **Investigation of microRNA alterations in leukemias and lymphomas**. *Meth. Enzymol* 2007, **427**:193-213.

41. Nicoloso MS, Calin GA: **MicroRNA involvement in brain tumors: from bench to bedside**. *Brain Pathol* 2008, **18**:122-129.

42. Calin GA, Croce CM: **MicroRNA signatures in human cancers**. *Nat. Rev. Cancer* 2006, **6**:857-866.

43. Iorio MV, Visone R, Di Leva G, Donati V, Petrocca F, Casalini P, Taccioli C, Volinia S, Liu C, Alder H, Calin GA, Ménard S, Croce CM: **MicroRNA signatures in human ovarian cancer**. *Cancer Res* 2007, **67**:8699-8707.

44. Barbarotto E, Schmittgen TD, Calin GA: **MicroRNAs and ca ncer: profile, profile, profile**. *Int. J. Cancer* 2008, **122**:969-977.

45. Sonkoly E, Ståhle M, Pivarcsi A: **MicroRNAs and immunity: novel players in the regulation of normal imm une function an d inflammation**. *Semin. Cancer Biol* 2008, **18**:131-140.

46. Lee YS, Dutta A: **MicroRNAs in cancer**. *Annu Rev Pathol* 2009, **4**:199-227.

47. Fabbri M, Croce CM, Calin GA: **MicroRNAs in the ontogeny of leukemias and lymphomas**. *Leuk. Lymphoma* 2009, **50**:160-170.

48. Rossi S, Kopetz S, Davuluri R, Hamilton SR, Calin GA: **MicroRNAs, ultraconserved genes and colorectal cancers** . *Int. J. Biochem. Cell Biol* 2010, **42**:1291-1297.

49. Buckland J: **Biomarkers: microRNAs un der the spotlight in inflammatory arthritis**. *Nat Rev Rheumatol* 2010, **6**:436.

50. Polikepahad S, Knight JM, Naghavi AO, Oplt T, Creighton CJ, Shaw C, Benham AL, Kim J, Soibam B, Harris RA, Coarfa C, Zariff A, Milosavljevic A, Batts LM, Kheradmand F, Gunaratne PH, Corry DB: **Pro-inflammatory r ole fo r le t-7 microRNAs in experimental asthma**. *J Biol Chem* 2010.

51. Xie L, Qian X, Liu B: **MicroRNAs: nov el biomarkers for gastrointestinal carcinomas**. *Mol. Cell. Biochem* 2010, **341**:291-299.

52. Wightman B, Ha I, Ruvkun G: **Posttranscriptional regulation of the heterochronic gene lin-14 by lin-4 mediat es temporal pattern for mation in C. elegans**. *Cell* 1993, **75**:855-862.

53. Filipowicz W, Bhattacharyya SN, Sonenberg N: **Mechanisms of pos t-transcriptional regulation by micr oRNAs: are the answers in sight?** *Nat. Rev. Genet* 2008, **9**:102-114.

54. Pillai RS, Bhattacharyya SN, Filipowicz W: **Repression of protein synthesis by miRNAs: how many mechanisms?** *Trends Cell Biol* 2007, **17**:118-126.

55. Bartel DP: **MicroRNAs: genomics, biogenesi s, mechanism, and function** . *Cell* 2004, **116**:281-297.

56. Lee RC, Feinbaum RL, Ambros V: **The C. elegans heterochron ic gene lin-4 encodes small RNAs with antise nse complementarity to lin-14** . *Cell* 1993, **75**:843-854.

57. Zhang B, Wang Q, Pan X: **MicroRNAs and their regulatory roles in animals and plants**. *J. Cell. Physiol* 2007, **210**:279-289.

58. Tsitsiou E, Lindsay MA: **microRNAs a nd the immune response** . *Curr Opin Pharmacol* 2009, **9**:514-520.

59. Xiao C, Rajewsky K: **MicroRNA contr ol in the immune system: basic principles**. *Cell* 2009, **136**:26-36.

60. Pauley KM, Chan EKL: **MicroRNAs and their emerging roles in immunology** . *Ann. N. Y. Acad. Sci* 2008, **1143**:226-239.

61. Davidson-Moncada J, Papavasiliou FN, Tam W: **MicroRNAs of the immune system: ro les in in flammation a nd cancer** . *Ann. N. Y. Acad. Sci* 2010, **1183**:183-194.

62. Wei B, Pei G: **microRNAs: critical regulators in Th17 cells and players in diseases**. *Cell. Mol. Immunol* 2010, **7**:175-181.

63. Carissimi C, Fulci V, Macino G: **MicroRNAs: novel regulators of immunity** . *Autoimmun Rev* 2009, **8**:520-524.

64. Garzon R, Calin GA, Croce CM: **MicroRNAs in Ca ncer**. *Annu. Rev. Med* 2009, **60**:167-179.

65. Lee Y, Ahn C, Han J, Choi H, Kim J, Yim J, Lee J, Provost P, Rådmark O, Kim S, Kim VN: **The nuclear RNase III Drosha initiates microRNA processing**. *Nature* 2003, **425**:415-419.

66. Denli AM, Tops BBJ, Plasterk RHA, Ketting RF, Hannon GJ: **Processing o f primary microRNAs by the Microprocessor complex**. *Nature* 2004, **432**:231-235.

67. Yi R, Qin Y, Macara IG, Cullen BR: **Exportin-5 mediates the nuclear export of pre-microRNAs and short hairpin RNAs**. *Genes Dev* 2003, **17**:3011-3016.

68. Zhang H, Kolb FA, Jaskiewicz L, Westhof E, Filipowicz W: **Single processing center models for human Dicer and bacterial RNase III**. *Cell* 2004, **118**:57-68.

69. Lippolis JD: **Immunological signaling netw orks: integratin g the body's immune response**. *J. Anim. Sci* 2008, **86**:E53-63.

70. Janeway CA Jr, Travers P, Walport M, Shlomchik MJ: *Immunobiology*. 6th edition. Garland Publishing; 2005.

71. Okada H, Kohanbash G, Lotze MT: **MicroRNAs in immune regulation--opportunities for cancer immunotherapy**. *Int. J. Biochem. Cell Biol* 2010, **42**:1256-1261.

72. Xiao C, Rajewsky K: **MicroRNA contr ol in the immune system: basic principles**. *Cell* 2009, **136**:26-36.

73. Tsitsiou E, Lindsay MA: **microRNAs a nd the immune response** . *Curr Opin Pharmacol* 2009, **9**:514-520.

74. Sonkoly E, Ståhle M, Pivarcsi A: **MicroRNAs and immunity: novel players in the regulation of normal imm une function an d inflamm ation**. *Semin. Cancer Biol* 2008, **18**:131-140.

75. Pauley KM, Chan EKL: **MicroRNAs and their emerging roles in immunology** . *Ann. N. Y. Acad. Sci* 2008, **1143**:226-239.

76. Davidson-Moncada J, Papavasiliou FN, Tam W: **MicroRNAs of the immune system: ro les in in flammation a nd cancer** . *Ann. N. Y. Acad. Sci* 2010, **1183**:183-194.

77. Carissimi C, Fulci V, Macino G: **MicroRNAs: novel regulators of immunity** . *Autoimmun Rev* 2009, **8**:520-524.

78. Yang L, Belaguli N, Berger DH: **MicroRNA a nd col orectal cancer** . *World J Surg* 2009, **33**:638-646.

79. Calin GA, Croce CM: **MicroRNA signatures in human cancers** . *Nat. Rev. Cancer* 2006, **6**:857-866.

80. Calin GA, Croce CM: **MicroRNA-cancer connection: the beg inning of a new tale**. *Cancer Res* 2006, **66**:7390-7394.

81. Subramanian S, Steer CJ: **MicroRNAs a s gatek eepers of apoptosis** . *J. Cell. Physiol* 2010, **223**:289-298.

82. Buckland J: **Biomarkers: microRNAs un der the spotlight in inflammatory arthritis**. *Nat Rev Rheumatol* 2010, **6**:436.

83. Lodish HF, Zhou B, Liu G, Chen C: **Micromanagement of the immune syste m by microRNAs**. *Nat. Rev. Immunol* 2008, **8**:120-130.

84. Barbarotto E, Schmittgen TD, Calin GA: **MicroRNAs and ca ncer: profile, profile, profile**. *Int. J. Cancer* 2008, **122**:969-977.

85. Bloomston M, Frankel WL, Petrocca F, Volinia S, Alder H, Hagan JP, Liu C, Bhatt D, Taccioli C, Croce CM: **MicroRNA expression patterns to differentiate pancreatic adenocarcinoma from norm al pancreas and chronic pancreatitis** . *JAMA* 2007, **297**:1901-1908.

86. Sasaki K, Kohanbash G, Hoji A, Ueda R, McDonald HA, Reinhart TA, Martinson J, Lotze MT, Marincola FM, Wang E, Fujita M, Okada H: **miR-17-92 expression in differentiated T cells - implic ations for cancer immuno therapy**. *J Transl Med* 2010, **8**:17.

87. Hurst DR, Edmonds MD, Welch DR: **Metastamir: the field of metas tasis-**

regulatory microRNA is spreading. *Cancer Res* 2009, **69**:7495-7498.

88. Akao Y, Nakagawa Y, Naoe T: **MicroRNA-143 and -145 in colon cancer** . *DNA Cell Biol* 2007, **26**:311-320.

89. Eisenberg I, Eran A, Nishino I, Moggio M, Lamperti C, Amato AA, Lidov HG, Kang PB, North KN, Mitrani-Rosenbaum S, Flanigan KM, Neely LA, Whitney D, Beggs AH, Kohane IS, Kunkel LM: **Distinctive patterns of microRNA express ion in primary muscular disorders** . *Proc. Natl. Acad. Sci. U.S.A* 2007, **104**:17016-17021.

90. Lu TX, Munitz A, Rothenberg ME: **MicroRNA-21 is up-regulated in allergic airway inflammation and regulates IL-12p 35 expression** . *J. Immunol* 2009, **182**:4994-5002.

91. Zhang B, Pan X, Cobb GP, Anderson TA: **microRNAs as oncogenes and tumor suppressors**. *Dev. Biol* 2007, **302**:1-12.

92. Shenouda SK, Alahari SK: **MicroRNA function in cancer: on cogene or a tumor suppressor?** *Cancer Metastasis Rev* 2009, **28**:369-378.

93. Esquela-Kerscher A, Slack FJ: **Oncomirs - microRNAs with a role in cancer** . *Nat. Rev. Cancer* 2006, **6**:259-269.

94. Cho WCS: **OncomiRs: the dis covery and progress o f microRNAs in cancers** . *Mol. Cancer* 2007, **6**:60.

95. Lu J, Getz G, Miska EA, Alvarez-Saavedra E, Lamb J, Peck D, Sweet-Cordero A, Ebert BL, Mak RH, Ferrando AA, Downing JR, Jacks T, Horvitz HR, Golub TR: **MicroRNA expression profiles classify human cancers**. *Nature* 2005, **435**:834-838.

96. Bloomston M, Frankel WL, Petrocca F, Volinia S, Alder H, Hagan JP, Liu C, Bhatt D, Taccioli C, Croce CM: **MicroRNA expression patterns to differentiate pancreatic adenocarcinoma from norm al pancreas and chronic pancreatitis** . *JAMA* 2007, **297**:1901-1908.

97. Schepeler T, Reinert JT, Ostenfeld MS, Christensen LL, Silahtaroglu AN, Dyrskjøt L, Wiuf C, Sørensen FJ, Kruhøffer M, Laurberg S, Kauppinen S, Ørntoft TF, Andersen CL: **Diagnostic and prognostic microRNAs in stage II colon cancer** . *Cancer Res* 2008, **68**:6416-6424.

98. Li W, Xie L, He X, Li J, Tu K, Wei L, Wu J, Guo Y, Ma X, Zhang P, Pan Z, Hu X, Zhao Y, Xie H, Jiang G, Chen T, Wang J, Zheng S, Cheng J, Wan D, Yang S, Li Y, Gu J: **Diagnostic and prognostic implicat ions of microRNAs in human hepatocellular carcinoma**. *Int. J. Cancer* 2008, **123**:1616-1622.

99. G P George, R D Mittal: **MicroRNAs: potential biomarkers in cancer** . *Ind J Clin Biochem* 2010, **25**:4-14.

100. Koturbash I, Zemp FJ, Pogribny I, Kovalchuk O: **Small molec ules w ith big**

effects: The role of the microRNAome in cancer and carcinogenesis . *Mutat Res* 2010.

101. Ueda R, Kohanbash G, Sasaki K, Fujita M, Zhu X, Kastenhuber ER, McDonald HA, Potter DM, Hamilton RL, Lotze MT, Khan SA, Sobol RW, Okada H: **Dicer-regulated microRNAs 222 and 339 promote resistance of cancer cells to cytotoxic T-lymphocytes by down-regulation of ICAM-1** . *Proc. Natl. Acad. Sci. U.S.A* 2009, **106**:10746-10751.

102. Michael MZ, O' Connor SM, van Holst Pellekaan NG, Young GP, James RJ: **Reduced accumulation of specific microRNAs in colorectal neoplasia**. *Mol. Cancer Res* 2003, **1**:882-891.

103. Akao Y, Nakagawa Y, Naoe T: **let-7 microRNA functions as a potential growth suppressor in human colon cancer cells** . *Biol. Pharm. Bull* 2006, **29**:903-906.

104. Yang L, Belaguli N, Berger DH: **MicroRNA and colorectal cancer** . *World J Surg* 2009, **33**:638-646.

105. Faber C, Kirchner T, Hlubek F: **The impact of micro RNAs on colorectal cancer**. *Virchows Arch* 2009, **454**:359-367.

106. Sturn A, Quackenbush J, Trajanoski Z: **Genesis: cluster analysis of microarray data**. *Bioinformatics* 2002, **18**:207-208.

107. Sturn A, Quackenbush J, Trajanoski Z: **Genesis: cluster analysis of microarray data**. *Bioinformatics* 2002, **18**:207-208.

108. Garzon R, Marcucci G, Croce CM: **Targeting microRNAs in cancer: rationale, strategies and challenges**. *Nat Rev Drug Discov* 2010, **9**:775-789.

109. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T: **Cytoscape: a software environment for integrated models of biomolecular interaction networks**. *Genome Res* 2003, **13**:2498-2504.

110. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G: **Gene ontology: tool for the unification of biology. The Gene Ontology Consortium** . *Nat. Genet* 2000, **25**:25-29.

111. Bindea G, Mlecnik B, Hackl H, Charoentong P, Tosolini M, Kirilovsky A, Fridman W, Pagès F, Trajanoski Z, Galon J: **ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks**. *Bioinformatics* 2009, **25**:1091-1093.

112. Schaefer A, Jung M, Mollenkopf H, Wagner I, Stephan C, Jentzmik F, Miller K, Lein M, Kristiansen G, Jung K: **Diagnostic and prognostic implications of microRNA profiling in prostate carcinoma**. *Int. J. Cancer* 2010, **126**:1166-1176.

113. Yanaihara N, Caplen N, Bowman E, Seike M, Kumamoto K, Yi M, Stephens RM, Okamoto A, Yokota J, Tanaka T, Calin GA, Liu C, Croce CM, Harris CC: **Unique microRNA molecular profiles in lung cancer diagnosis and prognosis**. *Cancer Cell* 2006, **9**:189-198.

114. Chen Y, Stallings RL: **Differential patterns of microRNA expression in neuroblastoma are correlated with prognosis, differentiation, and apoptosis**. *Cancer Res* 2007, **67**:976-983.

115. Heim S, Mitelman F: **Primary chromosome abnormalities in human neoplasia**. *Adv. Cancer Res* 1989, **52**:1-43.

116. Geigl JB, Speicher MR: **Single-cell isolation from cell suspensions and whole genome amplification from single cells to provide templates for CGH analysis**. *Nat Protoc* 2007, **2**:3173-3184.

117. van Beers EH, Nederlof PM: **Array-CGH and breast cancer**. *Breast Cancer Res* 2006, **8**:210.

118. Pinkel D, Albertson DG: **Array comparative genomic hybridization and its applications in cancer**. *Nat. Genet* 2005, **37 Suppl**:S11-17.

119. Mlecnik B, Tosolini M, Charoentong P, Kirilovsky A, Bindea G, Berger A, Camus M, Gillard M, Bruneval P, Fridman W, Pagès F, Trajanoski Z, Galon J: **Biomolecular network reconstruction identifies T-cell homing factors associated with survival in colorectal cancer**. *Gastroenterology* 2010, **138**:1429-1440.

120. Xin H, Kikuchi T, Andarini S, Ohkouchi S, Suzuki T, Nukiwa T, Huqun, Hagiwara K, Honjo T, Saijo Y: **Antitumor immune response by CX3 CL1 fractalkine gene transfer depends on both NK and T cells**. *Eur. J. Immunol* 2005, **35**:1371-1380.

121. Schee K, Fodstad Ø, Flatmark K: **MicroRNAs as biomarkers in colorectal cancer**. *Am. J. Pathol* 2010, **177**:1592-1599.

122. Lu J, Getz G, Miska EA, Alvarez-Saavedra E, Lamb J, Peck D, Sweet-Cordero A, Ebert BL, Mak RH, Ferrando AA, Downing JR, Jacks T, Horvitz HR, Golub TR: **MicroRNA expression profiles classify human cancers**. *Nature* 2005, **435**:834-838.

123. Volinia S, Calin GA, Liu C, Ambs S, Cimmino A, Petrocca F, Visone R, Iorio M, Roldo C, Ferracin M, Prueitt RL, Yanaihara N, Lanza G, Scarpa A, Vecchione A, Negrini M, Harris CC, Croce CM: **A microRNA expression signature of human solid tumors defines cancer gene targets**. *Proc. Natl. Acad. Sci. U.S.A* 2006, **103**:2257-2261.

124. Hayashita Y, Osada H, Tatematsu Y, Yamada H, Yanagisawa K, Tomida S, Yatabe Y, Kawahara K, Sekido Y, Takahashi T: **A polycistronic microRNA cluster, miR-17-92, is overexpressed in human lung cancers and enhances cell proliferation**. *Cancer Res* 2005, **65**:9628-9632.

125. Matsubara H, Takeuchi T, Nishikawa E, Yanagisawa K, Hayashita Y, Ebi H, Yamada H, Suzuki M, Nagino M, Nimura Y, Osada H, Takahashi T: **Apoptosis induction by antisense oligonucleotides against miR-17-5p and miR-20a in lung cancers overexpressing miR-17-92**. *Oncogene* 2007, **26**:6099-6105.

126. Bagga S, Bracht J, Hunter S, Massirer K, Holtz J, Eachus R, Pasquinelli AE: **Regulation by let-7 and lin-4 miRNAs results in target mRNA deg radation**. *Cell* 2005, **122**:553-563.

127. Cimmino A, Calin GA, Fabbri M, Iorio MV, Ferracin M, Shimizu M, Wojcik SE, Aqeilan RI, Zupo S, Dono M, Rassenti L, Alder H, Volinia S, Liu C, Kipps TJ, Negrini M, Croce CM: **miR-15 and miR-16 in duce apoptosis by targeting BCL2**. *Proc. Natl. Acad. Sci. U.S.A* 2005, **102**:13944-13949.

128. Grimson A, Farh KK, Johnston WK, Garrett-Engele P, Lim LP, Bartel DP: **MicroRNA targeting specificity in mammals : determinants beyond seed pairing**. *Mol. Cell* 2007, **27**:91-105.

129. Krek A, Grün D, Poy MN, Wolf R, Rosenberg L, Epstein EJ, MacMenamin P, da Piedade I, Gunsalus KC, Stoffel M, Rajewsky N: **Combinatorial microRNA target predictions**. *Nat. Genet* 2005, **37**:495-500.

130. Griffiths-Jones S, Grocock RJ, van Dongen S, Bateman A, Enright AJ: **miRBase: microRNA sequences, ta rgets and gene nomenclature**. *Nucleic Acids Res* 2006, **34**:D140-144.

131. Gaidatzis D, van Nimwegen E, Hausser J, Zavolan M: **Inference of miRNA targets using evolutionary conservation and pathw ay analysis**. *BMC Bioinformatics* 2007, **8**:69.

132. Miranda KC, Huynh T, Tay Y, Ang Y, Tam W, Thomson AM, Lim B, Rigoutsos I: **A pattern-based method for the identi fication of MicroRNA bind ing sites and their corresponding heteroduplexes**. *Cell* 2006, **126**:1203-1217.

133. Maragkakis M, Reczko M, Simossis VA, Alexiou P, Papadopoulos GL, Dalamagas T, Giannopoulos G, Goumas G, Koukis E, Kourtis K, Vergoulis T, Koziris N, Sellis T, Tsanakas P, Hatzigeorgiou AG: **DIANA-microT web server: elucidating microRN A function s through target prediction**. *Nucleic Acids Res* 2009, **37**:W273-276.

134. Huang JC, Babak T, Corson TW, Chua G, Khan S, Gallie BL, Hughes TR, Blencowe BJ, Frey BJ, Morris QD: **Using expression p rofiling data to identify human microRNA targets**. *Nat. Methods* 2007, **4**:1045-1049.

135. Saito T, Saetrom P: **MicroRNAs--targeting an d target prediction**. *N Biotechnol* 2010, **27**:243-249.

Pharmacology & Therapeutics 128 (2010) 488–498

Contents lists available at ScienceDirect

Pharmacology & Therapeutics

journal homepage: www.elsevier.com/locate/pharmthera

Associate Editor: G. Eisenhofer

Information technology solutions for integration of biomolecular and clinical data in the identification of new cancer biomarkers and targets for therapy

Hubert Hackl [a], Gernot Stocker [a], Pornpimol Charoentong [a], Bernhard Mlecnik [b], Gabriela Bindea [b], Jerome Galon [b,*], Zlatko Trajanoski [a,*]

[a] Biocenter, Section for Bioinformatics, Innsbruck Medical University, Schöpfstrasse 45, 6020, Innsbruck, Austria
[b] INSERM, Integrative Cancer Immunology Team, INSERM U872, Paris, France

ARTICLE INFO

Keywords:
Information technology
Data integration
Cancer
Networks
Modelling

ABSTRACT

The quest for new cancer biomarkers and targets for therapy requires not only the aggregation and analysis of heterogeneous biomolecular data but also integration of clinical data. In this review we highlight information technology solutions for the integration of biomolecular and clinical data and focus on a solution at the departmental level, i.e., decentralized and medium-scale solution for groups of labs working on a specific topic. Both, hardware and software requirements are described as well as bioinformatics methods and tools for the data analysis. The highlighted IT solutions include storage architecture, high-performance computing, and application servers. Additionally, following computational approaches for data integration are reviewed: data aggregation, integrative data analysis including methodological aspects as well as examples, biomolecular pathways and network reconstruction, and mathematical modelling. Finally, a case study in cancer immunology including the used computational methods is shown, demonstrating how IT solutions for integrating biomolecular and clinical data can help to identify new cancer biomarkers for improving diagnosis and predicting clinical outcome.

Contents

1. Introduction

Many developments have occurred in prevention and treatment of cancer, but death from this disease is still common. Of the 58 million people who died worldwide in 2005, 7.6 million died of cancer. Based on projections, cancer deaths will continue to rise with an estimated 9 million people dying from cancer in 2015, and 11.4 million dying in 2030. Despite extensive characterization of environmental, intrinsic and underlying mechanisms (Hanahan & Weinberg, 2000), markers of the oncogenic process remain so far poorly predictive of patient survival and fail to prove their reliability in clinical use.

Genetic and molecular tumor prognostic factors have been proposed to identify patients who may be at risk for recurrence. None has yet been sufficiently informative for inclusion in clinical practice (Locker et al., 2006). Identification of patients with a high-risk of recurrence is therefore a major clinical issue. However, in order to develop stratified or personalized strategies for such complex multifactorial disease it is of importance to understand how

* Corresponding authors. Galon is to be contacted at INSERM, Integrative Cancer Immunology Team, INSERM U872, Cordeliers Research Center, 15 rue de l'Ecole de Médecine, 75006 Paris, France. Trajanoski is to be contacted at Biocenter, Section for Bioinformatics, Innsbruck Medical University, Schöpfstrasse 45, 6020 Innsbruck, Austria Trajanoski, Tel.: +43 512 9003 71401; fax: +43 512 9003 74400.
E-mail addresses: jerome.galon@crc.jussieu.fr (J. Galon), zlatko.trajanoski@i-med.ac.at (Z. Trajanoski).

0163-7258/$ – see front matter © 2010 Published by Elsevier Inc.
doi:10.1016/j.pharmthera.2010.08.012

numerous and diverse elements function together in human pathology. With the advance of new technologies including high-throughput techniques for DNA sequencing, RNA expression profiling, protein quantification, multiplexed immunohistochemistry (tissue microarrays), cell sorting and analyses, we have now the means to comprehensively analyze cells and tissues for various biomolecules and identify new cancer biomarkers and targets for therapy.

There has been a growing desire to integrate these high-throughput data sets and make them publicly available. Considerable efforts were undertaken to integrate specific data types into a centralized database. For example gene expression data from heterogeneous platforms can be stored and retrieved in GEO (Edgar et al., 2002) or ArrayExpress (Parkinson et al., 2007), whereas PRIDE was designed for proteomics data (Martens et al., 2005) turning publicly available data into publicly accessible data. While these repositories which integrate heterogeneous data for single data type are of great value for the research community, it soon became clear that the ability to integrate data from multiple sources is becoming critical (Chaussabel et al., 2009).

The quest for new cancer biomarkers and targets for therapy requires not only the aggregation and analysis of heterogeneous biomolecular data but also the integration of clinical data with all relevant parameters (e.g., tumor staging, treatment, and cancer relapse). Recently, the National Institutes of Health (NIH) launched The Cancer Genome Atlas (TCGA) pilot project to integrate clinical data and high-throughput data for three tumor types: glioblastoma, ovarian (serous cystadenocarcinoma) and lung (squamous carcinoma). Encouraged by the broad use of publicly available data sets, the scope was expanded to include more than 20 tumor types and thousands of samples over the next 5 years. These systematic efforts to generate reference data sets are tremendously helpful. In combination with targeted research addressing cancer subgroups, employing additional large-, medium-, and small-scale techniques will facilitate the identification of new biomarkers.

In order to fully exploit the available TCGA reference data sets and newly generated data sets novel tools for integrative analysis and visualization of the biomolecular and clinical data are required. While the need for data integration has been appreciated in the past (Searls, 2005), specific solutions are rare and focus either on particular methodological aspects or technical solutions. Due to the complexity of the problem as well as logistical and legal issues, there are no out-of-the-box packages but rather customized and seldom transferable solutions. Additional difficulties are imposed by the fast development of information technology and short half-life time of software applications.

In general, there are three possibilities for addressing the problem depending on the size of the available resources: 1) centralized IT solution with customized hardware and software as available for large academic institutions (e.g., several hundred PIs) or pharmaceutical companies (see for example Mathew et al., 2007; Waller et al., 2007), 2) decentralized and medium-scale departmental solutions for groups of labs working on a specific topic, and 3) small scale solutions for individual research labs. In this paper, we present IT solutions for the integration of biomolecular and clinical data for the identification of new cancer biomarkers and targets for therapy at the departmental level. Additionally, we review the computational methods including data integration, data analysis, visualization and mathematical modelling which can be used alone or in combination at the departmental level or in single research labs.

2. IT solutions

The search for biomarkers and new targets for complex diseases in general and for cancer in particular requires several data sources (Fig. 1). As of today, the majority of the large-scale technologies like transcriptomics or proteomics deliver data that can be archived with reasonable resources. There is a common sense in the community that archiving data from next generation sequencing (or deep sequencing) instruments is best done using inexpensive hard disks. In both cases, raw data are preprocessed and imported into a data warehouse, i.e., a repository of stored data.

2.1. Data warehouse

The central process is data warehousing which includes three steps: 1) to develop a unified model that can accommodate all the information from single databases, 2) transformation of the data into this model and loading to the data warehouse, and 3) retrieval of data from the source databases in one environment ('one-stop shop') as well as integrated access to the data requiring knowledge, that the individual sources cannot provide (Stein, 2003). This concept has been applied for a long time to many different (business) applications and also found its way into biology (Ritter et al., 1994; Schonbach et al., 2000; Hu et al., 2004; Kasprzyk et al., 2004; Shah et al., 2005; Lee et al., 2006; Rhodes et al., 2007). For an in-depth overview of data warehouse technology we refer the reader to Chaudhuri and Dayal (1997), Devlin (1997), and Inmon (2002).

A centralized solution in which a data warehouse houses all relevant databases including public sources (e.g., Shah et al., 2005), interfaces to laboratory management systems, and holds patient records has several advantages. First, the solution can be customized to specific needs. Second, it can be optimized so that performance can be increased. And third, maintenance and adaptation can be done more effectively. However, such a solution requires a long planning and testing phase (e.g., several person-years), expensive installation and operation. For example interfacing laboratory information management systems (LIMS) is difficult (Stephan et al., 2010). In general, these sophisticated systems are able to manage and analyze data generated for only a single type or a limited number of instruments, and were designed for only one type of data (Maurer et al., 2005; Hartler et al., 2007). Thus, addressing a biological question relying on several complementary technologies requires a specific off-the-shelf database. It should be noted that such a database could absorb several person-years of software engineering and this effort tends to be underestimated. Thus, only large institutions can afford this type of solution.

At the departmental level a preferable setting is a local database hosting only the necessary data. In this case primary data are archived at separate locations. Only preprocessed and normalized data are stored in a dedicated database. Although it is tempting to upload and analyze all types of data in a single system, experience shows that primary data are mostly used once. This approach is even more advisable for large-scale data including microarrays, proteomics or sequence data. However, links to the primary data need to be secured so that later re-analyses using improved tools can be guaranteed. In this context it is noteworthy that in the majority of published studies the analyses were based on medium-throughput data, meaning that the number of analyzed molecular species was in the range of 100–1000 (after filtering and preselection). With this number of elements the majority of the tools perform satisfactorily on a standard desktop computer. Performance is a crucial issue if the number of molecules detected in a single patient sample increases to >10,000 i.e., in microarray studies or >100,000 (proteomics studies). In this situation the methods and the IT infrastructure require re-evaluation.

Incorporation of clinical data into the data warehouse (Hu et al., 2004) poses major challenges. Many institutions have electronic patient records and in principle, extracting the information could be straightforward. However, technical, ethical, and legal issues might delay or even prohibit the process of data collection. Heterogeneous clinical and departmental information systems, accessibility of patient data, and managing sensitive information can introduce several levels of complexity and require extensive stakeholder discussions.

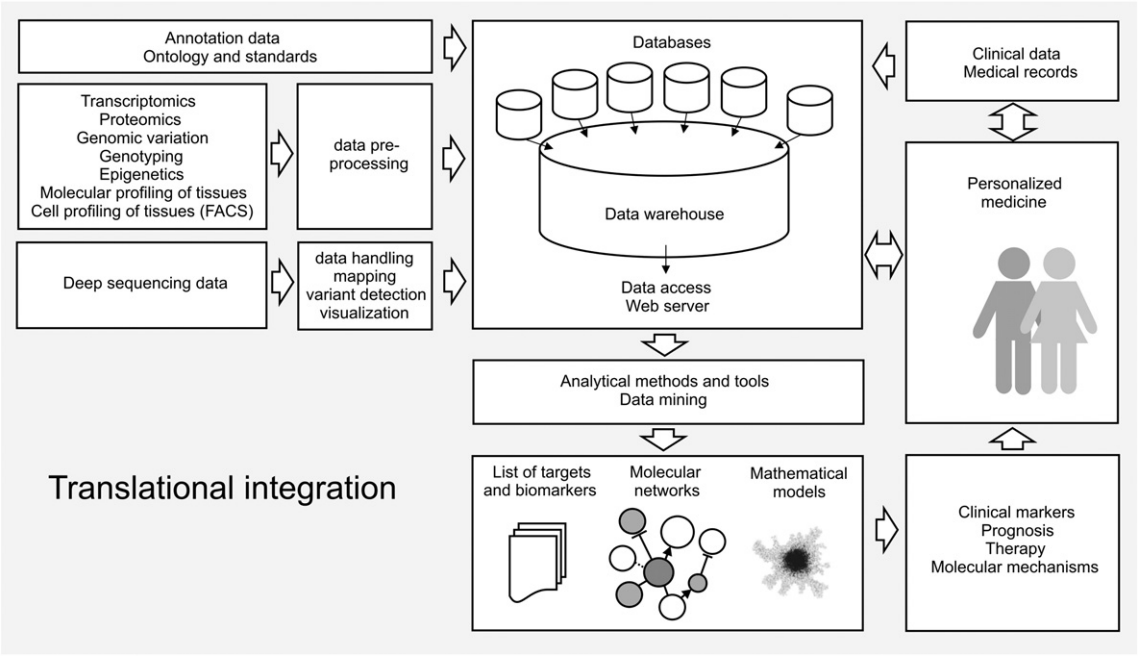


Fig. 1. Data flow in translational cancer research. Shown are the heterogeneous data sources from which large-scale data is generated. These data sets have to be integrated with clinical data and medical records. Applying analytical methods on the integrated data can provide list(s) of targets and biomarkers, highlight deregulated pathways, or establish predictive models, which can be then applied to treat patients individually.

Therefore the preferable IT solution for a department is based on a relatively small database for only a few specific cohorts. The patient data should be first de-identified, entered into the database, and then provided to the biologists and bioinformaticians.

The data warehouse depicted in Fig. 1 requires state-of-the-art IT infrastructure. IT basis for consecutive data management, data analysis and data integration consists of a reliable computational environment which can handle the storage of diverse data, deliver data to the computing cluster and provide raw and processed data to the end user. The computational infrastructure of a bioinformatics environment can be divided into three corner stones: 1) storage architecture, 2) high-performance computing infrastructure, and 3) application servers (Fig. 2).

2.2. Storage architecture

Over the last decade bioinformaticians have faced a considerable challenge of storage and processing of huge amounts of biological data. With the advance of next generation sequencing the data flood rose into the range of petabytes (1000 terabytes). Hence, appropriate storage solutions are of utmost importance. Storage devices can be categorized in several types (see Box 1).

In addition to the hardware requirements, the choice of a database management system is crucial. In the biological context it is necessary to differentiate between flat file databases and relational databases. Flat file databases are frequently used to exchange and archive biological information like sequences and annotations in a structured form using either binary or text format. The format can vary from tab delimited single line entries over text delimited multi-line entries to non-relational data stores termed NoSQL.

In contrast, relational database and the associated management system (RDBMS) is a software system which stores attributes and their associations in a non-redundant – normalized – manner. Connected attributes are grouped in tables and can be accessed using structured query language (SQL). In data warehouse systems the degree of normalization can be decreased in favor of improved performance. RDBMS in productive environments are usually installed on a central database server, which has sufficiently dimensioned storage. Often, this server acts as backend for remote applications.

2.3. High-performance computing infrastructure

The increasing need for intensive calculations is met by processor manufacturers by switching from single-processor/single-core architecture to multi-processor/multi-core architecture in one single computer. This paradigm shift away from the speed race for single processors towards processing in parallel using multiple processors influenced also the development of bioinformatics applications. The hardware used for HPC can significantly vary in terms of application, performance, and flexibility (Box 2).

The choice of the appropriate HPC environment depends on many parameters including available resources, expertise, amount of data generated, type of number-crunching applications, requirements concerning the data availability, or security issues. But the choice depends also on the biological applications, the nature of the data as well as the utilized algorithms, i.e., the question is if the problem can be parallelized (Stein, 2008; Schadt et al., 2010). E.g. if genetic associations between thousands of gene expression traits and hundreds of thousands of SNP genotypes are computed then each SNP-trait pair can be computed independently of the other pairs (Schadt et al., 2010). In another parallelized approach (relative quantitation by MS/MS analysis) it was shown that the computing time decreases linearly with the number of used processors (Hartler et al., 2007). Certain applications, such as microarray analysis or constructing weighted co-expression networks operate on the data most efficiently if they are held in a computer's random access memory (RAM).The analysis and management of the vast amount of high-throughput sequencing data is accompanied by a trend in using latest HPC developments such as cloud computing. Often the infrastructure at the departmental level is a mixture of available technologies: multi-core CPU for single workstations, local HPC cluster for the bulk of intensive calculations, and cloud computing for less critical data.

In general, in cases where a single high-throughput instrument is used, purchasing servers and hiring a system administrator

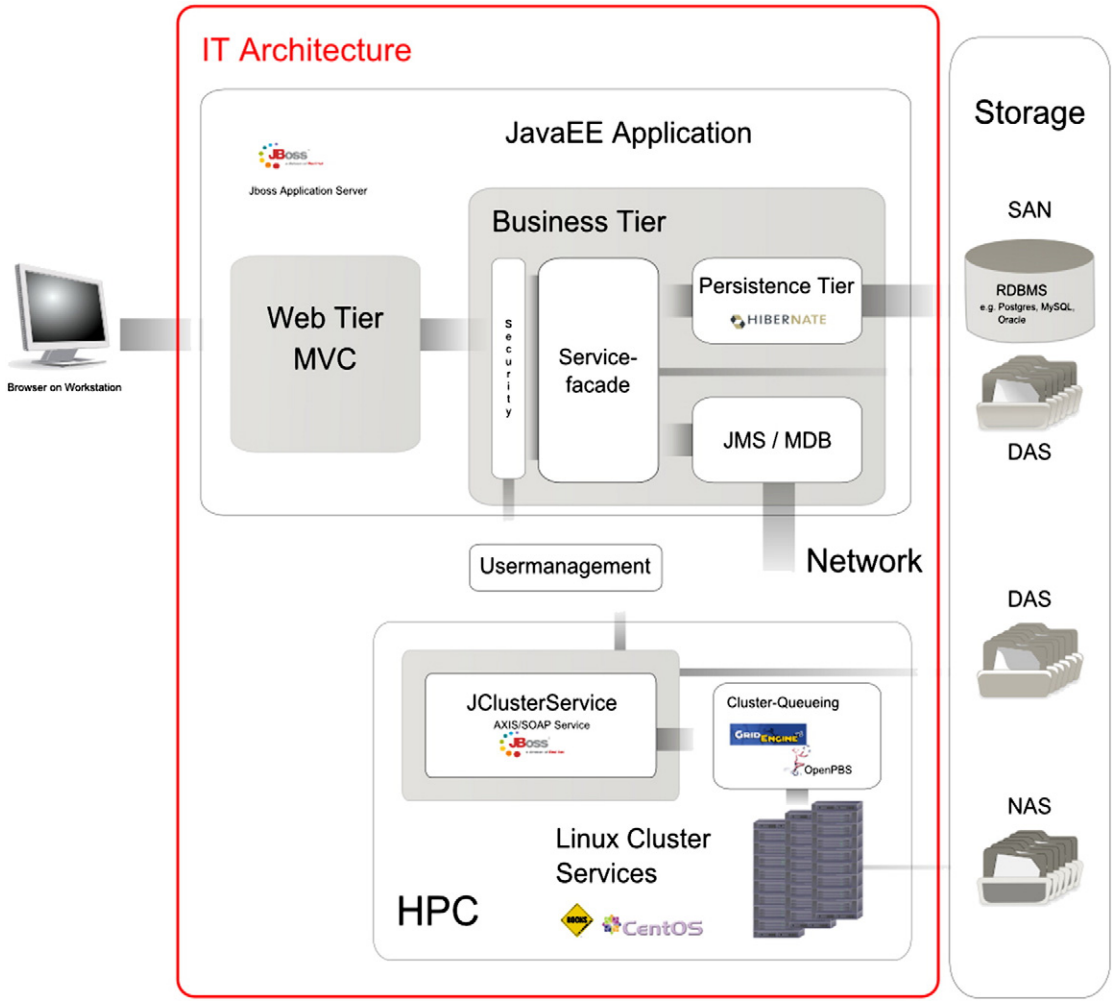


Fig. 2. IT infrastructure for translational cancer research. The IT infrastructure can be divided into three components: 1) storage component including database management system, 2) high-performance computing infrastructure (HPC), and 3) application servers providing services of applications to remote users. State-of-the-art implementation of application servers uses layered architecture separating different tiers (see text for details).

Box 1
Storage solutions

Direct attached storage (DAS): a DAS consists of one or more disks which are directly connected to the accessing computer using input/output interfaces.
Network attached storage (NAS): a NAS provides file-based access over a network protocol direct attached storage devices. Ready to use NAS boxes can be easily integrated into existing infrastructure and can be extended as well. NAS are usually located in a local area network.
Storage area network (SAN): a storage area network is a dedicated network specifically designed to offer computers block-wise access to storage devices like disks, disk arrays or backup-devices and libraries. This network is optimized for data transfer.
iSCSI: similar to the concept of SAN iSCSI offers block-wise access to remote storage devices but instead of using expensive fibre channel network, an inexpensive Ethernet network is applied.
DAS/NAS hybrids: there are special hybrid solutions which combine SAN and NAS into one single storage appliance. Hereby computers can access a central disk system using file-based protocols as well as block-based protocols.

are sufficient to process the generated raw data. Often core facilities are providing the appropriate infrastructure and a researcher can focus on the data analysis and interpretation. Multiple instruments and/or several high-throughput platforms require dedicated hardware and personnel and hence, medium to long-term commitment of the institution to maintain IT infrastructure including housing with appropriate power supply and air conditioning.

2.4. Application servers

An application server is a program, which provides services or applications to remote (web)-accessing clients. Basically it can provide dedicated services like file service, mail service or web service but can also be an execution environment for custom-made applications. In the latter case an application server offers applications a standardized environment and provides infrastructure services like connections to databases, web services, as well as coordinated and simultaneous data access. An application deployed into an application server consumes the services and implements its specific functionality like managing microarray data or running server-side processing steps. Examples for application server environments are Java EE or Microsoft .NET Server.

Box 2
High-performance computing solutions

Custom-made processor chips: the lowest level of hardware implementation is a specialized chip which processes the data through an integrated circuit and returns the data back to a control system. The control system can be a regular PC.
FPGA-boards: field programmable gate arrays (FPGA) are integrated circuits which can be programmed. For some commonly used bioinformatics algorithms there are commercially available FPGA-boards, which can be accessed by custom-made software.
Graphics processing unit (GPU): rendering of 3D graphics is a numerical intensive task for which manufacturers of graphic cards have already implemented parallel processing pipelines.
Single multi-processor/core computer: the more conventional and more frequently used approach of running HPC infrastructure is to use standard central processing units (CPU). Algorithms can be implemented with high-level programming languages and therefore development is much easier.
High-performance computing (HPC) cluster: a HPC cluster can be defined as a set of multiple computing devices, which communicate via a high-performance computing network. The communication- and file-service- networks are in most cases dedicated networks which can be reached from outside.
Grid computing: compared to local HPC cluster grid computing is a more generalized approach and describes a network of IT resources, which are spread over different locations and communicate with each other via Internet.
Cloud computing: cloud computing describes the approach of IT infrastructure which can be adapted on demand according to the specific needs of the cloud consumer. The services are billed according to the usage of the service and therefore cost effective usage without having a datacenter / HPC infrastructure. Currently cloud computing is offered by various datacenters like Amazon and Google.
Distributed computing: the most loosely coupled approach to achieve distributed data processing can be achieved by using computers attached to the internet. When the computer is not used, input data is fetched from a central dispatch server and calculation is performed on the remote computer. Results are sent back to the dispatching server.

It is a well-established software engineering to introduce a layered architecture, which separates presentation tier, middle tier (also called business tier) and persistence tier (Fig. 2). The advantage of this approach lays in the encapsulation and reusability of tiers. E.g. different presentation tiers like web applications or web services offer users customized views and functionality by accessing always the same middle tier. Examples for applications which use these technologies were recently published (Hartler et al., 2007; Stocker et al., 2009).

Other web-based application server environments are Zope or web servers with script extensions like PHP, Perl, Ruby, or Python. Most of these frameworks introduce tiered views of applications comparable to the approach described above. Recently, also cloud computing providers offered comparable application server environments known as "platform as a service" which are hosted in well-established data centers. In this case there is no need for expensive production server environments and hence, it is a preferable choice for public web-based applications which do not handle critical patient data. However, this solution is limited to the capabilities offered by the service provider to the application developer.

Alternative approach is a hybrid solution which combines local rich client applications with remote accessed web services. Hereby, locally installed applications are used for interactively composing analysis workflows which consume remote applications over the internet during execution. The used application is independent from the infrastructure behind the web services. A popular example for such a hybrid solution which integrates local analysis functionality and remote SOAP web services is Taverna (Oinn et al., 2004).

3. Computational methods

Not only the information technology aspects as described above are of importance but also the analytical methods for the integration of diverse datasets. There is a plethora of computational methods for the analyses of biomolecular and clinical data including bioinformatics and statistical tools. The number of bioinformatics tools developed in the past ten years is exponentially increasing. In Table 1 commonly used tools for gene expression and functional analysis are listed. For a summary of statistical tools we refer to textbooks (i.e., Altman, 1990). In the following chapter we are reviewing approaches for data integration: data aggregation, integrative data analysis, pathways and networks, and mathematical modelling.

3.1. Data aggregation and meta-analysis

One approach in data integration is the aggregation (meta-analysis) of the same type of data, which can increase sample size and hence improve statistical power (Mathew et al., 2007). Due to the maturity of the technology and the availability of the data, expression profiling data can be analyzed using this approach. Several platforms of microarray technology have been applied now for over a decade and gene expression profiles and datasets in many different tumor samples, heterogeneous cells within the tumor and its microenvironment, cancer types and subtypes, and other neoplastic events were performed and have been made partly available through public repositories like GEO or ArrayExpress.

Direct comparison and integration implies a number of issues to be considered: 1) normalization of raw data to exclude study-, platform-, batch-specific effects (e.g., see Orlov et al., 2007). As an example results across all studies have to be included for quantile normalization as used for Affymetrix GeneChips analysis. 2) Detection of differentially expressed genes (e.g., Motakis et al., 2009) and correction for multiple hypothesis testing, 3) a correct annotation of probesets, transcripts/genes is crucial to compare expression levels of the same entity (e.g., transcript isoform), and 4) sample nomenclature and consistency in clinical (histological) sample description across studies. There are two approaches: different microarray experiments are put together to form a single dataset (clustering or intersection operations can then be easily performed) or each individual microarray experiment is analyzed first and then the statistical results from all experiments are aggregated as for example in the rank aggregation approach (Pihur & Datta, 2008). A number of procedures exist for the combination of statistical results (e.g., p-values) and classical statistical methods as used for the meta-analysis in clinical trials (Whitehead & Whitehead, 1991) can be also applied to microarray data (Grutzmann et al., 2005). Other classical approaches include the Mantel–Haentzel method as used in the case of stratified groups (Mantel & Haenszel, 1959) or meta-regression to explore the relationship between study characteristics. There are also more advanced methods specifically developed for this purpose including the latent variable approach (Choi et al., 2003; Wang et al., 2004; Choi et al., 2007). Ultimately meta-analyses are indispensible for identification of robust prognosis signatures and (gene) biomarkers in cancer as demonstrated by several examples (Rhodes et al., 2004a; Grutzmann et al., 2005; Wirapati et al., 2008; Dreyfuss et al., 2009). But meta-analyses play

Table 1
Selected tools and resources for gene expression and functional analysis.

Function	Tools	Reference
Clustering, classification, visualization	Genesis	Sturn et al., 2002
	GeneSpring	(Agilent)
	PAM	Tibshirani et al., 2002
	TM4	Saeed et al., 2003
	OmniViz	Valk et al., 2004
	GEPAS	Herrero et al., 2003
	ExpressionProfiler	Kapushesky et al., 2004
	Cluster	Eisen et al., 1998
Gene ontology, enrichment analysis	GSEA	Subramanian et al., 2005
	DAVID,	Huang et al., 2009
	g:Profiler	Reimand et al., 2007
	AmiGO	Carbon et al., 2009
	Onto-Tools	Draghici et al., 2003
	ClueGO	Bindea et al., 2009
	Golorize	Garcia et al., 2007
	FatiGO	Al-Shahrour et al., 2004
	GoStat	Beissbarth and Speed, 2004
Pathways, mapping	KEGG	Kanehisa et al., 2004
	HumanCyc	Romero et al., 2005
	BioCarta Pathways	(BioCarta)
	Reactome	Joshi-Tope et al., 2005
	Cancer Cell Map	http://cancer.cellmap.org
	PathwayExplorer	Mlecnik et al., 2005
	GenMAPP	Dahlquist et al., 2002
	INOH	http://www.inoh.org
	PANTHER	Mi et al., 2005
	Science signaling maps	(http://stke.sciencemag.org/cm/)
Proteins, interactions	HPRD	Peri et al., 2003
	BIND	Bader et al., 2003
	DIP	Xenarios et al., 2002
	BioGRID	Stark et al., 2006
	STRING	Jensen et al., 2009
	Annotator	Schneider et al., 2010
	Pfam	Sonnhammer et al., 1997
	PROSITE	Hulo et al., 2008
	InterPro	Hunter et al., 2009
	ProDom	Servant et al., 2002
	SMART	Schultz et al., 1998
	BLOCKS	Henikoff et al., 2000
	UniProt	UniProt-Consortium, 2010
Networks	Cytoscape	Shannon et al., 2003
	ARACNe	Basso et al., 2005
	WGCNA	Langfelder and Horvath, 2008
	IPA	(Ingenuity)
	Bibliosphere	(Genomatix)
Databases, repositories	GEO	Edgar et al., 2002
	ArrayExpress	Parkinson et al., 2005
	SMD	Sherlock et al., 2001
	Oncomine	Rhodes et al., 2004b
Software environments, tool collections	R	(http://www.r-project.org)
	Bioconductor	Gentleman et al., 2004
	Matlab	(MathWorks)
	GenePattern	Reich et al., 2006

not only an important role in gene expression studies but also for genome-wide association studies in cancer (Landi et al., 2009). Further methods for meta-analyses and integration of genomics and genetics data are summarized in Guerrero et al. (2009).

An outstanding database for cancer gene expression data is Oncomine combining a compendium of >39,000 cancer genomic profiles (Rhodes et al., 2004b; Rhodes & Chinnaiyan, 2005; Rhodes et al., 2005a; Rhodes et al., 2007). It includes a microarray data pipeline, a gene annotation data warehouse, and analytical tools for differential expression, co-expression, enrichment modules and interaction networks. The integrated method Cancer Outlier Profile Analysis (COPA) (Tomlins et al., 2005) is useful to identify outlier expression profiles in cases where only a small subset of tumor samples shows overexpression. Further large-scale informatics initiatives and frameworks for cancer research are shown in Table 2.

3.2. Integrative data analysis: methodological aspects

The development of high-throughput technologies including technologies for measuring genetic variations, quantitation of gene expression, protein levels, posed challenges for the storage and analyses of the vast amount of generated data. One approach for integrating genomic data, transcriptomic data, and proteomic data is the concept of a data warehouse (often a relational database), where heterogeneous data are organized and merged to allow a consistent access for integrative data analysis, data mining (search for new patterns in the underlying data), and supervised machine learning (using patterns within the data to build classifiers for new data).

As previously shown for functional interaction of genes/proteins (Fraser & Marcotte, 2004; Rhodes et al., 2005b; Jensen et al., 2009), data sets from diverse experiments can be individually tested for their quality against a benchmark set and weighted accordingly. Various statistical approaches can then be used for the integration. In the Naïve Bayesian integration model the resulting likelihood ratio (LR) is basically the product of the LR ratio from each individual dataset, where the likelihood ratio shows the relation between the posterior odds and the prior odds given the evident interaction within each dataset (Rhodes et al., 2005b; Laubenbacher et al., 2009).

Another proposed integration methodology uses an optimization algorithm to minimize the numbers of false positives and false negatives (Hwang et al., 2005). It makes no assumptions about the number of data sets integrated and may be applied to data from any existing and future technologies. In a common situation when the raw data are not available, perhaps the simplest method to combine independent p-values is the one developed by Fisher (Fisher, 1954; Hamid et al., 2009). Also graph models (Steiner trees) were successfully applied for integration of proteomic, transcriptomic, and interactome data in signaling and regulatory networks (Huang & Fraenkel, 2009).

3.3. Integrative data analysis: examples

3.3.1. Integration of genetic data, genomic variation, and gene expression data

There are several successful examples for integrated analyses for the identification of cancer genetic events like gene fusion and gene rearrangement in prostate cancer and melanoma (Garraway et al., 2005; Tomlins et al., 2005). The heterogeneity between individual tumors, however, will make it difficult to apply multiple targeted therapies and patient stratification based on a mutational signature of defined key genes (Wood et al., 2007; Fox et al., 2009). Similar heterogeneity was observed by deep sequencing of the hematopoetic cancer genome: identified genes are mutated in only a small fraction of AML cases (Ley et al., 2008; Fox et al., 2009). In any case, the new sequencing technologies

Table 2
Large-scale cancer (informatics) initiatives and frameworks.

Initiative/framework	Abbreviation	Institution
Cancer Biomedical Informatics Grid	CaBIG	NCI, NIH
Cancer Genome Atlas	TCGA	NCI/NHGRI, NIH
Cancer Genome Anatomy Project	CGAP	NCI, NIH
Cancer Molecular Analysis Project	CMAP	NCI, NIH
Oncomine		U-M, Compendia
Biomedical Informatics Research Network	BIRN	NCRR, NIH
Repository for Molecular Brain Neoplasia Data	REMBRANDT	NCRR, NIH
Glioma Molecular Diagnostic Initiative	GMDI	NCI/NINDS, NIH
ASCENTA®		GeneLogic
BioExpress® (Oncology suite)		GeneLogic
National Cancer Research Institute (Informatics Initiative)	NCRI	NCRI

are instrumental to many aspects in cancer biology (transcriptome, genome, methylome, and genomic variations).

Genomic variations in cancer are another important data type for integrative analysis including single nucleotide polymorphisms (SNPs), insertions and deletions (Indels), copy number variations (CNVs), loss-of-heterozygosity, chromosomal aberrations and rearrangements. For example somatic copy number alterations were recently identified across multiple cancer types (Beroukhim et al., 2010). First step in this analysis is usually to map the variations to genomic positions within exons, genes, or the whole genome – comprising also regulatory sequences – or alignment to chromosomes or cytogenetic regions. An interesting approach is to combine genetic data with gene expression analysis (eQTL, eSNPs, and stepwise linkage analysis of microarray signatures — SLAMS) (Adler et al., 2006; Zhong et al., 2010).

3.3.2. Integrating clinical and gene expression data

Clinical data from patients are collected during standard treatment procedures and during clinical trials and include a number of parameters, i.e., cancer stages and scores (e.g., Gleason Score for prostate cancer), prognosis (survival time and relapse time), subtyping, or cancer biology parameters like ER-status for breast cancer (Mathew et al., 2007; Sims, 2009). Again, IT solutions and databases are inevitable for the access to these data and mining of electronic medical records. However, there are legal issues such as patient confidentiality. Lack of standardization between hospitals and institutions also makes gathering clinical data difficult (Baudot et al., 2009).

Integration of clinical and gene expression data can be divided into 2 approaches: unsupervised and supervised analysis (Quackenbush, 2006). Unsupervised clustering is used disregarding prior knowledge to subgroup tumors/patients by similarity in their expression profiles and can be used to identify (new) molecular subtypes (with cytogenetic or molecular abnormalities) and uncover biologically interesting patterns. In most cases hierarchical clustering or k-means clustering is used, where the latter needs a priori knowledge of the expected numbers of clusters. Factor analyses, which reduce the dimensionality of the expression data such as principal component analysis (PCA) (Raychaudhuri et al., 2000) or correspondence analysis (CA) (Fellenberg et al., 2001) are instrumental to identify most informative gene expression patterns and associated genes, indicating potential biomarkers. A plethora of clustering algorithms and distance measurements (e.g., Pearson correlation coefficient or Euclidean distance) has been developed so far and many of them are integrated in dedicated software applications (see Table 1). Either genes or samples/patients can be grouped based on their expression profiles across samples or across genes, respectively. In case of biclustering genes and samples are both clustered simultaneously.

Classification is a supervised approach, which takes external factors (clinical data) into account. The first need is a feature selection process to identify which genes best distinguishes two (or more) classes of patients/tumors in the data set. For this purpose a wide variety of statistical methods can be used including t-test, analysis of variance (ANOVA), and significance analysis of microarrays (SAM) (Tusher et al., 2001) and the selected genes and their patterns of expression can be used as biomarkers for diagnostic and prognostic applications. The question is whether a set of genes and their expression patterns in an initial set of patients can be used to classify disease in new patients (Valk et al., 2004; Quackenbush, 2006). Classification algorithms like support vector machines (SVM) or nearest shrunken centroids (Tibshirani et al., 2002) can be applied on a training set resulting in classifiers which then can be used on test data for class prediction.

In several studies these approaches turned out to be very successful in predicting cancer subclasses (Alizadeh et al., 2000; Perou et al., 2000; van 't Veer et al., 2002; Valk et al., 2004). In a leukemia study the authors demonstrated that a classification with 100% specificity and 100% sensitivity (Haferlach et al., 2005) can be achieved. Classification of different cancer types was not only shown for gene expression profiles (Ramaswamy et al., 2001) but also for microRNAs profiles (Lu et al., 2005).

Depending on which clinical data are used, feature selection (identification of biomarkers) can also be based on correlation of gene expression values with continuous clinical parameter of the corresponding patient or in case of binary data logistic regression. Prognosis (survival probability) can be estimated by survival analysis (Kaplan–Meier curves) for overall survival or event-free survival, taking censored data into account. Integration of survival times and gene expression can be done based on comparison of survival curves between gene sets or clusters from preceding clustering analyses (Valk et al., 2004). Significant difference in survival curves can be assessed by a log-rank test or a proportional hazard model and the magnitude of the difference by estimating the hazard rate. To test individual genes (or biomolecules) for the effect on prognosis, patients are dichotomized into two groups, one with high expression and the other group with low expression (based on median or minimal p cutoff).

3.4. Biomolecular pathways and networks

Biomolecular networks have an instrumental role in the integration of medical information for the translation of high-throughput genomics into a greater understanding of the disease and into personalized medicine (Baudot et al., 2009). In cancer a number of network modelling approaches showed to be very promising (Pujana et al., 2007; Tomlins et al., 2007; Wong et al., 2008; Kreeger & Lauffenburger, 2010; Mlecnik et al., 2010b).

The first step in network modelling is usually to construct gene co-expression networks, where two genes (nodes) are connected if the global correlation between their expression profiles (strength of relationship) over the tumors/patients is above a certain cutoff (e.g., Pearson correlation coefficient >0.6) or if there is significant correlation (e.g., $p<0.05$). For weighted co-expression networks these connections can be weighted with a (sigmoid) adjacency function. Hierarchical clustering of a topological overlap measure (taking also connections between neighbouring genes into account) was effective in detection of gene co-expression modules in glioblastoma (Horvath et al., 2006). Another association measure is mutual information used by two common approaches, namely Algorithm for the Reconstruction of Accurate Cellular Networks (ARACNe) (Basso et al., 2005) and Relevance Networks (Butte & Kohane, 2000).

It is impossible based on gene expression to identify causal relationships. However, there are some approaches such as data process inequality (Basso et al., 2005), partial correlation (de la Fuente et al., 2004), and conditional independence (Friedman et al., 2000) resulting in more reliable predictions. Common modifiers might be detected by an algorithm called MINDy, which identifies genome-wide post-translational modifiers based on gene expression profiles of a transcription factor, target, and the modulator (Wang et al., 2009). Global correlation analysis may miss patterns that only cover a subset of samples in each class, caused by heterogeneity of the disease cause and differential co-expression analyses might take this into account (Fang et al., 2010).

A network can be also built based on a measure such as the kappa-score if there is an agreement between two gene sets (e.g., target genes for transcription factors or microRNAs) or if they are sharing gene ontology terms as applied in ClueGO (Bindea et al., 2009) and DAVID (Huang et al., 2009). For cancer metastasis networks the co-occurrence measures Phi-correlation and relative risk were used (Chen et al., 2009). In addition to gene expression a number of different resources can be integrated into networks providing further

insights otherwise hidden in the complex data sets. Especially protein–protein interaction (association) data provide a meaningful complementary source for this purpose as evident from following examples.

Co-expression profiling and a network modelling strategy based on interactome networks among other data starting from 4 known breast cancer associated genes and their product resulted in identifying HMRR as a factor responsible for centrosome dysfunction (Pujana et al., 2007). Ideker and colleagues introduced an algorithm and demonstrated extracting relevant subnetworks from protein–protein interaction networks based on coherent expression patterns of their genes (also in case where genes were not significantly differentially expressed). This approach was successfully applied in the network based classification of breast cancer metastasis (Chuang et al., 2007).

A B-cell interactome (BCI) was determined by Naïve Bayes integration of protein–protein, protein–DNA, and modulatory interaction clues. Network dysregulation analyses was performed in that way that BCI edges with aberrant behaviour in phenotype show a difference in the mutual information between gene pairs Applying this method oncogenes and molecular perturbation targets in B-cell lymphomas could be identified (Mani et al., 2008; Laubenbacher et al., 2009). The STRING database (Jensen et al., 2009) was used in a recent study (Mlecnik et al., 2010b) to predict protein/gene associations (see Section 4). Probabilistic models to identify functional and regulatory modules from gene expression data showed not only to be relevant in yeast (Segal et al., 2003) but also applicable to cancer (Segal et al., 2004; Wong et al., 2008).

The limitations of the pathway analyses arise from the incomplete or incorrectly defined data. As a result, the network and pathway predictions are often becoming not reproducible soon after the publication. Hence, it is of utmost importance to experimentally verify at least some of the predictions. Nevertheless, pathway and network analyses are becoming increasingly popular for several reasons. First, proteins are social and do not act individually in a cellular context. Second, complex diseases are characterized by deregulated pathways and it is therefore necessary to study pathways instead of individual genes. And finally, in many cases the predicted interactions are robust and the likelihood for experimental validation is high.

3.5. Mathematical models

Another way of integrating data is mathematical modelling. Modelling has been successfully applied in physiology for many decades but only recently the quality and the quantity of biomolecular data became available for the development of causative and predictive models. Interactions between tumor cells and the surrounding cells are highly complex and mathematical models and computational simulation can help to delineate molecular processes and support the identification of novel key players. Computational tools can investigate mechanisms on different biological scales and predict tumor behaviour, which may highlight promising direction of the experimental work for cancer diagnosis and therapy.

3.5.1. Mathematical models of cancer

The number of mathematical models that describe solid tumor dynamics has increased dramatically. Mathematical models of cancer can be divided into two groups: descriptive and mechanistic (Anderson & Quaranta, 2008).

Descriptive models explain the regulation of growth such as size and cell numbers without emphasis on cell biological detail (Araujo & McElwain, 2004; Kozusko & Bourdeau, 2007; Anderson & Quaranta, 2008). There are several reviews of multi-scale mathematical models of tumor growth (Bellomo et al., 2003; Bellomo et al., 2008; Macklin et al., 2009). Zhang et al. (2009) explained applicability of a multi-scale tumor modelling platform that understands brain cancer. The spatial model for avascular tumor growth which is described by partial differential equations or cellular automata provides a pattern on the surface of multi-cell spheroids (Roose et al., 2007; Chaplain, 2008).

Mechanistic models focus on a specific aspect of tumor progression (e.g., molecular mechanisms) in order to understand biological processes that derive cancer therapy (Araujo & McElwain, 2004; Anderson & Quaranta, 2008; Joshi et al., 2009; Ribeiro & Pinto, 2009). Johnston et al. discussed two mechanisms that could regulate the growth of cell numbers and maintain the equilibrium that is normally observed in the crypt. Results show that an increase in cell renewal can lead to the growth of cancers and the long lag phases in tumor growth, during which new, higher equilibria are reached, before unlimited growth in cell numbers ensues (Johnston et al., 2007).

3.5.2. Mathematical models of cancer-immune cells interactions

Recently, a large number of studies have accumulated indicating that the immune system can recognize and eliminate tumor cells (Smyth et al., 2001; Parish, 2003; Eftimie et al., in press). There are still many unanswered questions about how the immune system interacts with the growing tumors and which components of the immune system play significant roles in responding to immunotherapy (De Pillis et al., 2005). Mathematical modelling of tumor–immune system interactions and chemotherapy treatment would provide an analytical predictive framework to address such questions. In this context, it is noteworthy that early model developments (Kuznetsov et al., 1994) already mimicked a number of phenomena that are seen in vivo. Later, numerical simulations lead to a deeper understanding of the solid cancer dormancy (Matzavinos et al., 2004).

In a recent study Kim et al. (2008) used a mathematical model together with the new experimental data to hypothesize that there may be a feasible, low-risk, clinical approach to enhancing the effects of imatinib treatment (Kim et al., 2008). Moore et al. modelled the interaction T cell subpopulations and CML cancer cells in the body, using a system of ordinary differential equations (ODEs). In doing so, parameters were determined which play a critical role in remission or clearance of the cancer (Moore & Li, 2004). Byrne et al. (2004) developed a simple mathematical model that described interactions between normal cells, tumor cells and infiltrating macrophages. De Pillis et al modelled the interaction of the NK and CD8+ T cells with various tumor cell lines using a system of differential equations (De Pillis & Radunskaya, 2003) and proposed new forms for the tumor-immune competition terms, and validate these forms through comparison with the experimental data (Diefenbach et al., 2001).

4. Case study: integrating biomolecular and clinical data for the identification of immunological marker in colorectal cancer

We have recently integrated biomolecular data and clinical data for colorectal cancer to identify new prognostic markers (Pages et al., 2005; Galon et al., 2006; Pages et al., 2009). The database incorporates >1700 patients with associated clinical data and biomolecular measurements. The detailed description of the database has been described elsewhere (Mlecnik et al., 2010a). Here we demonstrate the use of several computational methods to explore the data and formulate new hypotheses.

The infrastructure of choice was the departmental type, i.e., local database for the selected patient cohort. The database TME.db (Tumor MicroEnvironment database) (Mlecnik, et al., 2010a) includes only processed data and clinical data, whereas the raw data are stored elsewhere. The database provides R-based statistical tools implementing parametric and nonparametric tests and methods for survival analysis.

The high-dimensionality and complexity of the biomolecular data leads to a real interpretation challenge. In our approach (Mlecnik et al., 2010b), we integrated experimental data with prior knowledge

from publicly available databases and took advantage of publicly available tools (Sturn et al., 2002; Shannon et al., 2003; Jensen et al., 2009). After identifying the genes whose expression was significantly associated with patient disease-free survival, we reconstructed a gene–gene network. The reconstructed molecular interaction network was then visualized. These analyses enabled the investigation of deregulated pathways as well as detailed maps of interactions between genes/proteins/metabolites within a single pathway.

In the context of IT solutions for integrating clinical and biomolecular data this case study shows few points that need to be considered. First, as science is becoming driven by data as a source of hypotheses, data management should be made an integral part of the research activities. Retrospective data management not only takes considerable efforts but it is often hindered or even impossible. Second, managing data requires long-term commitment since the infrastructural and personnel resources are not negligible. In-house solutions are preferable due to changing requirements. And third, iterative cycles of computational and experimental work can not only improve the tools and leverage the data, but also provide answers to scientific questions.

5. Conclusion

In this paper we presented IT solutions and computational tools required for the integration of biomolecular and clinical data for the identification of cancer markers and targets for therapy. Although used to address cancer, the approach is generic and can be applied also to other multifactorial diseases such as diabetes or cardiovascular diseases.

It is evident that high-throughput genomic technologies rely on high-performance computing infrastructure. An increasing use of high-end computational infrastructure in a clinical setting will be seen in the future. Integration of patient archiving systems for imaging data (PACS), genomic and pharmacogenomic databases, as well as other laboratory and patient-relevant data will require novel solutions. Integration of patient databases represents significant challenges for designers and administrators of information management systems. The lack of international standards in patient care and management and different accounting systems will require the development and installation of country-specific (or even regional-specific) systems. Security issues arising from the sensitivity of certain types of information need to be addressed and solved in a proper manner. This development is inevitable and requires institutional commitment. However, the necessary resources should not be underestimated.

While institutional solutions are under way, it will take time until researchers are able to fully exploit the potential of integrating biomolecular and clinical data. Meanwhile, a more pragmatic approach is to establish a medium-scale solution at the departmental level as shown here. Such a focused infrastructure requires only a fraction of the costs and time as compared to an institutional one. It should be noted that due to the available technology the computational infrastructure can be dislocated from the actual site where the data are generated. Web-based interfaces to databases and software applications and appropriate security measures are common in many scientific as well as in other areas. As demonstrated in a series of studies with our collaborators, this lean IT solution for integrating biomolecular and clinical data can indeed identify new cancer biomarkers for improving diagnosis and clinical outcome (Pages et al., 2005; Galon et al., 2006; Pages et al., 2009; Mlecnik et al., 2009).

Although the design of the IT solution is specific for cancer immunology, the development and installation of similar settings is straightforward for several reasons. First, there are a number of academic packages for the storage and preprocessing of raw data for specific molecules, which can be easily installed. Second, bioinformatics analysis and visualization tools, which can handle various types of processed data are available. And third, software technology as well as computational infrastructure at the departmental level is affordable and does not represent a limitation for the analysis of medium-scale data sets — normally the case after preprocessing and filtering. We strongly believe that similar specific solutions will provide insights into molecular mechanisms of cancer and support the identification of novel cancer biomarkers and targets for therapy.

Acknowledgments

This work was supported by the Austrian Ministry for Science and Research, GEN-AU Project Bioinformatics Integration Network (BIN), INSERM, the National Cancer Institute (INCa), Association pour la Recherche sur le Cancer (ARC), the Cancéropole Ile de France, Ville de Paris, and by the European Commission (FP7, Geninca Consortium, grant number 202230).

References

Adler, A. S., Lin, M., Horlings, H., Nuyten, D. S., van de Vijver, M. J., & Chang, H. Y. (2006). Genetic regulators of large-scale transcriptional signatures in cancer. *Nat Genet 38*, 421—430.

Alizadeh, A. A., Eisen, M. B., Davis, R. E., Ma, C., Lossos, I. S., Rosenwald, A., et al. (2000). Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature 403*, 503—511.

Al-Shahrour, F., Diaz-Uriarte, R., & Dopazo, J. (2004). FatiGO: a web tool for finding significant associations of gene ontology terms with groups of genes. *Bioinformatics 20*, 578—580.

Altman (1990). *Practical Statistics for Medical Research.* : Chapman & Hall.

Anderson, A. R., & Quaranta, V. (2008). Integrative mathematical oncology. *Nat Rev Cancer 8*, 227—234.

Araujo, R. P., & McElwain, D. L. (2004). A history of the study of solid tumour growth: the contribution of mathematical modelling. *Bull Math Biol 66*, 1039—1091.

Bader, G. D., Betel, D., & Hogue, C. W. (2003). BIND: the Biomolecular Interaction Network Database. *Nucleic Acids Res 31*, 248—250.

Basso, K., Margolin, A. A., Stolovitzky, G., Klein, U., Dalla-Favera, R., & Califano, A. (2005). Reverse engineering of regulatory networks in human B cells. *Nat Genet 37*, 382—390.

Baudot, A., Gomez-Lopez, G., & Valencia, A. (2009). Translational disease interpretation with molecular networks. *Genome Biol 10*, 221.

Beissbarth, T., & Speed, T. P. (2004). GOstat: find statistically overrepresented gene ontologies within a group of genes. *Bioinformatics 20*, 1464—1465.

Bellomo, N., De Angelis, E., & Preziosi, L. (2003). Multiscale modeling and mathematical problems related to tumor evolution and medical therapy. *J Theor Med 5*, 111—136.

Bellomo, N., Li, N. K., & Maini, P. K. (2008). On the foundations of cancer modelling: selected topics, speculations, and perspectives. *Math Model Meth Appl Sci 18*, 593—646.

Beroukhim, R., Mermel, C. H., Porter, D., Wei, G., Raychaudhuri, S., Donovan, J., et al. (2010). The landscape of somatic copy-number alteration across human cancers. *Nature 463*, 899—905.

Bindea, G., Mlecnik, B., Hackl, H., Charoentong, P., Tosolini, M., Kirilovsky, A., et al. (2009). ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. *Bioinformatics 25*, 1091—1093.

Butte, A. J., & Kohane, I. S. (2000). Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. *Pac Symp Biocomput 5*, 415—426.

Byrne, H. M., Cox, S. M., & Kelly, C. E. (2004). Macrophage–tumour interactions: in vivo dynamics. *Discrete Cont Dyn B 4*, 81—98.

Carbon, S., Ireland, A., Mungall, C. J., Shu, S., Marshall, B., & Lewis, S. (2009). AmiGO: online access to ontology and annotation data. *Bioinformatics 25*, 288—289.

Chaplain, M. A. (2008). Modelling aspects of cancer growth: insight from mathematical and numerical analysis and computational simulation. In J. Banasiak, V. Capasso, M. A. J. Chaplain, M. Lachowicz, & J. Miekisz (Eds.), *Multiscale Problems in the Life Sciences*. Berlin: Springer.

Chaudhuri, S., & Dayal, U. (1997). *An Overview of Data Warehousing and OLAP Technology*. ACM SIGMOD Record.

Chaussabel, D., Ueno, H., Banchereau, J., & Quinn, C. (2009). Data management: it starts at the bench. *Nat Immunol 10*, 1225—1227.

Chen, L. L., Blumm, N., Christakis, N. A., Barabasi, A. L., & Deisboeck, T. S. (2009). Cancer metastasis networks and the prediction of progression patterns. *Br J Cancer 101*, 749—758.

Choi, H., Shen, R., Chinnaiyan, A. M., & Ghosh, D. (2007). A latent variable approach for meta-analysis of gene expression data from multiple microarray experiments. *BMC Bioinform 8*, 364.

Choi, J. K., Yu, U., Kim, S., & Yoo, O. J. (2003). Combining multiple microarray studies and modeling interstudy variation. *Bioinformatics 19*(Suppl 1), i84—i90.

Chuang, H. Y., Lee, E., Liu, Y. T., Lee, D., & Ideker, T. (2007). Network-based classification of breast cancer metastasis. *Mol Syst Biol 3*, 140.

Dahlquist, K. D., Salomonis, N., Vranizan, K., Lawlor, S. C., & Conklin, B. R. (2002). GenMAPP, a new tool for viewing and analyzing microarray data on biological pathways. *Nat Genet 31*, 19—20.

de la Fuente, A., Bing, N., Hoeschele, I., & Mendes, P. (2004). Discovery of meaningful associations in genomic data using partial correlation coefficients. *Bioinformatics 20*, 3565—3574.
De Pillis, L. G., & Radunskaya, A. (2003). A mathematical model of immune response to tumor invasion. *Comput Fluid Solid Mechanics 2*, 1661—1668.
De Pillis, L. G., Radunskaya, A. E., & Wiseman, C. L. (2005). A validated mathematical model of cell-mediated immune response to tumor growth. *Cancer Res 65*, 7950—7958.
Devlin, B. (1997). *Data Warehouse — From Architecture to Implementation.* : Addison Wesley.
Diefenbach, A., Jensen, E. R., Jamieson, A. M., & Raulet, D. H. (2001). Rae1 and H60 ligands of the NKG2D receptor stimulate tumour immunity. *Nature 413*, 165—171.
Draghici, S., Khatri, P., Bhavsar, P., Shah, A., Krawetz, S. A., & Tainsky, M. A. (2003). Onto-Tools, the toolkit of the modern biologist: Onto-Express, Onto-Compare, Onto-Design and Onto-Translate. *Nucleic Acids Res 31*, 3775—3781.
Dreyfuss, J. M., Johnson, M. D., & Park, P. J. (2009). Meta-analysis of glioblastoma multiforme versus anaplastic astrocytoma identifies robust gene markers. *Mol Cancer 8*, 71.
Edgar, R., Domrachev, M., & Lash, A. E. (2002). Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res 30*, 207—210.
Eftimie, R., Bramson, J. L., & Earn, D. J. (in press). Interactions between the immune system and cancer: a brief review of non-spatial mathematical models. *Bull Math Biol.* doi:10.1007/s11538-010-9526-3
Eisen, M. B., Spellman, P. T., Brown, P. O., & Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci USA 95*, 14863—14868.
Fang, G., Kuang, R., Pandey, G., Steinbach, M., Myers, C. L., & Kumar, V. (2010). Subspace differential coexpression analysis: problem definition and a general approach. *Pac Symp Biocomput*, 145—156.
Fellenberg, K., Hauser, N. C., Brors, B., Neutzner, A., Hoheisel, J. D., & Vingron, M. (2001). Correspondence analysis applied to microarray data. *Proc Natl Acad Sci USA 98*, 10781—10786.
Fisher, R. A. (1954). *Statistical Methods for Research Workers*, 12 ed. New York: Hafner.
Fox, E. J., Salk, J. J., & Loeb, L. A. (2009). Cancer genome sequencing—an interim analysis. *Cancer Res 69*, 4948—4950.
Fraser, A. G., & Marcotte, E. M. (2004). A probabilistic view of gene function. *Nat Genet 36*, 559—564.
Friedman, N., Linial, M., Nachman, I., & Pe'er, D. (2000). Using Bayesian networks to analyze expression data. *J Comput Biol 7*, 601—620.
Galon, J., Costes, A., Sanchez-Cabo, F., Kirilovsky, A., Mlecnik, B., Lagorce-Pages, C., et al. (2006). Type, density, and location of immune cells within human colorectal tumors predict clinical outcome. *Science 313*, 1960—1964.
Garcia, O., Saveanu, C., Cline, M., Fromont-Racine, M., Jacquier, A., Schwikowski, B., et al. (2007). GOlorize: a Cytoscape plug-in for network visualization with gene ontology-based layout and coloring. *Bioinformatics 23*, 394—396.
Garraway, L. A., Widlund, H. R., Rubin, M. A., Getz, G., Berger, A. J., Ramaswamy, S., et al. (2005). Integrative genomic analyses identify MITF as a lineage survival oncogene amplified in malignant melanoma. *Nature 436*, 117—122.
Gentleman, R. C., Carey, V. J., Bates, D. M., Bolstad, B., Dettling, M., Dudoit, S., et al. (2004). Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol 5*, R80.
Grutzmann, R., Boriss, H., Ammerpohl, O., Luttges, J., Kalthoff, H., Schackert, H. K., et al. (2005). Meta-analysis of microarray data on pancreatic cancer defines a set of commonly dysregulated genes. *Oncogene 24*, 5079—5088.
Guerrero, R., Allison, D. B., & Goldstein, D. (2009). *Meta-analysis and Combining Information in Genetics and Genomics.* : CRC Press.
Haferlach, T., Kohlmann, A., Schnittger, S., Dugas, M., Hiddemann, W., Kern, W., et al. (2005). Global approach to the diagnosis of leukemia using gene expression profiling. *Blood 106*, 1189—1198.
Hamid, J. S., Hu, P., Roslin, N. M., Ling, V., Greenwood, C. M. T., & Beyene, J. (2009). Data integration in genetics and genomics: methods and challenges. *Hum Genomics Proteomics 2009*, 1—13.
Hanahan, D., & Weinberg, R. A. (2000). The hallmarks of cancer. *Cell 100*, 57—70.
Hartler, J., Thallinger, G. G., Stocker, G., Sturn, A., Burkard, T. R., Korner, E., et al. (2007). MASPECTRAS: a platform for management and analysis of proteomics LC–MS/MS data. *BMC Bioinform 8*, 197.
Henikoff, J. G., Greene, E. A., Pietrokovski, S., & Henikoff, S. (2000). Increased coverage of protein families with the blocks database servers. *Nucleic Acids Res 28*, 228—230.
Herrero, J., Al-Shahrour, F., Diaz-Uriarte, R., Mateos, A., Vaquerizas, J. M., Santoyo, J., et al. (2003). GEPAS: a web-based resource for microarray gene expression data analysis. *Nucleic Acids Res 31*, 3461—3467.
Horvath, S., Zhang, B., Carlson, M., Lu, K. V., Zhu, S., Felciano, R. M., et al. (2006). Analysis of oncogenic signaling networks in glioblastoma identifies ASPM as a molecular target. *Proc Natl Acad Sci USA 103*, 17402—17407.
Hu, H., Brzeski, H., Hutchins, J., Ramaraj, M., Qu, L., Xiong, R., et al. (2004). Biomedical informatics: development of a comprehensive data warehouse for clinical and genomic breast cancer research. *Pharmacogenomics 5*, 933—941.
Huang, S. S., & Fraenkel, E. (2009). Integrating proteomic, transcriptional, and interactome data reveals hidden components of signaling and regulatory networks. *Sci Signal 2*, ra40.
Huang, D. W., Sherman, B. T., & Lempicki, R. A. (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc 4*, 44—57.
Hulo, N., Bairoch, A., Bulliard, V., Cerutti, L., Cuche, B. A., de Castro, E., et al. (2008). The 20 years of PROSITE. *Nucleic Acids Res 36*, D245—D249.
Hunter, S., Apweiler, R., Attwood, T. K., Bairoch, A., Bateman, A., Binns, D., et al. (2009). InterPro: the integrative protein signature database. *Nucleic Acids Res 37*, D211—D215.
Hwang, D., Rust, A. G., Ramsey, S., Smith, J. J., Leslie, D. M., Weston, A. D., et al. (2005). A data integration methodology for systems biology. *Proc Natl Acad Sci USA 102*, 17296—17301.
Inmon, W. (2002). *Building the Data Warehouse*, 3 ed. : Wiley.
Jensen, L. J., Kuhn, M., Stark, M., Chaffron, S., Creevey, C., Muller, J., et al. (2009). STRING 8—a global view on proteins and their functional interactions in 630 organisms. *Nucleic Acids Res 37*, D412—D416.
Johnston, M. D., Edwards, C. M., Bodmer, W. F., Maini, P. K., & Chapman, S. J. (2007). Mathematical modeling of cell population dynamics in the colonic crypt and in colorectal cancer. *Proc Natl Acad Sci USA 104*, 4008—4013.
Joshi, B., Wang, X., Banerjee, S., Tian, H., Matzavinos, A., & Chaplain, M. A. (2009). On immunotherapies and cancer vaccination protocols: a mathematical modelling approach. *J Theor Biol 259*, 820—827.
Joshi-Tope, G., Gillespie, M., Vastrik, I., D'Eustachio, P., Schmidt, E., de Bono, B., et al. (2005). Reactome: a knowledgebase of biological pathways. *Nucleic Acids Res 33*, D428—D432.
Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y., & Hattori, M. (2004). The KEGG resource for deciphering the genome. *Nucleic Acids Res 32*, D277—D280.
Kapushesky, M., Kemmeren, P., Culhane, A. C., Durinck, S., Ihmels, J., Korner, C., et al. (2004). Expression Profiler: next generation—an online platform for analysis of microarray data. *Nucleic Acids Res 32*, W465—W470.
Kasprzyk, A., Keefe, D., Smedley, D., London, D., Spooner, W., Melsopp, C., et al. (2004). EnsMart: a generic system for fast and flexible access to biological data. *Genome Res 14*, 160—169.
Kim, P. S., Lee, P. P., & Levy, D. (2008). Dynamics and potential impact of the immune response to chronic myelogenous leukemia. *PLoS Comput Biol 4*, e1000095.
Kozusko, F., & Bourdeau, M. (2007). A unified model of sigmoid tumour growth based on cell proliferation and quiescence. *Cell Prolif 40*, 824—834.
Kreeger, P. K., & Lauffenburger, D. A. (2010). Cancer systems biology: a network modeling perspective. *Carcinogenesis 31*, 2—8.
Kuznetsov, V. A., Makalkin, I. A., Taylor, M. A., & Perelson, A. S. (1994). Nonlinear dynamics of immunogenic tumors: parameter estimation and global bifurcation analysis. *Bull Math Biol 56*, 295—321.
Landi, M. T., Chatterjee, N., Yu, K., Goldin, L. R., Goldstein, A. M., Rotunno, M., et al. (2009). A genome-wide association study of lung cancer identifies a region of chromosome 5p15 associated with risk for adenocarcinoma. *Am J Hum Genet 85*, 679—691.
Langfelder, P., & Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. *BMC Bioinform 9*, 559.
Laubenbacher, R., Hower, V., Jarrah, A., Torti, S. V., Shulaev, V., Mendes, P., et al. (2009). A systems biology view of cancer. *Biochim Biophys Acta 1796*, 129—139.
Lee, T. J., Pouliot, Y., Wagner, V., Gupta, P., Stringer-Calvert, D. W., Tenenbaum, J. D., et al. (2006). BioWarehouse: a bioinformatics database warehouse toolkit. *BMC Bioinform 7*, 170.
Ley, T. J., Mardis, E. R., Ding, L., Fulton, B., McLellan, M. D., Chen, K., et al. (2008). DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. *Nature 456*, 66—72.
Locker, G. Y., Hamilton, S., Harris, J., Jessup, J. M., Kemeny, N., Macdonald, J. S., et al. (2006). ASCO 2006 update of recommendations for the use of tumor markers in gastrointestinal cancer. *J Clin Oncol 24*, 5313—5327.
Lu, J., Getz, G., Miska, E. A., Alvarez-Saavedra, E., Lamb, J., Peck, D., et al. (2005). MicroRNA expression profiles classify human cancers. *Nature 435*, 834—838.
Macklin, P., McDougall, S., Anderson, A. R., Chaplain, M. A., Cristini, V., & Lowengrub, J. (2009). Multiscale modelling and nonlinear simulation of vascular tumour growth. *J Math Biol 58*, 765—798.
Mani, K. M., Lefebvre, C., Wang, K., Lim, W. K., Basso, K., Dalla-Favera, R., et al. (2008). A systems biology approach to prediction of oncogenes and molecular perturbation targets in B-cell lymphomas. *Mol Syst Biol 4*, 169.
Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *J Natl Cancer Inst 22*, 719—748.
Martens, L., Hermjakob, H., Jones, P., Adamski, M., Taylor, C., States, D., et al. (2005). PRIDE: the proteomics identifications database. *Proteomics 5*, 3537—3545.
Mathew, J. P., Taylor, B. S., Bader, G. D., Pyarajan, S., Antoniotti, M., Chinnaiyan, A. M., et al. (2007). From bytes to bedside: data integration and computational biology for translational cancer research. *PLoS Comput Biol 3*, e12.
Matzavinos, A., Chaplain, M. A., & Kuznetsov, V. A. (2004). Mathematical modelling of the spatio-temporal response of cytotoxic T-lymphocytes to a solid tumour. *Math Med Biol 21*, 1—34.
Maurer, M., Molidor, R., Sturn, A., Hartler, J., Hackl, H., Stocker, G., et al. (2005). MARS: microarray analysis, retrieval, and storage system. *BMC Bioinform 6*, 101.
Mi, H., Lazareva-Ulitsky, B., Loo, R., Kejariwal, A., Vandergriff, J., Rabkin, S., et al. (2005). The PANTHER database of protein families, subfamilies, functions and pathways. *Nucleic Acids Res 33*, D284—D288.
Mlecnik, B., Sanchez-Cabo, F., Charoentong, P., Bindea, G., Pages, F., Berger, A., et al. (2010a). Data integration and exploration for the identification of molecular mechanisms in tumor-immune cells interaction. *BMC Genomics 11*(Suppl 1), S7.
Mlecnik, B., Scheideler, M., Hackl, H., Hartler, J., Sanchez-Cabo, F., & Trajanoski, Z. (2005). PathwayExplorer: web service for visualizing high-throughput expression data on biological pathways. *Nucleic Acids Res 33*, W633—W637.
Mlecnik, B., Tosolini, M., Charoentong, P., Kirilovsky, A., Bindea, G., Berger, A., et al. (2009). Biomolecular network reconstruction identifies t-cell homing factors associated with survival in colorectal cancer. *Gastroenterology 138*, 1429—1440.
Mlecnik, B., Tosolini, M., Charoentong, P., Kirilovsky, A., Bindea, G., Berger, A., et al. (2010b). Biomolecular network reconstruction identifies T-cell homing factors associated with survival in colorectal cancer. *Gastroenterology 138*, 1429—1440.
Moore, H., & Li, N. K. (2004). A mathematical model for chronic myelogenous leukemia (CML) and T cell interaction. *J Theor Biol 227*, 513—523.

Motakis, E., Ivshina, A. V., & Kuznetsov, V. A. (2009). Data-driven approach to predict survival of cancer patients: estimation of microarray genes' prediction significance by Cox proportional hazard regression model. *IEEE Eng Med Biol Mag 28*, 58—66.

Oinn, T., Addis, M., Ferris, J., Marvin, D., Senger, M., Greenwood, M., et al. (2004). Taverna: a tool for the composition and enactment of bioinformatics workflows. *Bioinformatics 20*, 3045—3054.

Orlov, Y. L., Zhou, J., Lipovich, L., Shahab, A., & Kuznetsov, V. A. (2007). Quality assessment of the Affymetrix U133A&B probesets by target sequence mapping and expression data analysis. *In Silico Biol 7*, 241—260.

Pages, F., Berger, A., Camus, M., Sanchez-Cabo, F., Costes, A., Molidor, R., et al. (2005). Effector memory T cells, early metastasis, and survival in colorectal cancer. *N Engl J Med 353*, 2654—2666.

Pages, F., Kirilovsky, A., Mlecnik, B., Asslaber, M., Tosolini, M., Bindea, G., et al. (2009). In situ cytotoxic and memory T cells predict outcome in patients with early-stage colorectal cancer. *J Clin Oncol 27*, 5944—5951.

Parish, C. R. (2003). Cancer immunotherapy: the past, the present and the future. *Immunol Cell Biol 81*, 106—113.

Parkinson, H., Kapushesky, M., Shojatalab, M., Abeygunawardena, N., Coulson, R., Farne, A., et al. (2007). ArrayExpress—a public database of microarray experiments and gene expression profiles. *Nucleic Acids Res 35*, D747—D750.

Parkinson, H., Sarkans, U., Shojatalab, M., Abeygunawardena, N., Contrino, S., Coulson, R., et al. (2005). ArrayExpress—a public repository for microarray gene expression data at the EBI. *Nucleic Acids Res 33*, D553—D555.

Peri, S., Navarro, J. D., Amanchy, R., Kristiansen, T. Z., Jonnalagadda, C. K., Surendranath, V., et al. (2003). Development of human protein reference database as an initial platform for approaching systems biology in humans. *Genome Res 13*, 2363—2371.

Perou, C. M., Sorlie, T., Eisen, M. B., van de Rijn, M., Jeffrey, S. S., Rees, C. A., et al. (2000). Molecular portraits of human breast tumours. *Nature 406*, 747—752.

Pihur, V., & Datta, S. (2008). Finding common genes in multiple cancer types through meta-analysis of microarray experiments: a rank aggregation approach. *Genomics 92*, 400—403.

Pujana, M. A., Han, J. D., Starita, L. M., Stevens, K. N., Tewari, M., Ahn, J. S., et al. (2007). Network modeling links breast cancer susceptibility and centrosome dysfunction. *Nat Genet 39*, 1338—1349.

Quackenbush, J. (2006). Microarray analysis and tumor classification. *N Engl J Med 354*, 2463—2472.

Ramaswamy, S., Tamayo, P., Rifkin, R., Mukherjee, S., Yeang, C. H., Angelo, M., et al. (2001). Multiclass cancer diagnosis using tumor gene expression signatures. *Proc Natl Acad Sci USA 98*, 15149—15154.

Raychaudhuri, S., Stuart, J. M., & Altman, R. B. (2000). Principal components analysis to summarize microarray experiments: application to sporulation time series. *Pac Symp Biocomput*, 455—466.

Reich, M., Liefeld, T., Gould, J., Lerner, J., Tamayo, P., & Mesirov, J. P. (2006). GenePattern 2.0. *Nat Genet 38*, 500—501.

Reimand, J., Kull, M., Peterson, H., Hansen, J., & Vilo, J. (2007). g:Profiler—a web-based toolset for functional profiling of gene lists from large-scale experiments. *Nucleic Acids Res 35*, W193—W200.

Rhodes, D. R., & Chinnaiyan, A. M. (2005). Integrative analysis of the cancer transcriptome.*Nat Genet 37*, S31—S37 Suppl.

Rhodes, D. R., Kalyana-Sundaram, S., Mahavisno, V., Barrette, T. R., Ghosh, D., & Chinnaiyan, A. M. (2005a). Mining for regulatory programs in the cancer transcriptome. *Nat Genet 37*, 579—583.

Rhodes, D. R., Kalyana-Sundaram, S., Mahavisno, V., Varambally, R., Yu, J., Briggs, B. B., et al. (2007). Oncomine 3.0: genes, pathways, and networks in a collection of 18, 000 cancer gene expression profiles. *Neoplasia 9*, 166—180.

Rhodes, D. R., Tomlins, S. A., Varambally, S., Mahavisno, V., Barrette, T., Kalyana-Sundaram, S., et al. (2005b). Probabilistic model of the human protein–protein interaction network. *Nat Biotechnol 23*, 951—959.

Rhodes, D. R., Yu, J., Shanker, K., Deshpande, N., Varambally, R., Ghosh, D., et al. (2004a). Large-scale meta-analysis of cancer microarray data identifies common transcriptional profiles of neoplastic transformation and progression. *Proc Natl Acad Sci USA 101*, 9309—9314.

Rhodes, D. R., Yu, J., Shanker, K., Deshpande, N., Varambally, R., Ghosh, D., et al. (2004b). ONCOMINE: a cancer microarray database and integrated data-mining platform. *Neoplasia 6*, 1—6.

Ribeiro, D., & Pinto, J. M. (2009). An integrated network-based mechanistic model for tumor growth dynamics under drug administration. *Comput Biol Med 39*, 368—384.

Ritter, O., Kocab, P., Senger, M., Wolf, D., & Suhai, S. (1994). Prototype implementation of the integrated genomic database. *Comput Biomed Res 27*, 97—115.

Romero, P., Wagg, J., Green, M. L., Kaiser, D., Krummenacker, M., & Karp, P. D. (2005). Computational prediction of human metabolic pathways from the complete human genome. *Genome Biol 6*, R2.

Roose, T., Chapman, S. J., & Maini, P. K. (2007). Mathematical models of avascular tumor growth. *SIAM Rev 49*, 179—208.

Saeed, A. I., Sharov, V., White, J., Li, J., Liang, W., Bhagabati, N., et al. (2003). TM4: a free, open-source system for microarray data management and analysis. *Biotechniques 34*, 374—378.

Schadt, E. E., Linderman, M. D., Sorenson, J., Lee, L., & Nolan, G. P. (2010). Computational solutions to large-scale data management and analysis. *Nat Rev Genet 11*, 647—657.

Schneider, G., Wildpaner, M., Sirota, F. L., Maurer-Stroh, S., Eisenhaber, B., & Eisenhaber, F. (2010). Integrated tools for biomolecular sequence-based function prediction as exemplified by the ANNOTATOR software environment. *Meth Mol Biol 609*, 257—267.

Schonbach, C., Kowalski-Saunders, P., & Brusic, V. (2000). Data warehousing in molecular biology. *Brief Bioinform 1*, 190—198.

Schultz, J., Milpetz, F., Bork, P., & Ponting, C. P. (1998). SMART, a simple modular architecture research tool: identification of signaling domains. *Proc Natl Acad Sci USA 95*, 5857—5864.

Searls, D. B. (2005). Data integration: challenges for drug discovery. *Nat Rev Drug Discov 4*, 45—58.

Segal, E., Friedman, N., Koller, D., & Regev, A. (2004). A module map showing conditional activity of expression modules in cancer. *Nat Genet 36*, 1090—1098.

Segal, E., Shapira, M., Regev, A., Pe'er, D., Botstein, D., Koller, D., et al. (2003). Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nat Genet 34*, 166—176.

Servant, F., Bru, C., Carrere, S., Courcelle, E., Gouzy, J., Peyruc, D., et al. (2002). ProDom: automated clustering of homologous domains. *Brief Bioinform 3*, 246—251.

Shah, S. P., Huang, Y., Xu, T., Yuen, M. M., Ling, J., & Ouellette, B. F. (2005). Atlas — a data warehouse for integrative bioinformatics. *BMC Bioinform 6*, 34.

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., et al. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res 13*, 2498—2504.

Sherlock, G., Hernandez-Boussard, T., Kasarskis, A., Binkley, G., Matese, J. C., Dwight, S. S., et al. (2001). The Stanford Microarray Database. *Nucleic Acids Res 29*, 152—155.

Sims, A. H. (2009). Bioinformatics and breast cancer: what can high-throughput genomic approaches actually tell us? *J Clin Pathol 62*, 879—885.

Smyth, M. J., Godfrey, D. I., & Trapani, J. A. (2001). A fresh look at tumor immunosurveillance and immunotherapy. *Nat Immunol 2*, 293—299.

Sonnhammer, E. L., Eddy, S. R., & Durbin, R. (1997). Pfam: a comprehensive database of protein domain families based on seed alignments. *Proteins 28*, 405—420.

Stark, C., Breitkreutz, B. J., Reguly, T., Boucher, L., Breitkreutz, A., & Tyers, M. (2006). BioGRID: a general repository for interaction datasets. *Nucleic Acids Res 34*, D535—D539.

Stein, L. D. (2003). Integrating biological databases. *Nat Rev Genet 4*, 337—345.

Stein, L. D. (2008). Towards a cyber infrastructure for the biological sciences: progress, visions and challenges. *Nat Rev Genet 9*, 678—688.

Stephan, C., Kohl, M., Turewicz, M., Podwojski, K., Meyer, H. E., & Eisenacher, M. (2010). Using Laboratory Information Management Systems as central part of a proteomics data workflow. *Proteomics 10*, 1230—1249.

Stocker, G., Fischer, M., Rieder, D., Bindea, G., Kainz, S., Oberstolz, M., et al. (2009). iLAP: a workflow-driven software for experimental protocol development, data acquisition and analysis. *BMC Bioinform 10*, 390.

Sturn, A., Quackenbush, J., & Trajanoski, Z. (2002). Genesis: cluster analysis of microarray data. *Bioinformatics 18*, 207—208.

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., et al. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci USA 102*, 15545—15550.

Tibshirani, R., Hastie, T., Narasimhan, B., & Chu, G. (2002). Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proc Natl Acad Sci USA 99*, 6567—6572.

Tomlins, S. A., Mehra, R., Rhodes, D. R., Cao, X., Wang, L., Dhanasekaran, S. M., et al. (2007). Integrative molecular concept modeling of prostate cancer progression. *Nat Genet 39*, 41—51.

Tomlins, S. A., Rhodes, D. R., Perner, S., Dhanasekaran, S. M., Mehra, R., Sun, X. W., et al. (2005). Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. *Science 310*, 644—648.

Tusher, V. G., Tibshirani, R., & Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci USA 98*, 5116—5121.

UniProt-Consortium (2010). The Universal Protein Resource (UniProt) in 2010. *Nucleic Acids Res 38*, D142—D148.

Valk, P. J., Verhaak, R. G., Beijen, M. A., Erpelinck, C. A., van Waalwijk, Barjesteh, van Doorn-Khosrovani, S., et al. (2004). Prognostically useful gene-expression profiles in acute myeloid leukemia. *N Engl J Med 350*, 1617—1628.

van 't Veer, L. J., Dai, H., van de Vijver, M. J., He, Y. D., Hart, A. A., Mao, M., et al. (2002). Gene expression profiling predicts clinical outcome of breast cancer. *Nature 415*, 530—536.

Waller, C. L., Shah, A., & Nolte, M. (2007). Strategies to support drug discovery through integration of systems and data. *Drug Discov Today 12*, 634—639.

Wang, J., Coombes, K. R., Highsmith, W. E., Keating, M. J., & Abruzzo, L. V. (2004). Differences in gene expression between B-cell chronic lymphocytic leukemia and normal B cells: a meta-analysis of three microarray studies. *Bioinformatics 20*, 3166—3178.

Wang, K., Saito, M., Bisikirska, B. C., Alvarez, M. J., Lim, W. K., Rajbhandari, P., et al. (2009). Genome-wide identification of post-translational modulators of transcription factor activity in human B cells. *Nat Biotechnol 27*, 829—839.

Whitehead, A., & Whitehead, J. (1991). A general parametric approach to the meta-analysis of randomized clinical trials. *Stat Med 10*, 1665—1677.

Wirapati, P., Sotiriou, C., Kunkel, S., Farmer, P., Pradervand, S., Haibe-Kains, B., et al. (2008). Meta-analysis of gene expression profiles in breast cancer: toward a unified understanding of breast cancer subtyping and prognosis signatures. *Breast Cancer Res 10*, R65.

Wong, D. J., Nuyten, D. S., Regev, A., Lin, M., Adler, A. S., Segal, E., et al. (2008). Revealing targeted therapy for human cancer by gene module maps. *Cancer Res 68*, 369—378.

Wood, L. D., Parsons, D. W., Jones, S., Lin, J., Sjoblom, T., Leary, R. J., et al. (2007). The genomic landscapes of human breast and colorectal cancers. *Science 318*, 1108—1113.

Xenarios, I., Salwinski, L., Duan, X. J., Higney, P., Kim, S. M., & Eisenberg, D. (2002). DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. *Nucleic Acids Res 30*, 303—305.

Zhang, L., Wang, Z., Sagotsky, J. A., & Deisboeck, T. S. (2009). Multiscale agent-based cancer modeling. *J Math Biol 58*, 545—559.

Zhong, H., Yang, X., Kaplan, L. M., Molony, C., & Schadt, E. E. (2010). Integrating pathway analysis and genetics of gene expression for genome-wide association studies. *Am J Hum Genet 86*, 581—591.

BMC Genomics

BioMed Central

Research

Open Access

Data integration and exploration for the identification of molecular mechanisms in tumor-immune cells interaction

Bernhard Mlecnik[1,2], Fatima Sanchez-Cabo[1,3], Pornpimol Charoentong[1], Gabriela Bindea[1,2], Franck Pagès[2,4], Anne Berger[4], Jerome Galon*[2,4] and Zlatko Trajanoski*[1]

Addresses: [1]Institute for Genomics and Bioinformatics, Graz University of Technology, Petersgasse 14, 8010 Graz, Austria, [2]INSERM, U872, Integrative Cancer Immunology, Paris, France, [3]Genomics Unit, Spanish National Centre for Cardiovascular Research, Madrid, Spain and [4]AP-HP, Georges Pompidou European Hospital, Paris, France

E-mail: Bernhard Mlecnik - bernhard.mlecnik@crc.jussieu.fr; Fatima Sanchez-Cabo - fsanchezcabo@gmail.com; Pornpimol Charoentong - p.charoentong@student.tugraz.at; Gabriela Bindea - gabriela.bindea@crc.jussieu.fr; Franck Pagès - Franck.PAGES@hop.egp.ap-hop-paris.fr; Anne Berger - anne.berger@hop.egp.ap-hop-paris.fr; Jerome Galon* - jerome.galon@crc.jussieu.fr; Zlatko Trajanoski* - zlatko.trajanoski@tugraz.at
*Corresponding author

from International Workshop on Computational Systems Biology Approaches to Analysis of Genome Complexity and Regulatory Gene Networks Singapore 20-25 November 2008

Published: 10 February 2010

BMC Genomics 2010, **11**(Suppl 1):S7 doi: 10.1186/1471-2164-11-S1-S7

This article is available from: http://www.biomedcentral.com/1471-2164/11/S1/S7

Publication of this supplement was made possible with help from the Bioinformatics Agency for Science, Technology and Research of Singapore and the Institute for Mathematical Sciences at the National University of Singapore.

Abstract

Cancer progression is a complex process involving host-tumor interactions by multiple molecular and cellular factors of the tumor microenvironment. Tumor cells that challenge immune activity may be vulnerable to immune destruction. To address this question we have directed major efforts towards data integration and developed and installed a database for cancer immunology with more than 1700 patients and associated clinical data and biomolecular data. Mining of the database revealed novel insights into the molecular mechanisms of tumor-immune cell interaction. In this paper we present the computational tools used to analyze integrated clinical and biomolecular data. Specifically, we describe a database for heterogenous data types, the interfacing bioinformatics and statistical tools including clustering methods, survival analysis, as well as visualization methods. Additionally, we discuss generic issues relevant to the integration of clinical and biomolecular data, as well as recent developments in integrative data analyses including biomolecular network reconstruction and mathematical modeling.

Background

Despite extensive characterization of environmental and intrinsic and underlying mechanisms [1,2], markers of the oncogenic process remain so far poorly predictive of patient survival and fail to prove their reliability in clinical use. For example, colorectal cancer is one of the most common malignancies for both men and women [3]. The rate of localized cancers (stage I-II; UICC-TNM classification) is about 40% [4,5]. Despite surgery with curative intent, the risk of recurrence of these early-stage patients is high (approximately 20-30%). To subject all of these patients to post-operative chemotherapy may be inappropriate and costly [6]. Genetic and molecular tumor prognostic factors have been proposed to identify patients who may be at risk for recurrence. None has yet been sufficiently informative for inclusion in clinical practice [5]. Identification of patients with high-risk of recurrence is therefore a major clinical issue. However, in order to develop stratified or personalized strategies for such complex multifactorial disease it is of importance to understand how numerous and diverse elements function together in human pathology. A comprehensive understanding of cancer requires the integration and analysis of data not only from the tumor but also its microenvironment including the immune cells.

Tumors are composed of a complex network of tumor cells, immune cells, stromal components including fibroblasts, and a complex vasculature. To grow, invade, and metastasize, a tumor interacts with its microenvironment, composed of diverse cells of various origins. The microenvironment contains cells of the immune system, including inflammatory infiltrates of innate immunity and infiltrates of the adaptive immune response. In colorectal cancer, previous studies have suggested a clinical role of the immune infiltrates [7-11]. In order to investigate the role of the immune infiltrates and analyze the tumor immunological microenvironment in humans we developed and installed a database for cancer immunology with more than 1700 patients and associated clinical data and biomolecular data. By analyzing the data we showed the importance of early-metastatic invasion in colorectal cancer and could pinpoint a novel prognostic marker for survival [10]. We evidenced that the recently characterized immune cell subpopulation of effector-memory T cells (T_{EM}), may have a central role in the control of tumor spreading to lymphovascular and perineural structures but also to lymph node or distant organs. In subsequent study we demonstrated the role of the adaptive immune system for predicting clinical outcome [9]. Furthermore, we revealed the importance for patient prognosis of the nature, the functional orientation, the density and the localization of immune cell populations within the primary tumor. Thus, adaptive immune reaction and intratumoral T-cell subpopulations were better predictor of survival than traditional staging based on a cancer's size and spread [9].

In the light of these studies it was of utmost importance to integrate the data and develop tools for analysis and visualization. In this paper, we present the solutions developed to analyze the tumor immunological microenvironment in humans including database, analytical tools, and tools for visualization. Specifically, we describe here the database for clinical and biomolecular data, the interfacing bioinformatics and statistical tools including clustering methods, survival analysis, as well as visualization methods. Furthermore, we discuss upcoming developments for integrative data analyses including biomolecular network reconstruction and mathematical modeling.

Bioinformatics and statistics tools for cancer immunology

Database for cancer immunology

The database developed for cancer immunology (Tumor Microenvironment (TME)) integrates clinical and biomolecular data. The underlying relational database model is designed as a cancer patient oriented database which takes all the patients anamnesis and clinical and medical history information into account whereby all patients are linked to a speci?c hospital. Security issues were treated in regard to the interest of patients. Ethical, Legal and Social Implications (ELSI) have been fulfilled (agreement #903434), security modules implemented, and anonymous information stored. The patient information additionally includes medical problems, surgery and detailed cancer information. Additionally TME.db allows the storage of a variety of different high-throughput experiments including:

- Real-Time TaqMan qPCR gene expression data (Low density arrays, single probes, T-cell repertoire analysis)
- Microsatellite instability (MSI) and mutations data
- Flow cytometric (FACS) phenotyping data
- Protein quantification (ELISA, Quantibody, cytometric beads assays) data
- Functional data (proliferation, survival, apoptosis, migration assays)
- Immunohistochemical data (Tissue Micro Array (TMA) and whole slide analysis)

TME.db joins and integrates all different types of data and stores them in a common place where all the determined analysis parameters are linked in a clear way dependent on the sample material and the experiment type. For accessing all the stored information again sophisticated query methods were developed in order to retrieve the data in a pre-modi?ed way, already prepared for statistical analysis. As of May 2009, the database incorporates 1784 patients with associated clinical data with 60 parameters (e.g. tumor staging, treatment, cancer relapse) and 16400 different material information as well as biomolecular measurements (including qPCR for 400 genes from 125 patients, 820 FACS parameters from 40 patients, 20 tissue microarray assays for 600 patients).

Software architecture

TME is a multi-tier client-server application and can be subdivided into different functional modules which interact as self-contained units according to their defined responsibilities: presentation tier, business tier and runtime environment. The presentation tier within TME is formed by a Web interface, which allows programming access to parts of the application logic. Thus, on the client side, a user requires an Internet connection and a recent Web browser with Java support, available for almost every platform. The business tier is realized as view-independent application logic, which stores and retrieves datasets by communicating with the persistence layer. The internal management of files is also handled from a central service component, which persists the meta-information for acquired files to the database. All services of this layer are implemented as STRUTS and are using SITEMESH.

Model driven development

In order to reduce coding and to increase the long term maintainability, the model driven development environment AndroMDA is used to generate components of the persistence layer and recurrent parts from the above mentioned business layer. AndroMDA accomplishes this by translating an annotated UML-model into a JEE-platform-specific implementation using Enterprise Java Beans (EJB), STRUTS and SITEMESH. Due to the flexibility of AndroMDA, application external services, such as the user management system, have a clean integration in the model. Dependencies of internal service components on such externally defined services are cleanly managed by its build system. By changing the build parameters in the AndroMDA configuration, it is also possible to support different relational database management systems. This is because platform specific code with the same functionality is generated for data retrieval. Furthermore, technology lock-in regarding the implementation of the service layers was also addressed by using AndroMDA, as the implementation of the service facade can be switched during the build process from Spring based components to distributed Enterprise Java Beans. At present, TME is operating on one local machine and, providing the usage scenarios do not demand it, this architectural configuration will remain. However, chosen technologies are known to work on Web server farms and crucial distribution of the application among server nodes is transparently performed by the chosen technologies.

Data retrieval, collaboration and data sharing

TME offers search masks which allow keyword based searching in the recorded projects, experiments and notes. These results are often discussed with collaboration partners to gain different opinions on the same raw data. In order to allow direct collaboration between scientists TME is embedded into a central user management system which offers multiple levels of access control to projects and their associated experimental data. The sharing of projects can be done on a per-user basis or on an institutional basis. For small or local single-user installations, the fully featured user management system can be replaced by a file-based user management which still offers the same functionalities from the sharing point of view, but lacks institute-wide functionalities.

Bioinformatics analysis tools

The database was mined using standard bioinformatics tools. Specifically, qPCR and FACS data were explored using two-dimensional hierarchical clustering of correlation matrices (i.e. gene-wise correlation of the respective patient groups [9]). Genesis clustering software was used to visualize the correlation matrix and to perform Pearson un-centered hierarchical clustering [12]. This tool was developed for large-scale gene expression cluster analysis and integrates various tools for microarray data analysis such as filters, normalization and visualization tools, distance measures as well as common clustering algorithms including hierarchical clustering, self-organizing maps, k-means, principal component analysis, and support vector machines [12].

Statistical analysis

Survival analysis provides a statistical framework for the modeling and statistical analysis of the time to event for a cohort of patients [13]. Since the distribution of survival times might have an unusual and often unknown form, nonparametric Kaplan-Meier estimates are widely used when censoring is present for the characterization of groups of patients with different underlying characteristics, i.e. calculating median survival times and patients at risk after a given period.

Similarly, the log-rank non-parametric test is used to check the null hypothesis that at any time point there is no difference in the probability of the event of interest between the groups [14]. The magnitude of the difference and its confidence interval can be calculated using a Cox proportional hazards model. Furthermore the effect of a novel biomarker can be adjusted for traditional parameters if this modeling strategy is used on several covariates.

TME implements the previous tests within a statistical analysis module. Calculations are done using the survival package from R [15] to which TME connects using RServe [16]. The aim is the automatic detection of biomarkers or sets of biomarkers that - alone or in combination with other parameters - are able to discriminate groups of colorectal cancer patients with good prognosis from those with bad prognosis for both, overall and disease-free survival. In particular, TME provides:

- Kaplan-Meier curves, estimates of the median survival time and number of patients at risk after a certain time period for the different groups of patients

- Log-rank test for the analysis of the differences in survival between groups of patients with different underlying characteristics

- Univariate Cox proportional hazards model to estimate the magnitude of the effect of the covariate in survival

- Tools for the categorization of numeric covariates into a fixed number of levels. This can be useful for the classification of the patients into groups based on the biomolecular markers stored in TME for each patient, such as the expression level of a gene or the number of cells of a given type found at different locations of the tumor sample.

Although categorization of the patients into groups might result in loss of information [17], this is often done in clinical practice. The way the cut-off is set for dichotomizing a continuous variable is also controversial: A previously described value or a biologically justified level can be used as suggested by Altman *et al* [18]. In the absence of a biologically sound cut-off value, using a statistic of the sample (such as the median) balances the number of cases per group but results in different levels across studies making the comparison of results from different groups difficult [17]. Hence, the analysis must be repeated in an independent cohort of patients categorized using the cut-off previously selected. The same is true when using the "minimum p-value" approach [19], i.e. taking the point yielding the "maximum" significance between groups. This approach has additional important problems such as the overestimation of the prognostic importance of the covariate and multiple testing issues that might be accounted for [18]

TME allows the inspection of the covariates dichotomizing them based in any of the previous options. In particular, if the minimum p-value approach is used the log-rank p-value can be corrected using either the formula proposed by Altman *et al* [18] or with cross-validation as proposed by Faraggi & Simon [20]. Additionally, TME implements the shrinkage method proposed by Holländer *et al* [21] to correct the hazard ratios.

Next version of TME will also include multivariate analysis using a Cox proportional hazards model and decision trees, which can easily accommodate heterogeneous variables and have yielded already satisfactory results in the discovery of biomarkers for breast cancer [22].

Data visualization

Data visualization was carried out using the publicly available software tools Cytoscape, ClueGO, and GOlorize. Cytoscape is free software package for visualizing, modeling and analyzing molecular and genetic interaction networks [23-26]. In Cytoscape, the nodes represent genes or proteins and they are connected with edges which representing interactions. Typical biological networks at the molecular level are gene regulation networks, signal transduction networks, protein interaction networks, and metabolic networks. In order to capture biological information, ClueGO [25], a Cytoscape plug-in, uses Gene Ontology [27] categories that are over-represented in selected one or two lists of genes. ClueGO takes advantage of GOlorize [26] plug-in, an efficient tool to the same class node-coloring and the class-directed layout algorithm for advanced network visualization.

Discussion

In this paper we described computational tools developed specifically to address biological questions in cancer immunology. The computational tools include: 1) a database for clinical and biomolecular data comprising >1700 patients with associated clinical information, FACS data, qPCR data, tissue microarray data; 2) bioinformatics tools developed for the analyses of medium and large-scale data, 3) statistical tools for the survival analysis; and 4) tools for visualization of the data. The power of the dedicated informatics solution is leveraged by the integration of all computational

resources using various interfaces. During the course of the development of the database, the implementation of the analytical tools, and the analysis of the data we have learned several important lessons.

Lessons learned

First, development of a dedicated database is time-consuming but indispensable task. In recent years, the biology community has expended considerable effort to confront the challenges of managing heterogeneous data in a structured and organized way and as a result developed information management systems for both raw and processed data. Laboratory information management systems (LIMS) have been implemented for handling data entry from robotic systems and tracking samples as well as data management systems for processed data including microarrays, proteomics data, and microscopy data. In general, these sophisticated systems are able to manage and analyze data generated for only a single type or a limited number of instruments, and were designed for only a specific type of molecule. Thus, addressing a biological question relying on several complementary technologies requires a specific off-the-shelf database. It should be noted that such a database could absorb several person-years of software engineering and this effort tends to be underestimated.

Second, incorporation of clinical data poses additional challenges. Many institutions have electronic patient records and in principle, extracting the information could be straightforward. However, technical, ethical, and legal issues might delay or even prohibit the process of data collection. Heterogeneous clinical and departmental information systems, accessibility of patient data, and managing sensitive information can introduce several levels of complexity and require extensive stakeholder discussions. A complex information management system that captures in a secure way the relevant data is suggestive only for large (i.e. several hundred PIs) institutions. The majority of the labs are better off with a design of a relatively small, departmental database for only few specific cohorts. The patient data should be first de-identified and then provided to the biologists and bioinformaticians.

Third, primary data should be archived at a separate location and only preprocessed and normalized data should be stored in the dedicated database. Although it is tempting to upload and analyze all types of data in a single system, experience shows that primary data is mostly used once. This approach is even more advisable for large-scale data including microarrays, proteomics of sequence data. However, links to the primary data need to be secured so that later re-analyses using improved tools can be guaranteed. In this context it is noteworthy that in the analyses we have performed so far only medium-throughput data was used, meaning that the number of analyzed molecular species was in the range of 100-1000. With this number of elements the majority of the tools perform satisfactorily on a standard desktop computer. Performance is a crucial issue if the number of molecules detected in a single patient sample increases to >10.000 (like in microarray studies) or >100.000 (proteomics studies) and the used methods need to be re-evaluated.

In this paper we show a powerful approach for integrative analyses of heterogenous biomolecular data and clinical data. Although powerful, our approach was sequential, i.e. the data was integrated in the database and the query masks allowed sequential analyses of specific biomolecular data, and their correlation with clinical data. We strongly believe that integrative data analyses methods will provide additional insights otherwise hidden in the complex data sets. Several approaches were suggested previously (e.g. [23-26,28-30]). However, normalization of the data, availability of reference datasets, and scarcity of the data (specific measurements are not available for all patients) are non-trivial issues which are difficult to address. In this context, novel data integration approaches are highly desirable. In the following paragraphs we highlight two approaches, namely biomolecular network reconstruction and mathematical modelling, which have the potential to provide mechanistic insights and ultimately translation of this knowledge to clinical applications.

Biomolecular network reconstruction

One emerging field, which was not addressed in this paper is biomolecular network reconstruction. The data we have so far used are actual measurements and are limited to the available technology and/or samples. There is a wealth of information stored in public databases on protein-protein interactions, text mining, two-hybrid screens, or gene silencing using siRNA. The integration of this datasets in databases like STRING [31] and the visualization tools like Cytoscape [23] and associated-software such as ClueGO [25] opens new avenues of exploration of biomolecular networks.

Mathematical modeling

Since the pathophysiological mechanisms underlying cancer are highly complex and involve many different cell types and processes, mathematical modeling is becoming an important tool to integrate the biological information and enhance our understanding of

interaction between cancer and immune system. Moreover, mathematical modeling may direct direction of experimental work for treatment and diagnosis. Here we briefly describe relevant modeling efforts for tumour-immune cells interaction.

Mathematical models of cancer

Traditionally, mathematical models of cancer fall into two broad camps: descriptive and mechanistic [32]. Descriptive models tend to focus on reproducing the gross characteristic of tumors such as size and cell numbers, are generally used to investigate tumor cell population dynamics, without emphasis on cell biological detail [32-34]. Over the last decades, many mathematical models have been proposed that focus on tumor growth. Macklin *et al*. [35] performed a new multiscale mathematical model for solid tumor growth which couples an improved model of tumor invasion with a model of tumor-induced angiogenesis. A large number of studies have described deterministic models which have been used to model the spatio-temporal spread of tumors [36]. By contrast, mechanistic models focus on specific aspects of tumor progression in order to explain the underlying biological processes that drive them [32,33,37].

Mathematical models of immune response

The regulation of immune system involves the interaction between populations of pathogen and immune cell. Immunological memory and specificity are property of the immune system. This ability to respond more rapidly and effective than to the first exposure [38]. Understanding of these aspects requires quantitative models of proliferation and differentiation of T lymphocytes. Mathematical modeling can describe these behaviors as deterministic or stochastic models. De Boer *et al*. proposed the simple mathematical model in which parameters can be estimated (proliferation and death rate) during clonal expansion and contraction phase [39,40]. Three models have been proposed by Ganusov [41] to discriminate between alternative memory cell differentiation pathways.

Mathematical models of cancer-immune interactions

Mathematical modeling of tumor growth that includes the immune response and chemotherapy treatment would provide an analytical predictive framework. Kim *et al*. developed a mathematical model with the new experimental data to gain insights into the dynamics and potential impact of the resulting anti-leukemia immune response on chronic myelogenous leukemia (CML) [42]. Moore *et al*. modeled the interaction T cell subpopulations and CML cancer cells in the body, using a system of ordinary differential equations [43]. Steffen *et al*. presented a mathematical model of melanoma invasion into healthy tissue with an immune response. They used this model as a framework with which to investigate primary tumor invasion and treatment by surgical excision [44].

Conclusion

In this paper we presented computational tools developed to manage and explore clinical and biomolecular data for the identification of molecular mechanisms in the tumor microenvironment. The presented bioinformatics and statistics solutions were applied on a patient cohort with colorectal cancer and revealed novel insights in the tumor-immune cells interaction. Although used to address a specific question, the approach is generic and can be applied also to different cancers as well as to other multifactorial diseases like diabetes or cardiovascular diseases.

List of abbreviations used

JavaEE: Java Enterprise Edition platform; MDA: Model Driven Architecture; SOAP: Simple Object Access Protocol

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

BM developed the database. BM, FSC, GB, and PC carried out the analyses. FP and AB collected and annotated the clinical data. JG and ZT coordinated the project. All authors contributed to the drafting of the manuscript, and read and approved the final manuscript.

Acknowledgements

This work was supported by the Austrian Ministry for Science and Research, GEN-AU Project Bioinformatics Integration Network (BIN), Austrian Science Fund (SFB Project Lipotoxicity), INSERM, the National Cancer Institute (INCa), Association pour la Recherche sur le Cancer (ARC), the Cancéropole Ile de France, Ville de Paris, and by the European Commission (FP7, Geninca Consortium, grant number 202230).

This article has been published as part of *BMC Genomics* Volume 11 Supplement 1, 2010: International Workshop on Computational Systems Biology Approaches to Analysis of Genome Complexity and Regulatory Gene Networks. The full contents of the supplement are available online at http://www.biomedcentral.com/1471-2164/11?issue=S1.

References

1. Steeg PS, Ouatas T, Halverson D, Palmieri D and Salerno M: **Metastasis suppressor genes: basic biology and potential clinical use.** *Clin Breast Cancer* 2003, **4:**51–62.
2. Hanahan D and Weinberg RA: **The hallmarks of cancer.** *Cell* 2000, **100:**57–70.
3. Parkin DM, Bray F, Ferlay J and Pisani P: **Global cancer statistics, 2002.** *CA Cancer J Clin* 2005, **55:**74–108.
4. Sobin LWC: *TNM classification of malignant timors* Wiley-Liss; 2000.

5. Locker GY, S H, J H, J J, N K, J M, M S, D H, RJ B and ASCO: **ASCO 2006 update of recommendations for the use of tumor markers in gastrointestinal cancer.** *J Clin Oncol* 2009, **24:**5313–5327.
6. Benson AB III, Schrag D, Somerfield MR, Cohen AM, Figueredo AT, Flynn PJ, Krzyzanowska MK, Maroun J, McAllister P and Van Cutsem E, *et al*: **American Society of Clinical Oncology recommendations on adjuvant chemotherapy for stage II colon cancer.** *J Clin Oncol* 2004, **22:**3408–3419.
7. Dalerba P, Maccalli C, Casati C, Castelli C and Parmiani G: **Immunology and immunotherapy of colorectal cancer.** *Crit Rev Oncol Hematol* 2003, **46:**33–57.
8. Atreya I and Neurath MF: **Immune cells in colorectal cancer: prognostic relevance and therapeutic strategies.** *Expert Rev Anticancer Ther* 2008, **8:**561–572.
9. Galon J, Costes A, Sanchez-Cabo F, Kirilovsky A, Mlecnik B, Lagorce-Pages C, Tosolini M, Camus M, Berger A and Wind P, *et al*: **Type, density, and location of immune cells within human colorectal tumors predict clinical outcome.** *Science* 2006, **313:**1960–1964.
10. Pages F, Berger A, Camus M, Sanchez-Cabo F, Costes A, Molidor R, Mlecnik B, Kirilovsky A, Nilsson M and Damotte D, *et al*: **Effector memory T cells, early metastasis, and survival in colorectal cancer.** *N Engl J Med* 2005, **353:**2654–2666.
11. Galon J, Fridman WH and Pages F: **The adaptive immunologic microenvironment in colorectal cancer: a novel perspective.** *Cancer Res* 2007, **67:**1883–1886.
12. Sturn A, Quackenbush J and Trajanoski Z: **Genesis: cluster analysis of microarray data.** *Bioinformatics* 2002, **18:**207–208.
13. Harrel FE: *Regression modeling strategies: with applications to Linear Models, Logistic Regression and Survival analysis* Springer Series in Statistics; 2001.
14. Bland JM and Altman DG: **The logrank test.** *BMJ* 2004, **328:**1073.
15. http://www.r-project.org.
16. http://rosuda.org/Rserve/.
17. Altman DG and Royston P: **The cost of dichotomising continuous variables.** *BMJ* 2006, **332:**1080.
18. Altman DG, Lausen B, Sauerbrei W and Schumacher M: **Dangers of using "optimal" cutpoints in the evaluation of prognostic factors.** *J Natl Cancer Inst* 1994, **86:**829–835.
19. Heinzl HTC: **A cautionary note on segmenting a cyclical covariate by minimum P-value search.** *Computational Statistics & Data Analysis* 2009, **35:**451–461.
20. Faraggi D and Simon R: **A simulation study of cross-validation for selecting an optimal cutpoint in univariate survival analysis.** *Stat Med* 1996, **15:**2203–2213.
21. Hollander N, Sauerbrei W and Schumacher M: **Confidence intervals for the effect of a prognostic factor after selection of an 'optimal' cutpoint.** *Stat Med* 2004, **23:**1701–1713.
22. Pittman J, Huang E, Nevins J, Wang Q and West M: **Bayesian analysis of binary prediction tree models for retrospectively sampled outcomes.** *Biostatistics* 2004, **5:**587–601.
23. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B and Ideker T: **Cytoscape: a software environment for integrated models of biomolecular interaction networks.** *Genome Res* 2003, **13:**2498–2504.
24. Cline MS, Smoot M, Cerami E, Kuchinsky A, Landys N, Workman C, Christmas R, Avila-Campilo I, Creech M and Gross B, *et al*: **Integration of biological networks and gene expression data using Cytoscape.** *Nat Protoc* 2007, **2:**2366–2382.
25. Bindea G, Mlecnik B, Hackl H, Charoentong P, Tosolini M, Kirilovsky A, Fridman WH, Pages F, Trajanoski Z and Galon J: **ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks.** *Bioinformatics* 2009, **25:**1091–1093.
26. Garcia O, Saveanu C, Cline M, Fromont-Racine M, Jacquier A, Schwikowski B and Aittokallio T: **GOlorize: a Cytoscape plug-in for network visualization with Gene Ontology-based layout and coloring.** *Bioinformatics* 2007, **23:**394–396.
27. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS and Eppig JT, *et al*: **Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.** *Nat Genet* 2000, **25:**25–29.
28. Hwang D, Rust AG, Ramsey S, Smith JJ, Leslie DM, Weston AD, de Atauri P, Aitchison JD, Hood L and Siegel AF, *et al*: **A data integration methodology for systems biology.** *Proc Natl Acad Sci USA* 2005, **102:**17296–17301.
29. Liang S, Fuhrman S and Somogyi R: **Reveal, a general reverse engineering algorithm for inference of genetic network architectures.** *Pac Symp Biocomput* 1998, 18–29.
30. Margolin AA, Nemenman I, Basso K, Wiggins C, Stolovitzky G, Dalla FR and Califano A: **ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context.** *BMC Bioinformatics* 2006, **7(Suppl 1):**S7.
31. von Mering C, Jensen LJ, Snel B, Hooper SD, Krupp M, Foglierini M, Jouffre N, Huynen MA and Bork P: **STRING: known and predicted protein-protein associations, integrated and transferred across organisms.** *Nucleic Acids Res* 2005, **33:** D433–D437.
32. Anderson AR and Quaranta V: **Integrative mathematical oncology.** *Nat Rev Cancer* 2008, **8:**227–234.
33. Araujo RP and McElwain DL: **A history of the study of solid tumour growth: the contribution of mathematical modelling.** *Bull Math Biol* 2004, **66:**1039–1091.
34. Kozusko F and Bourdeau M: **A unified model of sigmoid tumour growth based on cell proliferation and quiescence.** *Cell Prolif* 2007, **40:**824–834.
35. Macklin P, McDougall S, Anderson AR, Chaplain MA, Cristini V and Lowengrub J: **Multiscale modelling and nonlinear simulation of vascular tumour growth.** *J Math Biol* 2009, **58:**765–798.
36. Roose T, Chapman SJ and Maini PK: **Mathematical models of avascular tumor growth.** *Siam Review* 2007, **49:**179–208.
37. Anderson ACMRK: *Single-Cell-Based Models in Biology and Medicine (Mathematics and Biosciences in Interaction)* Birkhauser Basel; 12001.
38. Beverley PC: **Primer: making sense of T-cell memory.** *Nat Clin Pract Rheumatol* 2008, **4:**43–49.
39. De Boer RJ, Oprea M, Antia R, Murali-Krishna K, Ahmed R and Perelson AS: **Recruitment times, proliferation, and apoptosis rates during the CD8(+) T-cell response to lymphocytic choriomeningitis virus.** *J Virol* 2001, **75:**10663–10669.
40. De Boer RJ, Homann D and Perelson AS: **Different dynamics of CD4+ and CD8+ T cell responses during and after acute lymphocytic choriomeningitis virus infection.** *J Immunol* 2003, **171:**3928–3935.
41. Antia R, Ganusov VV and Ahmed R: **The role of models in understanding CD8+ T-cell memory.** *Nat Rev Immunol* 2005, **5:**101–111.
42. Kim PS, Lee PP and Levy D: **Dynamics and potential impact of the immune response to chronic myelogenous leukemia.** *PLoS Comput Biol* 2008, **4:**e1000095.
43. Moore H and Li NK: **A mathematical model for chronic myelogenous leukemia (CML) and T cell interaction.** *J Theor Biol* 2004, **227:**513–523.
44. Eikenberry S, Thalhauser C and Kuang Y: **Tumor-immune interaction, surgical treatment, and cancer recurrence in a mathematical model of melanoma.** *PLoS Comput Biol* 2009, **5:** e1000362.

Biomolecular Network Reconstruction Identifies T-Cell Homing Factors Associated With Survival in Colorectal Cancer

BERNHARD MLECNIK,[*,‡,§,‖] MARIE TOSOLINI,[*,‡,§] PORNPIMOL CHAROENTONG,[‖] AMOS KIRILOVSKY,[*,‡,§] GABRIELA BINDEA,[*,‡,§,‖] ANNE BERGER,[¶] MATTHIEU CAMUS,[*,‡,§] MÉLANIE GILLARD,[*,‡,§] PATRICK BRUNEVAL,[#] WOLF–HERMAN FRIDMAN,[*,‡,§,**] FRANCK PAGÈS,[*,‡,§,**] ZLATKO TRAJANOSKI,[‖] and JÉRÔME GALON[*,‡,§,¶]

[*]INSERM, Integrative Cancer Immunology Team, INSERM U872, Paris, France; [‡]Université Paris Descartes, Paris, France; [§]Cordeliers Research Center, Université Pierre et Marie Curie Paris 6, Paris, France; [‖]Institute for Genomics and Bioinformatics, Graz University of Technology, Graz, Austria; [¶]Department of General and Digestive Surgery, [#]Department of Pathology, and [**]Department of Immunology, Georges Pompidou European Hospital, Assistance Publique-Hôpitaux de Paris, AP-HP, Paris, France

BACKGROUND & AIMS: Colorectal cancer is a complex disease involving immune defense mechanisms within the tumor. Herein, we used data integration and biomolecular network reconstruction to generate hypotheses about the mechanisms underlying immune responses in colorectal cancer that are relevant to tumor recurrence. **METHODS:** Mechanistic hypotheses were formulated on the basis of data from 108 patients and tested using different assays (gene expression, phenome mapping, tissue-microarrays, T-cell receptor [TCR] repertoire). **RESULTS:** This integrative approach revealed that chemoattraction and adhesion play important roles in determining the density of intratumoral immune cells. The presence of specific chemokines (*CX3CL1*, *CXCL10*, *CXCL9*) and adhesion molecules (*ICAM1*, *VCAM1*, *MADCAM1*) correlated with different subsets of immune cells and with high densities of T-cell subpopulations within specific tumor regions. High expression of these molecules correlated with prolonged disease-free survival. Moreover, the expression of certain chemokines associated with particular TCR repertoire and specific TCR use predicted patient survival. **CONCLUSIONS: Data integration and biomolecular network reconstruction is a powerful approach to uncover molecular mechanisms. This study shows the utility of this approach for the investigation of malignant tumors and other diseases. In colorectal cancer, the expression of specific chemokines and adhesion molecules were found as being critical for high densities of T-cell subsets within the tumor and associated with particular TCR repertoire. Intratumoral-specific TCR use correlated with the prognosis of the patients.**

Keywords: Integrative Biology; Colorectal Cancer; Chemokines; Immune Reaction.

To develop stratified or personalized strategies for complex multifactorial diseases it is important to understand how numerous and diverse elements function together in human pathology.[1,2] A comprehensive understanding of diseases such as cancer not only will require the integration and analysis of data from the tumor in its microenvironment, but also of other data sources from model organisms stored in public databases.[1,2] Cancer is the result of an accumulation of genetic alterations that allows growth of neoplastic cells.[3,4] The adenoma–carcinoma sequence underlies the development of colorectal cancer (CRC), and distinct pathways (microsatellite instability and chromosomal instability pathways) have been identified.[5] The natural evolution of a cancer also involves antagonistic interactions of the tumor with the defense mechanisms of the host.[6,7] Inflammatory mediators can promote tumor progression and metastases.[8] The innate and adaptive immune systems also can protect the host against tumor development through mechanisms of immunosurveillance.[9] The increased susceptibility of immunodeficient mice to carcinogen-induced and spontaneous tumors showed the role of innate and adaptive immunity in the control of tumor development.[9–11] More recent data provide support for a role for adaptive immunity also during the equilibrium phase of cancer.[12]

In human CRC, adaptive immune reaction was found, and densities of immune cells are very different from patient to patient.[7] Numerous HLA-restricted T cells specific for tumor peptides have been described.[13] Lymphocytes infiltrating solid tumors have been associated with improved prognosis.[14–16] Tumors from CRC patients containing a high density of infiltrating memory and effector memory T cells were found to be less likely to disseminate to lymphovascular and perineural structures and to regional lymph nodes.[17] Tumor recurrence and overall patient survival times correlated broadly with the immune context and the presence of memory T cells within the tumor.[18,19] Tumors also contain a variety of cytokines, chemokines, and inflammatory and cytotoxic

Abbreviations used in this paper: CT, center; CRC, colorectal cancer; DFS, disease-free survival; HR, hazard ratio; IM, invasive margin; PCR, polymerase chain reaction; TCR, T-cell receptor; T_H, T-helper–specific.

0016-5085/10/$36.00
doi:10.1053/j.gastro.2009.10.057

mediators. This complex network reflects the heterogeneity underlying tumor biology and tumor–host interactions.[7,9] The reasons for the very different densities of immune cells found within tumors, however, remain unknown.

To gain an improved understanding of tumor–host interactions in human CRC, we developed and applied an intuitive data integration strategy to analyze immune reaction in CRC. We used a method that effectively created hypotheses permitting us to detect an immune network relevant to prognosis. Predicted molecules involved in lymphocyte chemoattraction and adhesion were analyzed together with immune populations in situ. Biological hypotheses then were validated in a large cohort of patients by a combination of high-throughput approaches. The novel aspects of our study revealed mechanisms resulting in high or low densities of specific immune cells at the tumor site. Chemokines and adhesion molecules associated with immune effector T cells with particular TCR repertoire. Furthermore, the presence of a specific intratumoral TCR repertoire correlated with the survival of the patient. Thus, we provided a framework for predicting effective host-immune reaction against cancer in human beings. This study shows the utility of data integration and biomolecular network reconstruction for the investigation of malignant tumors and other diseases.

Materials and Methods

Patients and Database

The records of CRC patients who underwent a primary resection of their tumor at the Laennec George Pompidou European Hospital (HEGP) Hospitals between 1996 and 2004 were reviewed and described previously.[18] Histopathologic and clinical findings were scored according to the International Union Against Cancer (UICC)-TNM staging system. For details, see the Supplementary Materials and Methods section and Supplementary Table 1. A secure web-based database Tumor Microenvironment Database (TME.db) was built in our lab for the management of patient data. Ethical, legal, and social implications were approved by the ethical review board.

Gene Expression Analysis

Frozen tumor samples (cohort 1, n = 108; revalidation cohort 2, n = 27) of randomly selected patients available from Laennec-HEGP Hospitals (1996–2004), with sufficient RNA quality and quantity, were selected for gene expression analysis. Total RNA was isolated by homogenization with the RNeasy isolation kit (Qiagen, Valencia, CA). Quantitative real-time TaqMan polymerase chain reaction (PCR) was performed using low-density arrays and the 7900 robotic real-time PCR system (Applied Biosystems, Foster City, CA). Data were analyzed using SDS Software v2.2 (Applied Biosystems) and the TME.db statistical module.

Large-Scale Flow Cytometric Analysis

After mechanical dispersion, cells from fresh tumors were washed and subjected to 4-color flow cytometry. Cells were resuspended in phosphate-buffered saline/0.5% bovine serum albumin and incubated for 30 minutes at 4°C with antibodies and relevant isotype controls. Forty thousand cells were analyzed per run. Analyses were performed with a FACScalibur flow cytometer and CellQuest software (Becton Dickinson, San Diego, CA).

T-Cell Receptor Repertoire Analysis

The complementarity determining region 3 (CDR3) length distribution analysis was achieved by performing reverse-transcription of V and V–J gene composition and transcripts into complementary DNA; CDR3-encoding messenger RNA (mRNA) was amplified by PCR using specific V and C primers. The intratumoral T-cell repertoire was performed on 10 randomly selected colorectal tumors using the TcLandscape technology (TcLand, Nantes, France).

Tissue Microarray and Immunohistochemistry

By using a tissue-array instrument (Beecher Instruments, Alphelys, Plaisir, France), 2 representative regions of the tumor (center [CT] and invasive margin [IM]) were punched from paraffin-embedded tissue blocks. Tissue-microarray sections were incubated with monoclonal antibodies against CD3 (SP7), CD8 (4B11), CD45RO (OPD4), GZMB (GrB-7), CD57 (NK1), CD1A (O10), cytokeratin (AE1AE3), and cytokeratin-8 (Neomarkers, Fremont, CA), T-bet (4B10) (Santa Cruz Biotechnology, Santa Cruz, CA), and CD68 (PG-M1) (Dako, Copenhagen, Denmark). Envision+ system and 3.3′-diaminobenzidine tetrahydrochloride–chromogen (DAB) were applied (Dako). Slides were analyzed using an image analysis workstation (Spot Browser; Alphelys).

Statistical Analysis

The gene prediction network in Figure 1 was created using the Search Tool for the Retrieval of Interacting Proteins (STRING) database and Gene Ontology (GO). Correlation matrix was performed using Pearson uncentered hierarchical clustering. For pairwise comparisons of parametric and nonparametric data the Student *t* test and the Wilcoxon rank-sum test were used, respectively. Kaplan–Meier estimators of survival were used to visualize the survival curves. Hazard ratio (Cox proportional hazards model) and the log-rank test were used to compare disease-free and overall survival between patients in different groups. To avoid overfitting, hazard ratios obtained by the minimum *P* value approach were corrected.[18] *P* values for gene combination analysis with high gene expression in CT and IM (HiHi) vs low expression in those two regions (LoLo) were corrected for multiple testing using the Benjamini-Hochberg method. We

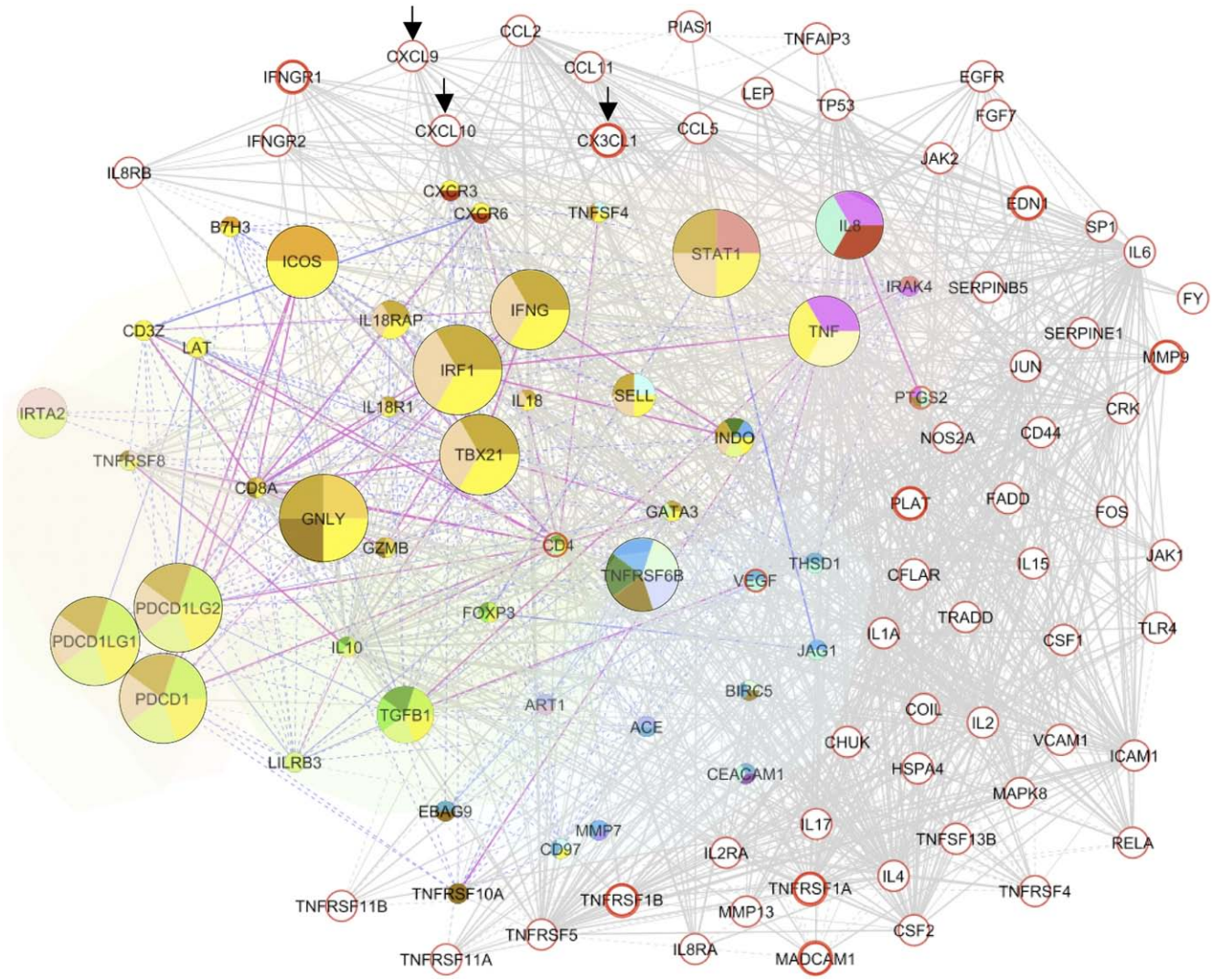


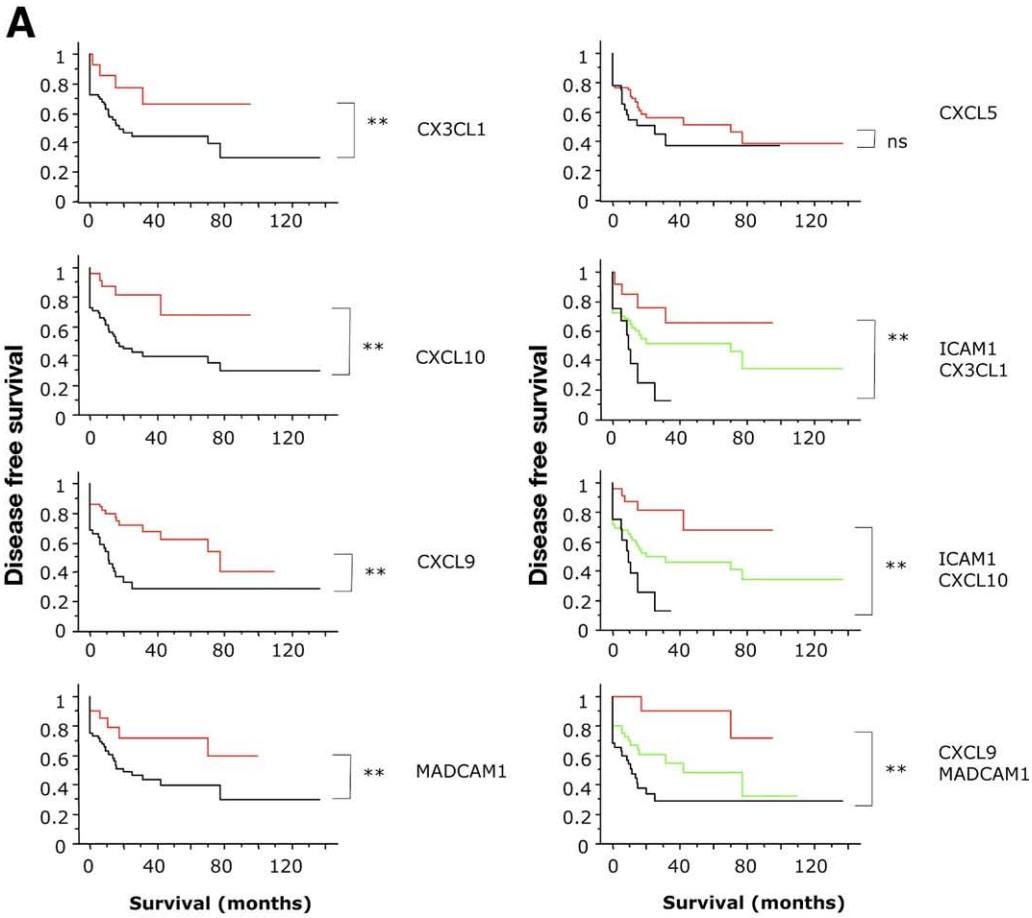
Figure 1. Biomolecular network using gene expression data in a cohort of patients with CRC and predicted gene–gene interactions based on available knowledge. The network illustrated experimental data (*colored nodes*) and in silico prediction (*white nodes* surrounded by a *red border*). The gene expression data were acquired by a reverse-transcription PCR study for 47 genes in a cohort of 108 CRC patients (Supplementary Table 1). The network was reconstructed based on a subset of 12 genes, which reached a significant log-rank level for DFS. The network shows the top genes predicted in silico plus the genes analyzed by reverse-transcription PCR. *CX3CL1* was the top predicted gene. All nodes surrounded by a *red border* were predicted by STRING. The node sizes of the network are based on the HR for DFS (Supplementary Table 1). Nodes surrounded by a *black border* had significant log-rank *P* values ($P < .05$). The edge weights of the network are based on the integrated score of the pairwise uncentered Pearson correlation value between the 47 reverse-transcription PCR genes and the combined edge scores for all genes predicted in silico provided by STRING (see Supplementary Materials and Methods section for details). The network node layout was based on Gene Ontology (GO), gene expression correlations (*blue lines*), STRING scores (*gray lines*), and the integrated association strength between genes (*edge thickness*). Edge thickness levels show the relation strength based on the integrated score value between the nodes. Nodes are colored based on multiple occurrences in different GO categories (Supplementary Table 2).

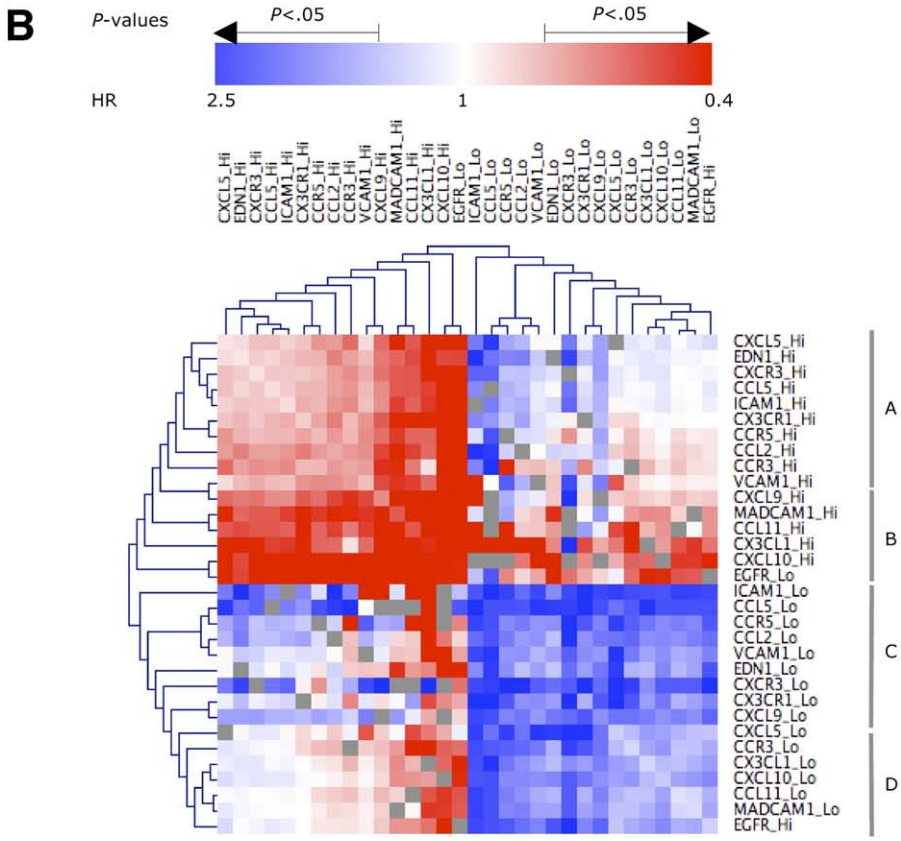
applied the Kruskal–Wallis 1-way analysis of variance to determine if any of the patient cohorts was significantly different regarding the clinical parameters; no significant difference was found between cohorts. All through this article a *P* value less than .05 was considered statistically significant. All analyses were performed with the statistical software R (survival package) and Statview (Cary, NC). For details, see the Supplementary Materials and Methods section.

Results

Immune-Related Genes Are Associated With the Absence of Tumor Recurrence

We first investigated gene expression in colorectal tumors. We determined the median cut-off values for each gene, and performed survival analysis for up to 10 years after primary tumor resection. Log-rank *P* values associated with disease-free survival then were calculated and hazard ratios were illustrated by the size of each node


A
Disease free survival
Survival (months)
CX3CL1
CXCL5
CXCL10
ICAM1 CX3CL1
CXCL9
ICAM1 CXCL10
MADCAM1
CXCL9 MADCAM1
**
ns


B
P-values
P<.05
P<.05
HR
2.5
1
0.4
A
B
C
D

in a network (Figure 1). The expression of genes associated with tumor invasion (*CEACAM1*, *CD97*), metastasis spreading (*ACE*, *EBAG9*, *MMP7*), tumor anti-apoptotic (survivin/*BIRC5*), and angiogenesis (vascular endothelial growth factor) was assessed. Surprisingly, the duration of disease-free survival (DFS) did not correlate significantly with the expression of these tumor-related genes. Host-immune response-related genes, particularly proinflammatory, immunosuppressive, T-helper–specific (T_H1, T_H2), innate, and adaptive immune response—related genes also were assessed. The patterns of expression of proinflammatory-related (*PTGS2*, *IRAK4*), T_H2-related (*GATA3*), and immunosuppression-related (*IL10*, *FoxP3*) genes did not vary according to tumor recurrence. In contrast, innate and adaptive immunity-related genes Granulysin (*GNLY*), Signal Transducer and Activator of Transcription 1 (*STAT1*), Interferon Regulatory Factor 1 (*IRF1*), Interferon Gamma (*IFNG*), T-Box 21/T-bet (*TBX21*), Interleukin 18 Receptor (*IL18RAP*), Inducible T-cell costimulator (*ICOS*). T_H1 (*STAT1*, *IRF1*, *IFNG*, *TBX21*), as well as genes involved in T-cell activation, T_H1, and negative regulation of the immune response (*PDCD1*, *PDCD1LG1*, *PDCD1LG2*) stratified patients into groups with statistically different DFS rates ($P < .05$).

Reconstructed Biomolecular Network Predicts Interacting Chemokines and Adhesion Molecules

Based on the gene expression data we reconstructed a gene–gene network (see Supplementary Materials and Methods section). By using the subset of genes relevant to tumor recurrence and with statistically different DFS (Supplementary Table 2), we further combined publicly available databases and prior knowledge[20] to enrich the network. The prediction of genes was based on conserved genomic neighborhood, phylogenetic profiling, co-expression analysis, protein–protein interaction, functional genomic public databases, and literature co-occurrence. The reconstructed network was visualized (Figure 1) using a network layout visualization that uses GO annotations as a source of external class information to direct the network layout process and to emphasize the biological function of the nodes (Supplementary Table 3).[21]

This integration and visualization of both experimental and in silico data on the network revealed putative functional interactions and new groups of genes associated with the patient's prognosis. Among the prediction of network membership (nodes with red border) were molecules involved in leukocyte and myeloid cell differentiation, the regulation of apoptosis, the protein kinases cascade, adhesion, and chemotaxis (Supplementary Figure 1).[22] To test whether the association between predicted genes and patient survival might be a result of their correlation with the seed genes (used for prediction), we performed STRING analysis without co-expression data. The network constructed without using co-expression information was highly similar to the initial network prediction (Supplementary Table 4).

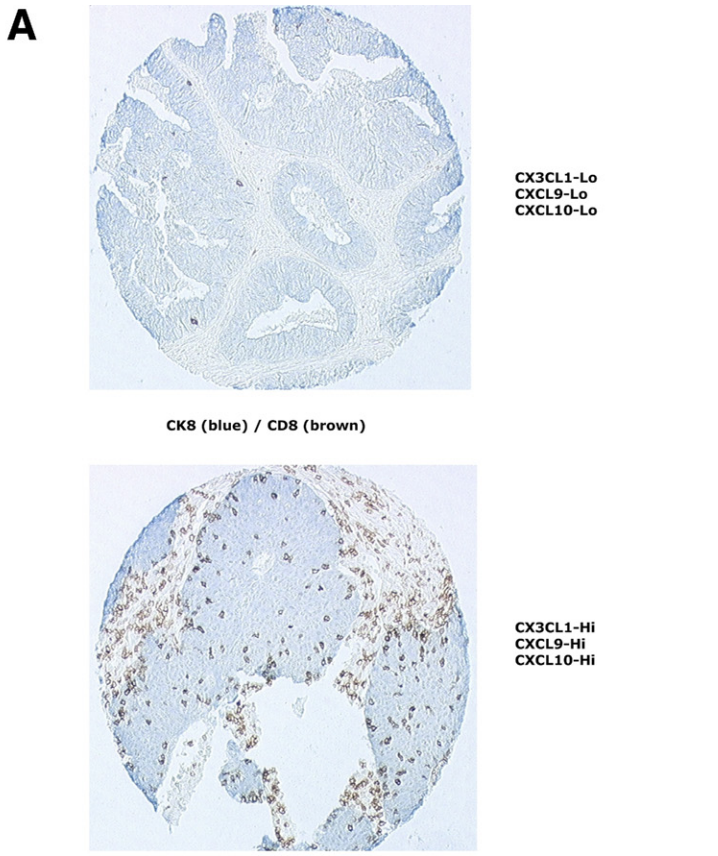
The first top-ranked predicted gene was *CX3CL1*. Other chemokines, such as *CXCL9*, *CXCL10*, *CCL2*, *CCL5*, and *CCL11*, and adhesion molecules, such as *MADCAM1*, *ICAM1*, and *VCAM1*, were predicted to be interacting molecules (Figure 1).

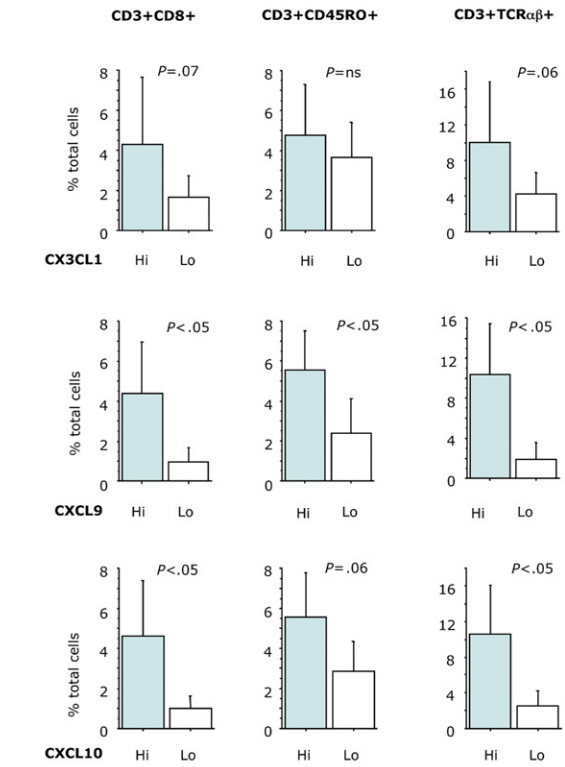
Chemoattractants and Adhesion Molecules Are Associated With Improved Prognosis

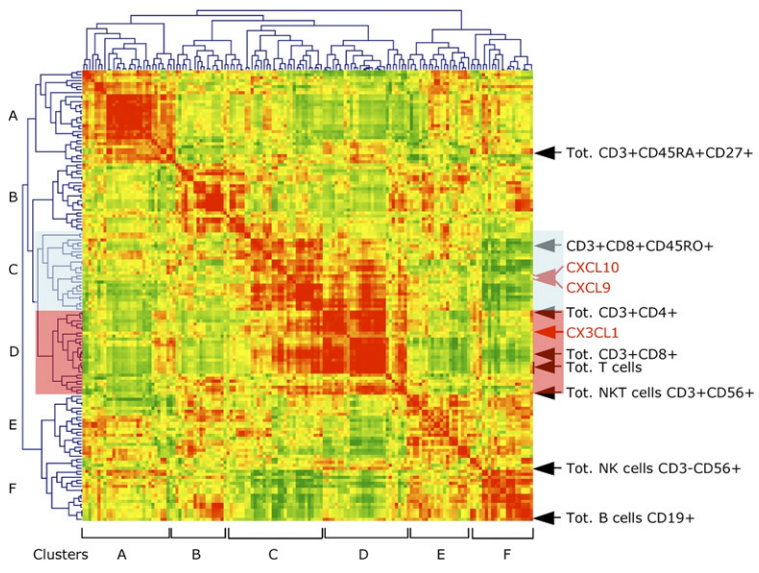
This reconstructed biomolecular network both generated testable hypotheses and predicted novel interactions. Among the predictions of network memberships were chemokines. These molecules could attract distinct cell subpopulations associated with patient survival. To validate the predictions, we analyzed the gene expression of *CX3CL1*, *CXCL9*, and *CXCL10* in primary tumors in 2 independent cohorts (n = 108 and n = 27). In the first cohort, an association between high chemokine expression and improved patient survival (cut-off level at median of the dataset hazard ratio [HR], 2.06, 1.78, and 1.76, respectively; $P < .05$) was observed for each marker (*CX3CL1*, *CXCL9*, and *CXCL10*). The HRs for *CX3CL1*, *CXCL9*, and *CXCL10* were increased (2.21, 2.38, and 2.92, respectively) by using the cut-off value that yielded the minimum *P* value for DFS (Figure 2*A*). Similar results also were found in the second cohort (Supplementary Figure 2).

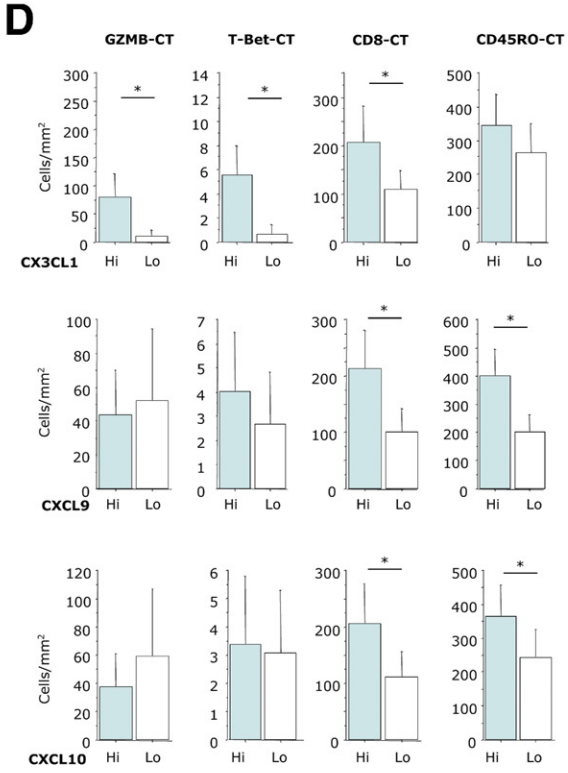
Patients with increased expression of other predicted chemokines and adhesion molecules, such as *CCL2*, *CCL5*, *CCL11*, *ICAM1*, and *MADCAM1*, showed prolonged DFS (HR Lo vs Hi, 1.81–2.21). In contrast, patients with decreased expression of Epidermal Growth Factor Receptor (EGFR) showed prolonged DFS. For control purposes we tested the expression of a nonpredicted chemokine, *CXCL5*. The expression of this chemokine did not vary according to tumor recurrence (Figure 2*A*). To investigate whether the combined analysis of predicted genes could improve the prediction of patient prognosis, we plotted a 2-dimensional hierarchical cluster matrix according to

Figure 2. Gene expression levels from 108 colorectal tumors (cohort 1) were analyzed by real-time quantitative PCR. (*A*) Kaplan–Meier curves for the duration of DFS, according to the expression of the predicted genes (*CX3CL1*, *CXCL9*, *CXCL10*, *CCL2*, *CCL5*, *CCL11*, and *MADCAM1*) were performed. Patients with high (Hi) expression for both genes (*red line*) or low (Lo) expression for both gene densities (*black line*), and heterogeneous expression (HiLo or LoHi, *green line*) are represented. (*B*) HRs were calculated for high and low gene expression compared with the whole cohort (108 patients). A HR-matrix (*heatmap*) followed by unsupervised hierarchical clustering was represented from favorable prognosis: HR, 0.4 (*red*) to poor prognosis HR, 2.5 (*blue*). All HR with HR less than 0.55 or HR greater than 1.66, were significant.


A
CX3CL1-Lo
CXCL9-Lo
CXCL10-Lo
CK8 (blue) / CD8 (brown)
CX3CL1-Hi
CXCL9-Hi
CXCL10-Hi


B
CD3+CD8+
CD3+CD45RO+
CD3+TCRαβ+
% total cells
P=.07
P=ns
P=.06
P<.05
P<.05
P<.05
P=.06
P<.05
CX3CL1
CXCL9
CXCL10
Hi
Lo


C
A
B
C
D
E
F
Clusters
Tot. CD3+CD45RA+CD27+
CD3+CD8+CD45RO+
CXCL10
CXCL9
Tot. CD3+CD4+
CX3CL1
Tot. CD3+CD8+
Tot. T cells
Tot. NKT cells CD3+CD56+
Tot. NK cells CD3-CD56+
Tot. B cells CD19+


D
GZMB-CT
T-Bet-CT
CD8-CT
CD45RO-CT
Cells/mm²
CX3CL1
CXCL9
CXCL10
Hi
Lo
*

the HR between patient groups (Figure 2*B*). By using this method, the DFS time differences found between patients (HiHi vs LoLo) were larger than those found by single chemokine analysis (Hi vs Lo) (HR, 2.30 and 1.78, respectively). The combined analysis of predicted adhesion molecules and chemokines revealed a statistically different DFS as illustrated for *ICAM1*/*CX3CL1*, *ICAM1*/*CXCL10*, and *CXCL9*/*MADCAM1* (Figure 2). Combined analysis revealed 4 major clusters. High levels of genes from cluster A or a combination of genes from cluster A had little impact on patient survival. Patients with low levels of EGFR or high levels of *CXCL10*, *CX3CL1*, *CCL11*, *MADCAM1*, or *CXCL9* (cluster B), had a very favorable prognosis compared with the whole cohort (all HR, <0.61). High levels of the chemokine *CX3CL1*, together with high levels of *CXCL9*, *CXCL10*, *CCL2*, *CCL5*, *CCL11*, *VCAM1*, *ICAM1*, or *MADCAM1*, further increased DFS (HR range, 0.54–0.29). In contrast, patients with a combination of genes from clusters C and D had a poor prognosis compared with the whole cohort (HR, 1.33–2.37). Thus, Kaplan–Meier curves and HR matrix displaying the duration of DFS according to the gene combinations showed that an improved prognosis is associated with the expression of specific chemoattractants and adhesion molecules.

Phenotypes of Intratumoral Immune Cells Correlated With DFS

The intratumoral immune cell infiltrate varies greatly between patients with CRC. To understand the reasons for this heterogeneity, the distribution of infiltrating immune cells was determined in tumors with high and low chemokine gene expression levels (*CX3CL1*, *CXCL9*, *CXCL10*). Immunostaining of tissue microarrays for CD8 T-cell effectors was performed as illustrated in Figure 3 for 2 representative patients with high and low levels of chemokine gene expression, respectively. The tumors of patients with high expression levels of *CX3CL1*, *CXCL9*, and *CXCL10* were found to contain a significantly higher density of CD8 T cells ($P < .05$) (Figure 3*A*). To further investigate functional patterns and the coordination of immune cell populations within the primary tumor in relation to *CX3CL1*, *CXCL9*, and *CXCL10* gene expression, we performed a phenotypic analysis of cell surface markers of the same tumors by flow cytometry. Increased memory CD3+CD45RO+ T-cell infiltration was observed for patients with high expression levels of *CXCL9* and *CXCL10*, but not of *CX3CL1*. Increased CD3+CD8+ and CD3+TCR$\alpha\beta$+ T-cell infiltration furthermore was detected for patients with high *CX3CL1*, *CXCL9*, and *CXCL10* expression levels (Figure 3*B*). Pairwise comparisons were performed by measuring the similarity between profiles using Pearson correlation coefficients. The results of the correlation analysis were visualized using hierarchical clustering of a correlation matrix[18] of 149 cell surface markers analyzed by flow cytometry (Figure 3*C*). The correlation matrix revealed 6 major clusters. *CX3CL1* showed positive correlations (cluster D, $P < .05$ for all combinations) with the total density of infiltrating T cells (total T cells, total CD3+CD8+, total CD3+CD4+). *CXCL9* and *CXCL10* showed a strong positive correlation (cluster C, $R = 0.76$; $P < .01$) and clustered together with a subpopulation of memory CD8 T cells (CD3+CD8+CD45RO+). Other immune cell populations were located in different clusters. For example, naive T cells (total CD3+CD45RA+CD27+) were located in cluster A, and natural killer (NK) (total CD3−CD56+) and B cells (total CD19+) were located in cluster F. These results thus indicate a high degree of functional coordination of specific types of intratumoral immune cells with the expression levels of specific subsets of chemokines.

To confirm these results, we investigated the density of immune cell populations in 2 specific regions of colorectal tumors using tissue microarrays: the CT and the IM of the tumor. Total T lymphocytes (CD3), CD8 T-cell effectors and their associated cytotoxic molecule (Granzyme B, GZMB), subset of activated cytotoxic T cells and NK cells (CD57), memory T cells (CD45RO), T_H1 cells (T-Bet), immature dendritic cells (CD1A), and macrophages (CD68), were in each case quantified by immunostaining (Figure 3*D*). A significant correlation was observed between specific immune cell densities and chemokine expression levels (Table 1, Supplementary Figure 3, and Supplementary Table 5). The tumors of patients with high *CX3CL1* expression levels contained a significantly increased density of effector-activated cytotoxic T cells

Figure 3. (*A*) Tissue-microarray spots of representative tumors with low (top) or high (bottom) *CX3CL1*, *CXCL9*, and *CXCL10* gene expression are illustrated. Stainings for tumor cells (cytokeratin-8+, *blue staining*) and cytotoxic T cells (CD8+, *brown staining*) were performed. (*B*) Comparison of immune cell densities as measured by flow cytometry from 39 freshly resected tumors. Cytotoxic (CD3+CD8+), memory (CD3+CD45RO+), and TCR$\alpha\beta$ (CD3+TCR$\alpha\beta$+) T cells were analyzed in the tumors from patients with high - (*blue histogram*) or low- (*white histogram*) *CX3CL1*, *CXCL9*, or *CXCL10* gene expression. (*C*) Hierarchical clustering of correlation matrix of the flow cytometry data from 39 freshly resected tumors. Pearson correlation coefficients (R) were calculated between the combination of 149 markers for major immune cell populations ("total" prefix) and specific T-cell subpopulations and *CX3CL1*, *CXCL9*, and *CXCL10*. Correlation coefficients were plotted with negative correlation (*green*), positive correlation (*red*), and R = 0 (*yellow*), in matrix representation followed by unsupervised Spearman hierarchical clustering. Six major clusters (A–F) are illustrated. (*D*) Comparison of the mean of immune cell densities (cell/mm^2) as measured by tissue-microarrays from 108 paraffin-embedded tumors. Cytotoxic (CD8+, GZMB+), Th1 (T-Bet+), and memory (CD45RO+) T cells were analyzed in the tumors from patients with high- (*blue histogram*) or low- (*white histogram*) *CX3CL1*, *CXCL9*, or *CXCL10* gene expression. *$P < .05$.

Table 1. Median Immune Cell Densities According to Chemokine Expression Levels

Cells	CCL5-Hi	CCL5-Lo	*P* value	CXCL9-Hi	CXCL9-Lo	*P* value	CXCL10-Hi	CXCL10-Lo	*P* value	CX3CL1-Hi	CX3CL1-Lo	*P* value
CD3-CT	416.5	188	.0007**	367.8	188	.0094**	380.5	191.6	.0046**	377	188	.0163*
CD3intra-CT	42	13	.3912	52	10	.0992(*)	42	10	.3015	53.5	10	.0359*
T.Bet-CT	2.6	1.6	.2976	4.6	0	.0265*	2	1.6	.4149	6.5	0	.0020**
T.Bet-IM	2.5	0	.5264	3.9	0	.4081	3.9	0	.3282	5.3	0	.0315*
GZMB-CT	44.5	10.7	.0539(*)	40	7.3	.1566	40	22	.4298	52.7	0	.0011**
GZMB-IM	124	63.4	.1014	83	71.3	.3000	124	71.3	.1713	139	19.5	.0036**
CD57-CT	44.1	14.8	.0025**	53.1	15.8	.0027**	49.3	13.7	.0013**	40.4	15.9	.0020**
CD8-CT	166.2	46	.0004**	134	46.7	.0042**	134	37.6	.0013**	118.4	53.8	.0542(*)
CD45RO-CT	416	135.5	<.0001**	366.7	136.5	.0025**	380.2	141.8	.0059**	282	143.9	.3483
CD68-CT	921.3	323.3	<.0001**	914.6	375.7	.0001**	864.2	329.3	<.0001**	613.2	438.1	.0967(*)
CD1a-CT	3.1	2	.7600	3.2	1.8	.2560	3.1	1.8	.8923	3.4	1.8	.3298

*.1 > $P \geq .05$
**.05 > $P \geq .01$

and T_H1 cells in situ ($P < .05$ for $GZMB_{CT}$, $T\text{-}Bet_{CT}$, $CD8_{CT}$, $CD57_{CT}$, $CD3_{CT}$, and for $GZMB_{IM}$, $T\text{-}Bet_{IM}$, and $CD57_{IM}$). The tumors of patients with high *CXCL9* and *CXCL10* expression levels, in contrast, contained a significantly increased number of memory CD8 T cells and macrophages in situ ($P < .05$ for *CXCL9* and $CD8_{CT}$, $CD45RO_{CT}$, $CD68_{CT}$, $CD8_{IM}$, $CD45RO_{IM}$, and $P < .05$ for *CXCL10* and $CD3_{CT}$, $CD8_{CT}$, $CD57_{CT}$, $CD68_{CT}$, $CD45RO_{IM}$, and $CD68_{IM}$). Thus, the chemokines *CXCL9*, *CXCL10*, and *CX3CL1* may attract different subsets of immune cells to different locations in the tumor. We performed Cox multivariate analysis combining soluble factors and immune cells. Each chemokine was dependent on the density of T-cell subsets, whereas the density of T-cell subsets remained significant, indicating the relationships between these factors (data not shown). In summary, these data showed that the phenotypes of the intratumoral immune cells are associated strongly with specific chemokines and adhesion molecules, indicating a high degree of functional coordination.

Intratumoral T-Cell Repertoire, Chemokines, and Prognosis

Although total T-cell subpopulations correlated with chemokine levels, one might expect that certain T cells, that selectively are attracted to the tumor site, are specific in nature. To test this hypothesis, we analyzed the T-cell repertoire of tumor-infiltrating T cells for 10 randomly selected patients. All TCR Vβ chain families (Vβ1–Vβ24) and 13 different CDR3 lengths for each Vβ were quantified. Three-dimensional density plots representing the total quantity of each TCR are represented for each patient. The results showed that all Vβ and most CDR3 lengths are present within the tumor of the patients (Figure 4*A* and *C*). They also showed that the amount of each TCR was different for each patient. Notably, the TCR repertoire of patients with a high *CX3CL1* level was clearly distinguishable from the repertoire of patients with a low *CX3CL1* expression level. An example is illustrated for Vβ2 (Figure 4*B* and *D*). The levels of Vβ2 were higher in patients with high *CX3CL1* levels than in patients with low *CX3CL1* levels. By extending the analysis of the specificity of the Vβ2 T cells to the CDR3 lengths, it was found that all patients with high *CX3CL1* levels possessed significantly increased levels of the Vβ2L03, Vβ2L07, Vβ2L06, and Vβ2L09 T-cell receptors than patients with low *CX3CL1* levels ($P < .01$, $P < .008$, $P < .05$, and $P < .05$, respectively) (Figure 5*B*, and data not shown). The other CDR3 lengths for Vβ2 did not differ significantly between patients. To analyze the T-cell repertoire in a global manner, we calculated a correlation matrix between all TCR rearrangements together with chemokine expression levels (Figure 5*E*, Supplementary Table 6). This visualization method revealed 6 major clusters of specific comodulated TCR rearrangements. Subsets of specific T cells thus simultaneously are overrepresented in the same patients. The majority of T cells were found in the same cluster (cluster C). Interestingly, *CX3CL1*, *CXCL9*, and *CXCL10* gene expression levels correlated with specific T-cell rearrangements within another cluster (cluster E). This indicated that these chemokines may attract subsets of T cells with a distinct TCR.

We lastly investigated whether the densities of these specific T cells correlated with patient survival. Specific T cells (eg, Vβ2L04, Vβ5.2L05, or Vβ5.2L04) found in cluster C whose intratumoral densities did not differ according to the level of *CX3CL1*, *CXCL9*, and *CXCL10* gene expression (Figure 5*A*) were not associated with patient survival (Figure 5*C*). In contrast, specific T cells found in cluster E or whose density was higher in tumors with high expression levels of *CX3CL1*, *CXCL9*, or *CXCL10* (eg, Vβ5.2L08, Vβ2L03, or Vβ2L07) (Figure 5*B*) did correlate with patient survival (Figure 5*D*). Overall survival rates at 3 years were 100% and 28%, respectively, in patients with tumors with high and low levels of infiltration by Vβ5.2L08, Vβ2L03, or Vβ2L07 T cells. Kaplan–Meier plots displaying the duration of overall survival (7 years of follow-up evaluation) according to intratumoral T-cell

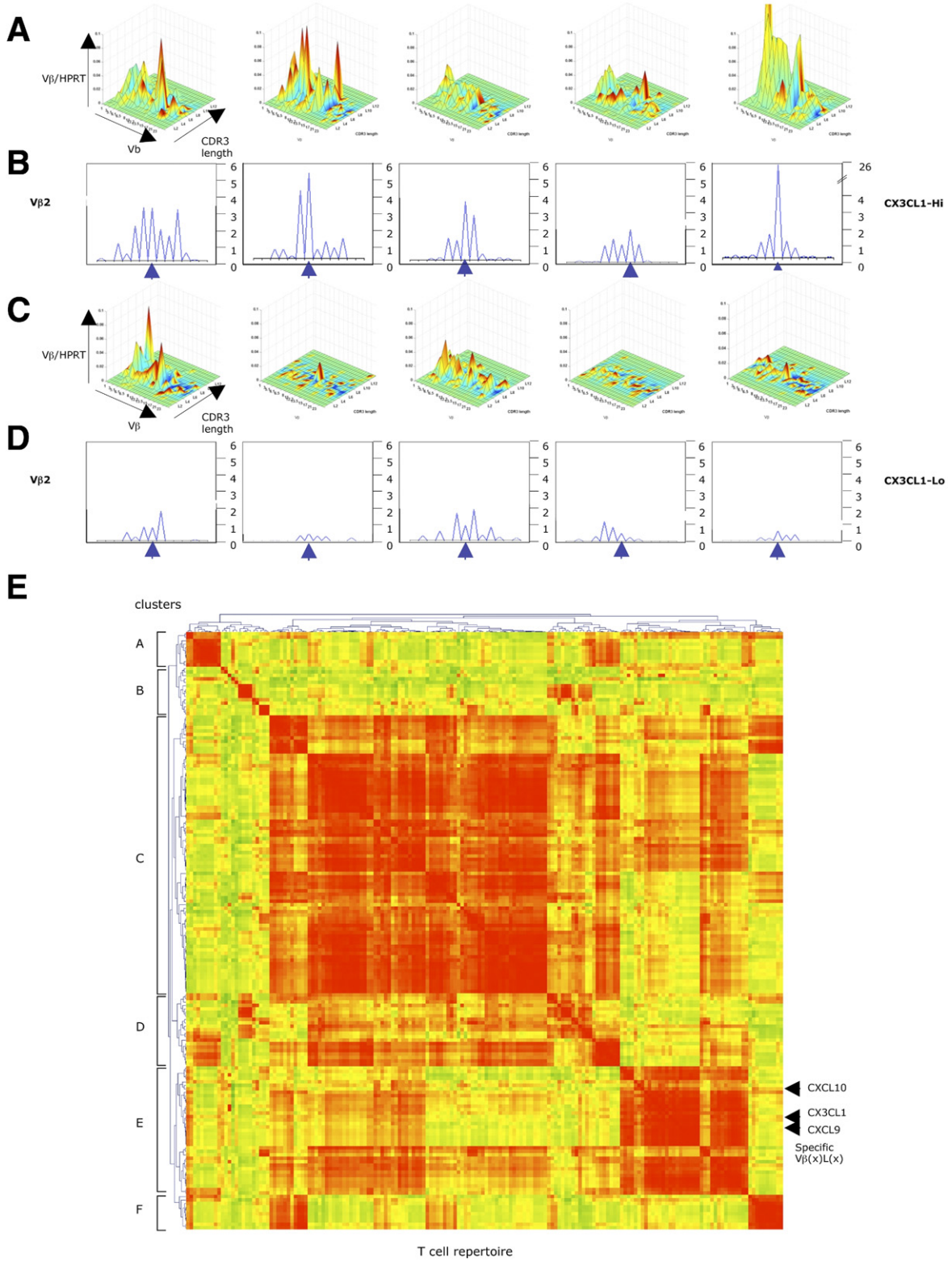


Figure 4. T-cell repertoire analyses were performed on tumors from 10 randomly selected CRC patients using the quantitative expression of the 26 TCR Vβ chain families. Results were expressed in a Vβ/hypoxanthine phosphoribosyltransferase (HPRT) ratio. These quantitative data were represented by the height of the peak in the (*A* and *C*) *TcLandscapes* and on the (*B* and *D*) *histograms*. T-cell repertoire analysis was performed by combining qualitative alterations of Vβ use at the CDR3 length level (13 different CDR3 lengths) with the magnitude of expression of each Vβ mRNA species. (*A* and *C*) The CDR3 lengths were represented in the TcLandscapes. (*E*) Pearson correlation coefficients (R) were calculated between the quantity of all Vβ for each of the 13 different CDR3 lengths and *CX3CL1*, *CXCL9*, and *CXCL10* gene expression. Correlation coefficients were plotted with negative correlation (*green*), positive correlation (*red*), and R = 0 (*yellow*), in matrix representation followed by unsupervised Spearman hierarchical clustering. Six major clusters (A–F) were represented on the matrix.

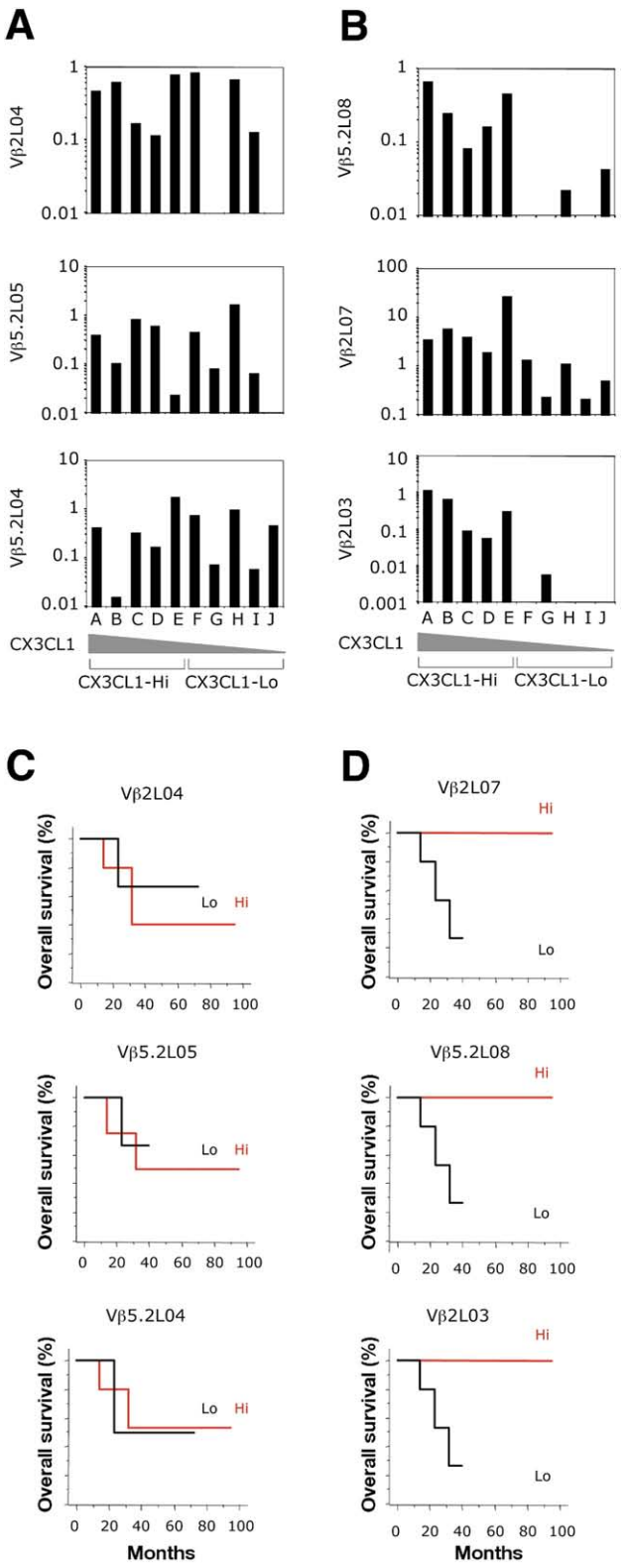


Figure 5. (*A* and *B*) The CDR3 lengths of the TCRs from 10 randomly selected CRC patients (A–J) are represented in the *histograms*. (*C* and *D*) Kaplan–Meier curves illustrate the overall survival of patients according to particular TCR expression at the median of the dataset. (*C*) Three TCRs (Vβ2L04, Vβ5.2L05, or Vβ5.2L04) not associated with *CX3CL1* gene expression, and (*D*) 3 TCRs (Vβ5.2L08, Vβ2L03, and Vβ2L07) significantly increased in patients with high expression of *CX3CL1*.

densities showed the improved prognosis associated with specific T cells (Figure 5*D*).

Discussion

The staggering complexity of multifactorial diseases such as cancer poses significant challenges for the development of stratified or personalized therapies. The integrated analysis of diverse datasets might circumvent these challenges and provide an enhanced understanding of complex systems, such as the tumor microenvironment. We applied such an integrated approach and performed global analyses of the phenome (large-scale flow cytometry experiments), transcriptome, tissue microarrays of specific tumor regions, and T-cell repertoire analysis in the tumor microenvironment of patients with CRC. Our data revealed mechanisms resulting in high or low densities of specific immune cells at the tumor site. Chemokines and adhesion molecules appeared to target immune effector T cells with a specific TCR repertoire within the tumor. Furthermore, the presence within a tumor of T cells with a specific TCR repertoire correlated with patient survival.

It has been proposed that the limitations of individual studies that are owing to experimental design can be overcome by analyzing data obtained from 2 or more different approaches.[23] Our study, in which we have obtained concordant results using different experimental approaches, shows the potential of this general approach. We provide support for hypothesis-driven research in human beings using prior knowledge and integrative biology. The prediction of genes associated with prognosis was based on biomolecular network reconstruction using different data sets (conserved genomic neighborhood, phylogenetic profiling, co-expression analysis, protein–protein interaction, functional genomic public database, literature co-occurrence). Interpretation of our data was facilitated by a novel visualization method combining HR values, a structured description of known biologic information at different levels of granularity (GO), and tools for data integration and visualization (Cytoscape). The functional patterns of biological markers that we uncovered led us to formulate hypotheses associating specific chemokines with immune cells found at the tumor site. Specifically, predictions for the roles of specific chemokines (*CX3CL1*, *CXCL9*, *CXCL10*, *CCL2*, *CCL5*, and *CCL11*) and adhesion molecules (*MADCAM1*, *ICAM1*, and *VCAM1*) were corroborated experimentally.

Binding of chemokines and adhesion molecules such as *CX3CL1* and *MADCAM1* to their receptors (*CX3CR1* and *A4β7* integrin, respectively)[24,25] controls the recruitment and the adhesion of effector CD8+ T cells to the intestine. *CX3CL1* binds to CD8+ T cells, displaying strong cytotoxicity and expressing Perforin+, GZMB+, CD57+, CD11a+, CCR7-, and CD62L-. We showed that the combined expression of *CX3CL1* and other predicted chemokines or *MADCAM1* induced a significant delay in

tumor recurrence. Our data suggest that, in CRC, the majority of cells attracted by *CX3CL1* were T_H1 and effector-activated cytotoxic T cells, whereas *CXCL9* and *CXCL10* attracted mostly memory CD45RO+ T cells. We previously showed that patients without tumor recurrence had higher memory T-cell densities than those whose tumors had recurred.[18] Patients with increased densities of CD57+ or T-Bet+ cells in each tumor region presented with statistically prolonged DFS. Combinations of effector-activated cytotoxic T cells (CD8, GZM, CD57, and T-Bet) and memory T cells (CD45RO) led to increasing HRs for DFS (data not shown). Several reports analyzed chemokines or adhesion molecules in CRC. Chemokine receptors (*CXCR4*, *CXCR5*, *CCR7*) and adhesion molecules (*CD44*, E-cadherin) previously have been associated with tumor invasion and bad prognosis.[26,27] In 2 studies, chemokines (*CX3CL1*, *CXCL16*) were associated with good prognosis.[28,29] However, the power of the reconstructed predictive biomolecular network approach we described herein is that it simultaneously integrates multiple molecules, including specific chemokines and adhesion molecules that could act in a coordinated manner. These results suggested possible mechanisms resulting in high or low densities of specific immune cells in CRC. Thus, the combined expression of chemokines and adhesion molecules, and the resultant density of T-cell subsets within the primary tumor, may prevent tumor expansion and recurrence.

The constant genomic metamorphosis of tumor cells[30] eventually may give rise to new phenotypes that display increased or reduced immunogenicity. The subset of effector cytotoxic T cells found within the tumor, however, is likely to recognize multiple antigens expressed by the tumor cells. Thus, tumor infiltration by cytotoxic and memory T lymphocytes could reflect a level of antitumor immunity shaped by multiple tumor parameters, such as altered expression level of HLA molecules, the expression pattern of tumor antigens, and the mutational pathways (microsatellite instability, chromosomal unstability methylator phenotype (CIMP), chromosomal instability).[5,7] Several studies previously reported oligoclonal or polyclonal intratumoral T-cell repertoire in solid tumors.[31] However, none of these studies analyzed the intratumoral TCR repertoire in relation with chemokine expression or prognosis. First, by analyzing the TCR use, we clearly found a highly polyclonal intratumoral T-cell repertoire because all Vβ and most of the CDR3 lengths were present within the tumor. Second, it can be hypothesized that a majority of T cells are inflammation-related (cluster C), whereas *CXCL9*, *CXCL10*, and *CX3CL1* chemokines attract a subset of specific antitumor T cells with a particular repertoire (cluster E). Third, the presence of a specific intratumoral TCR repertoire correlated with the survival of the patient.

In summary, we revealed mechanisms determining the densities (high or low) of specific immune cells in colorectal tumors. Chemokines and adhesion molecules appeared to target immune effector T cells with particular TCR repertoire within the tumor. Furthermore, the presence of a specific intratumoral TCR repertoire correlated with the survival of the patient. This study also suggests that the creation of similar predictive networks could be used to focus biologic research on specific molecules or pathways in a broad range of physiologic and pathologic processes.

Supplementary Material

Note: To access the supplementary material accompanying this article, visit the online version of *Gastroenterology* at www.gastrojournal.org, and at doi: 10.1053/j.gastro.2009.10.057.

References

1. Benoist C, Germain RN, Mathis D. A plaidoyer for "systems immunology". Immunol Rev 2006;210:229–234.
2. Oltvai ZN, Barabasi AL. Systems biology. Life's complexity pyramid. Science 2002;298:763–764.
3. Hanahan D, Weinberg RA. The hallmarks of cancer. Cell 2000; 100:57–70.
4. Fearon ER, Vogelstein B. A genetic model for colorectal tumorigenesis. Cell 1990;61:759–767.
5. Banerjea A, Bustin SA, Dorudi S. The immunogenicity of colorectal cancers with high-degree microsatellite instability. World J Surg Oncol 2005;3:26.
6. Finn OJ. Human tumor antigens, immunosurveillance, and cancer vaccines. Immunol Res 2006;36:73–82.
7. Finn OJ. Cancer immunology. N Engl J Med 2008;358:2704–2715.
8. Coussens LM, Werb Z. Inflammation and cancer. Nature 2002; 420:860–867.
9. Dunn GP, Old LJ, Schreiber RD. The three Es of cancer immunoediting. Annu Rev Immunol 2004;22:329–360.
10. Shankaran V, Ikeda H, Bruce AT, et al. IFNgamma and lymphocytes prevent primary tumour development and shape tumour immunogenicity. Nature 2001;410:1107–1111.
11. Smyth MJ, Thia KY, Street SE, et al. Differential tumor surveillance by natural killer (NK) and NKT cells. J Exp Med 2000;191: 661–668.
12. Koebel CM, Vermi W, Swann JB, et al. Adaptive immunity maintains occult cancer in an equilibrium state. Nature 2007;450: 903–907.
13. Atreya I, Neurath MF. Immune cells in colorectal cancer: prognostic relevance and therapeutic strategies. Expert Rev Anticancer Ther 2008;8:561–572.
14. Clemente CG, Mihm MC Jr, Bufalino R, et al. Prognostic value of tumor infiltrating lymphocytes in the vertical growth phase of primary cutaneous melanoma. Cancer 1996;77:1303–1310.
15. Diederichsen AC, Hjelmborg JB, Christensen PB, et al. Prognostic value of the CD4+/CD8+ ratio of tumour infiltrating lymphocytes in colorectal cancer and HLA-DR expression on tumour cells. Cancer Immunol Immunother 2003;52:423–428.
16. Zhang L, Conejo-Garcia JR, Katsaros D, et al. Intratumoral T cells, recurrence, and survival in epithelial ovarian cancer. N Engl J Med 2003;348:203–213.
17. Pages F, Berger A, Camus M, et al. Effector memory T cells, early metastasis, and survival in colorectal cancer. N Engl J Med 2005;353:2654–2666.

18. Galon J, Costes A, Sanchez-Cabo F, et al. Type, density, and location of immune cells within human colorectal tumors predict clinical outcome. Science 2006;313:1960–1964.
19. Galon J, Fridman WH, Pages F. The adaptive immunologic microenvironment in colorectal cancer: a novel perspective. Cancer Res 2007;67:1883–1886.
20. von Mering C, Jensen LJ, Snel B, et al. STRING: known and predicted protein-protein associations, integrated and transferred across organisms. Nucleic Acids Res 2005;33:D433–D437.
21. Garcia O, Saveanu C, Cline M, et al. GOlorize: a Cytoscape plug-in for network visualization with Gene Ontology-based layout and coloring. Bioinformatics 2007;23:394–396.
22. Bindea G, Mlecnik B, Hackl H, et al. ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. Bioinformatics 2009;25:1091–1093.
23. Vidal M. A biological atlas of functional maps. Cell 2001;104:333–339.
24. Berlin C, Berg EL, Briskin MJ, et al. Alpha 4 beta 7 integrin mediates lymphocyte binding to the mucosal vascular addressin MAdCAM-1. Cell 1993;74:185–195.
25. Imai T, Hieshima K, Haskell C, et al. Identification and molecular characterization of fractalkine receptor CX3CR1, which mediates both leukocyte migration and adhesion. Cell 1997;91:521–530.
26. Gunther K, Leier J, Henning G, et al. Prediction of lymph node metastasis in colorectal carcinoma by expression of chemokine receptor CCR7. Int J Cancer 2005;116:726–733.
27. Schimanski CC, Schwald S, Simiantonaki N, et al. Effect of chemokine receptors CXCR4 and CCR7 on the metastatic behavior of human colorectal cancer. Clin Cancer Res 2005;11:1743–1750.
28. Hojo S, Koizumi K, Tsuneyama K, et al. High-level expression of chemokine CXCL16 by tumor cells correlates with a good prognosis and increased tumor-infiltrating lymphocytes in colorectal cancer. Cancer Res 2007;67:4725–4731.
29. Ohta M, Tanaka F, Yamaguchi H, et al. The high expression of Fractalkine results in a better prognosis for colorectal cancer patients. Int J Oncol 2005;26:41–47.
30. Sjoblom T, Jones S, Wood LD, et al. The consensus coding sequences of human breast and colorectal cancers. Science 2006;314:268–274.
31. Matsutani T, Shiiba K, Yoshioka T, et al. Evidence for existence of oligoclonal tumor-infiltrating lymphocytes and predominant production of T helper 1/T cytotoxic 1 type cytokines in gastric and colorectal tumors. Int J Oncol 2004;25:133–141.

Received August 4, 2009. Accepted October 29, 2009.

Reprint requests

Address requests for reprints to: Jérôme Galon PhD, Research Director (INSERM), Integrative Cancer Immunology Team, INSERM U872, Paris, France. e-mail: jerome.galon@crc.jussieu.fr; fax: +33 1 4051 0420.

Acknowledgments

The authors are grateful to Dr Ion Gresser and Dr John Nelson for providing helpful comments and critical review of the manuscript.

Conflicts of interest

The authors disclose no conflicts.

Funding

This work was supported by grants from the Association pour la Recherche sur le Cancer, the National Cancer Institute, the Canceropole Ile de France, Ville de Paris, INSERM, the Austrian Federal Ministry of Science and Research (Genome Programme Austria [GEN-AU] project Bioinformatics Integration Network), the Austrian Science Fund (SpezialForschungBereich [SFB] project Lipotoxicity), and the European Commission (Seventh Framework Programme [7FP], Geninca Consortium, grant number 202230).

BIOINFORMATICS APPLICATIONS NOTE Vol. 25 no. 8 2009, pages 1091–1093
doi:10.1093/bioinformatics/btp101

Systems biology

ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks

Gabriela Bindea[1−4,†], Bernhard Mlecnik[1−3,†], Hubert Hackl[4], Pornpimol Charoentong[4], Marie Tosolini[1−3], Amos Kirilovsky[1−3], Wolf-Herman Fridman[1−3,5], Franck Pagès[1−3,5], Zlatko Trajanoski[4] and Jérôme Galon[1−3,5,*]

[1]INSERM, AVENIR Team, Integrative Cancer Immunology, U872, 75006 Paris, [2]Université Paris Descartes, [3]Université Pierre et Marie Curie Paris 6, Cordeliers Research Center, Paris, France, [4]Institute for Genomics and Bioinformatics, Graz University of Technology, Graz, Austria and [5]Assistance Publique-Hôpitaux de Paris, HEGP, Paris, France

Received on November 13, 2008; revised on February 8, 2009; accepted on February 16, 2009

Advance Access publication February 23, 2009

Associate Editor: Trey Ideker

ABSTRACT

Summary: We have developed ClueGO, an easy to use Cytoscape plug-in that strongly improves biological interpretation of large lists of genes. ClueGO integrates Gene Ontology (GO) terms as well as KEGG/BioCarta pathways and creates a functionally organized GO/pathway term network. It can analyze one or compare two lists of genes and comprehensively visualizes functionally grouped terms. A one-click update option allows ClueGO to automatically download the most recent GO/KEGG release at any time. ClueGO provides an intuitive representation of the analysis results and can be optionally used in conjunction with the GOlorize plug-in.

Availability: http://www.ici.upmc.fr/cluego/cluegoDownload.shtml

Contact: jerome.galon@crc.jussieu.fr

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

Since the number of genes that can be analyzed by high-throughput experiments by far exceeded what can be interpreted by a single person, different attempts have been initiated in order to capture biological information and systematically organize the wealth of data. For example Gene Ontology (GO) (Ashburner *et al.*, 2000) annotates genes to biological/cellular/molecular terms in a hierarchically structured way, whereas Kyoto encyclopedia of genes and genomes (KEGG) (Kanehisa *et al.*, 2002) and BioCarta assigns genes to functional pathways. Several functional enrichment analysis tools (e.g. Boyle *et al.*, 2004; Huang *et al.*, 2007; Maere *et al.*, 2005; Ramos *et al.*, 2008; Zeeberg *et al.*, 2003) and algorithms (e.g. Li *et al.*, 2008) were developed to enhance data interpretation.

As most of these tools mainly present their results as long lists or complex hierarchical trees, we aimed to develop ClueGO a Cytoscape (Shannon *et al.*, 2003) plug-in to facilitate the biological interpretation and to visualize functionally grouped terms in the form of networks and charts. Other tools like BiNGO (Maere *et al.*, 2005) or PIPE (Ramos *et al.*, 2008) assess overrepresented GO terms and reconstruct the hierarchical ontology tree, whereas ClueGO uses kappa statistics to link the terms in the network. Compared with the approach of Ramos *et al.* (2008) which creates an *in silico* annotation network based on pathways and protein interaction data and maps the gene list of interest afterwards, ClueGO generates a dynamical network structure by already initially considering the gene lists of interest. ClueGO integrates GO terms as well as KEGG/BioCarta pathways and creates a functionally organized GO/pathway term network. A variety of flexible restriction criteria allow for visualizations in different levels of specificity. In addition, ClueGO can compare clusters of genes and visualizes their functional differences. ClueGO takes advantage of Cytoscape's versatile visualization framework and can be used in conjunction with the GOlorize plug-in (Garcia *et al.*, 2007).

2 METHODS AND IMPLEMENTATION

ClueGO has two major features: it can be either used for the visualization of terms corresponding to a list of genes, or the comparison of functional annotations of two clusters.

2.1 Data import

Gene identifier sets can be directly uploaded in simple text format or interactively derived from gene network graphs visualized in Cytoscape. ClueGO supports several gene identifiers and organisms by default and is easy extendable for additional ones in a plug-in like manner (Supplementary Material).

2.2 Annotation sources

To allow a fast analysis, ClueGO uses precompiled annotation files including GO, KEGG and BioCarta for a wide range of organisms. A one-click update feature automatically downloads the latest ontology and annotation sources and creates new precompiled files that are added to the existing ones. This ensures an up-to-date functional analysis. Additionally ClueGO can easily integrate new annotation sources in a plug-in like way (Supplementary Material).

2.3 Enrichment tests

ClueGO offers the possibility to calculate enrichment/depletion tests for terms and groups as left-sided (Enrichment), right-sided (Depletion) or two-sided (Enrichment/Depletion) tests based on the hypergeometric distribution.

*To whom correspondence should be addressed.

†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

Furthermore it provides options to calculate mid-*P*-values and doubling for two-sided tests to deal with discreetness and conservatism effects as suggested by (Rivals *et al.*, 2007). To correct the *P*-values for multiple testing several standard correction methods are proposed (Bonferroni, Bonferroni step-down and Benjamini-Hochberg).

2.4 Network generation and visualization

To create the annotations network ClueGO provides predefined functional analysis settings ranging from general to very specific ones. Furthermore, the user can adjust the analysis parameters to focus on terms, e.g. in certain GO level intervals, with particular evidence codes or with a certain number and percentage of associated genes. An optional redundancy reduction feature (Fusion) assesses GO terms in a parent–child relation sharing similar associated genes and preserves the more representative parent or child term. The relationship between the selected terms is defined based on their shared genes in a similar way as described by Huang *et al.* (2007). ClueGO creates first a binary gene-term matrix with the selected terms and their associated genes. Based on this matrix, a term–term similarity matrix is calculated using chance corrected kappa statistics to determine the association strength between the terms. Since the term–term matrix is of categorical origin, kappa statistic was found to be the most suitable method. Finally, the created network represents the terms as nodes which are linked based on a predefined kappa score level. The kappa score level threshold can initially be adjusted on a positive scale from 0 to 1 to restrict the network connectivity in a customized way. The size of the nodes reflects the enrichment significance of the terms. The network is automatically laid out using the Organic layout algorithm supported by Cytoscape. The functional groups are created by iterative merging of initially defined groups based on the predefined kappa score threshold. The final groups are fixed or randomly colored and overlaid with the network. Functional groups represented by their most significant (leading) term are visualized in the network providing an insightful view of their interrelations. Also other ways of selecting the group leading term, e.g. based on the number or percentage of genes per term are provided. As an alternative to the kappa score grouping the GO hierarchy using parent–child relationships can be used to create functional groups.

When comparing two gene clusters, another original feature of ClueGO allows to switch the visualization of the groups on the network to the cluster distribution over the terms. Besides the network, ClueGO provides overview charts showing the groups and their leading term as well as detailed term histograms for both, cluster specific and common terms.

Like BiNGO, ClueGO can be used in conjuntion with GOlorize for functional analysis of a Cytoscape gene network. The created networks, charts and analysis results can be saved as project in a specified folder and used for further analysis.

3 CASE STUDY

To demonstrate how ClueGO assesses and compares biological functions for clusters of genes we selected up- and down-regulated natural killer (NK) cell genes in healthy donors from an expression profile of human peripheral blood lymphocytes (GSE6887, Gene Expression Omnibus). For upregulated NK genes ClueGO revealed specific terms like ‘Natural killer cell mediated cytotoxicity’ in the group ‘Cellular defense response’. Downregulated in NK cells compared with the reference (a pool of all immune cell types) were genes involved in the innate immune response (Macrophages), but also in the adaptive immune response (T and B cell). The common functionality refers to characteristics of leukocytes (chemotaxis), besides other terms involved in cell division and metabolism (Fig. 1).

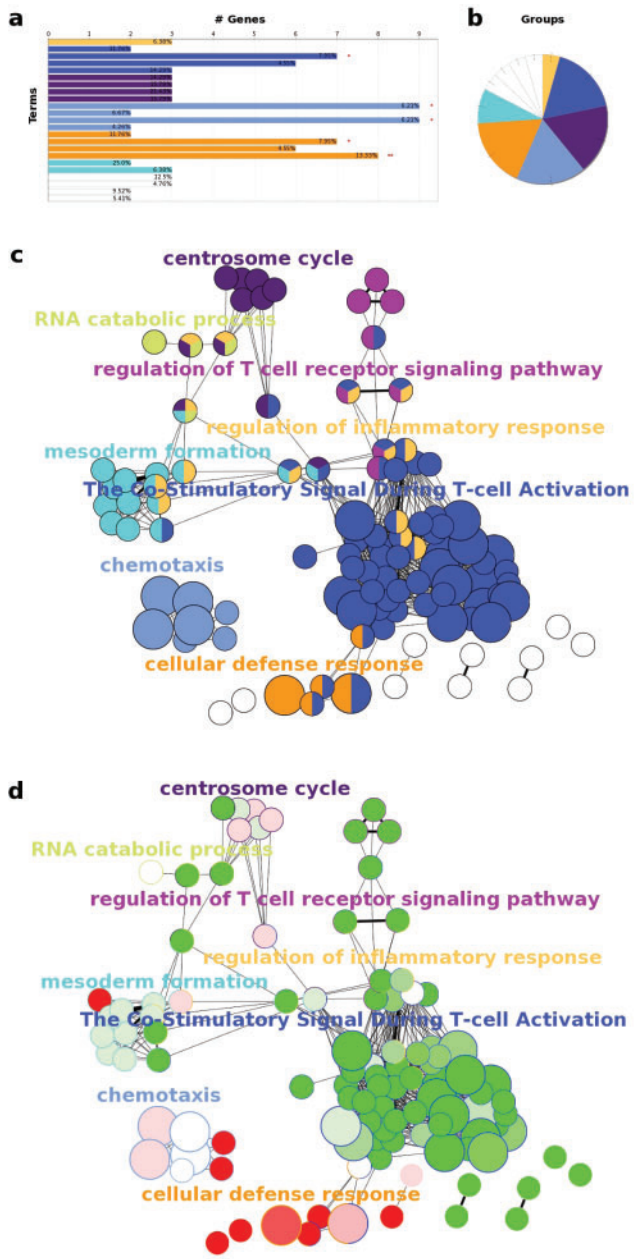


Fig. 1. ClueGO example analysis of up- and down-regulated NK cell genes in peripheral blood from healthy human donors. (a) GO/pathway terms specific for upregulated genes. The bars represent the number of genes associated with the terms. The percentage of genes per term is shown as bar label. (b) Overview chart with functional groups including specific terms for upregulated genes. (c) Functionally grouped network with terms as nodes linked based on their kappa score level ($\geq$0.3), where only the label of the most significant term per group is shown. The node size represents the term enrichment significance. Functionally related groups partially overlap. Not grouped terms are shown in white. (d) The distribution of two clusters visualized on network (c). Terms with up/downregulated genes are shown in red/green, respectively. The color gradient shows the gene proportion of each cluster associated with the term. Equal proportions of the two clusters are represented in white.

4 SUMMARY

ClueGO is a user friendly Cytoscape plug-in to analyze interrelations of terms and functional groups in biological networks. A variety of flexible adjustments allow for a profound exploration of gene clusters in annotation networks. Our tool is easily extendable to new organisms and identifier types as well as new annotation sources which can be included in a transparent, plug-in like manner. Furthermore, the one-click update feature of ClueGO ensures an up-to-date analysis at any time.

ACKNOWLEDGEMENTS

We thank A Van Cortenbosch for the name of the tool.

Funding: INSERM; Ville de Paris; INCa; the Austrian Ministry for Science and Research, Project GEN-AU; BINII; the European 7FP Grant Agreement 202230 (GENINCA).

Conflict of Interest: none declared.

REFERENCES

Ashburner,M. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.

Boyle,E.I. *et al.* (2004) GO::TermFinder–open source software for accessing Gene Ontology information and finding significantly enriched gene ontology terms associated with a list of genes. *Bioinformatics*, **20**, 3710–3715.

Garcia,O. *et al.* (2007) GOlorize: a cytoscape plug-in for network visualization with Gene Ontology-based layout and coloring. *Bioinformatics*, **23**, 394–396.

Huang,D.W. *et al.* (2007) The DAVID Gene Functional Classification Tool: a novel biological module-centric algorithm to functionally analyze large gene lists. *Genome Biol.*, **8**, R183–R183.

Kanehisa,M. *et al.* (2002) The KEGG databases at GenomeNet. *Nucleic Acids Res.*, **30**, 42–46.

Li,Y. *et al.* (2008) A global pathway crosstalk network. *Bioinformatics*, **24**, 1442–1447.

Maere,S. *et al.* (2005) BiNGO: a Cytoscape plugin to assess overrepresentation of Gene Ontology categories in biological networks. *Bioinformatics*, **21**, 3448–3449.

Ramos,H. *et al.* (2008) The protein information and property explorer: an easy-to-use, rich-client web application for the management and functional analysis of proteomic data. *Bioinformatics*, **24**, 2110–2111.

Rivals,I. *et al.* (2007) Enrichment or depletion of a GO category within a class of genes: which test? *Bioinformatics*, **23**, 401–407.

Shannon,P. *et al.* (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.*, **13**, 2498–2504.

Zeeberg,B.R. *et al.* (2003) GoMiner: a resource for biological interpretation of genomic and proteomic data. *Genome Biol.*, **4**, R28–R28.

Coordination of Intratumoral Immune Reaction and Human Colorectal Cancer Recurrence

Matthieu Camus,[1,2,3] Marie Tosolini,[1,2,3] Bernhard Mlecnik,[1,2,3] Franck Pagès,[1,2,3,4] Amos Kirilovsky,[1,2,3] Anne Berger,[5] Anne Costes,[1,2,3] Gabriela Bindea,[1,2,3,7] Pornpimol Charoentong,[7] Patrick Bruneval,[6] Zlatko Trajanoski,[7] Wolf-Herman Fridman,[1,2,3,4] and Jérôme Galon[1,2,3]

[1]Integrative Cancer Immunology INSERM AVENIR Team 15, INSERM U872; [2]Cordeliers Research Centre, Université Pierre et Marie Curie Paris 6; [3]Université Paris-Descartes; Departments of [4]Immunology, [5]General and Digestive Surgery, and [6]Pathology, Georges Pompidou European Hospital, Paris, France; and [7]Institute for Genomics and Bioinformatics, Graz University of Technology, Graz, Austria

Abstract

A role for the immune system in controlling the progression of solid tumors has been established in several mouse models. However, the effect of immune responses and tumor escape on patient prognosis in the context of human cancer is poorly understood. Here, we investigate the cellular and molecular parameters that could describe *in situ* immune responses in human colorectal cancer according to clinical parameters of metastatic lymph node or distant organ invasion (META− or META+ patients). Primary tumor samples of colorectal carcinoma were analyzed by integrating large-scale phenotypic (flow cytometry, 39 patients) and gene expression (real time reverse transcription-PCR, 103 patients) data sets related to immune and protumoral processes. In META− colorectal cancer primary tumors with high densities of T cells, we observed significant positive correlations between markers of innate immune cells [tumor-associated macrophages, dendritic cells, natural killer (NK) cells, and NKT cells] and markers of early-activated T cells. Significant correlations were also observed between markers of cytotoxic and effector memory T-cell subpopulations. These correlation profiles were absent in tumors with low T-cell infiltrates and were altered in META+ tumors with high T-cell infiltrates. We show that the coexpression of genes mediating cytotoxicity (*GNLY*) and Th1 adaptive immune responses (*IRF1*) accurately predicted patient survival independently of the metastatic status. High intratumoral mRNA expression of the proangiogenic mediator vascular endothelial growth factor was associated with significantly reduced survival rates in patients expressing high mRNA levels of *GNLY*. Investigation of the colorectal cancer primary tumor microenvironment allowed us to uncover the association of favorable outcomes with efficient coordination of the intratumoral immune response. [Cancer Res 2009;69(6):2685–93]

Introduction

Cancer progression is a complex process involving host-tumor interactions through multiple molecular and cellular factors of the tumor microenvironment (1). Tumors may be vulnerable to immune destruction. As revealed by experiments in immune-deficient mice, immune responses mediated by IFNγ (2, 3) and cytotoxic mediators such as perforin (4, 5) secreted by lymphocytes are involved in cancer immunosurveillance (6, 7). In human cancer, complex tumor-host interactions are less well documented. However, lymphocytes were also shown to participate in antitumoral responses (8). Consistent with findings in melanoma (9) and ovarian cancer (10, 11), tumor-infiltrating T cells were associated with improved clinical outcome and survival in colorectal cancer patients (12–16).

We recently highlighted intratumoral memory T cells as the major immune effector cells significantly associated with the decrease of early metastatic events (tumor emboli) and the prevention of relapse in colorectal cancer patients (17). Furthermore, we revealed the importance to patient prognosis of the nature, functional orientation, density, and localization of immune cell populations within the primary tumor. Multivariate Cox analysis showed that immune patterns remained the unique parameter significantly associated with prognosis, whereas T stage, N stage, and differentiation of the tumor were not significant when adjusted to immune patterns (18). Patients with cancers at nonmetastatic stages had prognoses as bleak as patients with metastatic tumors, if presenting a low intratumoral adaptive immune reaction. Conversely, patients with metastatic tumors eliciting a high intratumoral immune reaction were of better prognosis. Thus, the amplitude of adaptive immune reaction within the primary tumor was a better predictor of survival than traditional clinical parameters (19).

However, the intrinsic capability of tumor cells to promote their own development (20) may allow tumors to overwhelm immune system activity. For instance, angiogenesis mediated by vascular endothelial growth factor (VEGF) is critical to the growth (by providing oxygen and nutrients) and malignant dissemination (providing a route for metastases) of solid tumors (21, 22). Furthermore, under the pressure of antitumoral immune activity, selection and outgrowth of variant tumor cells with reduced immunogenicity could occur (8, 22, 23). Thus, during cancer progression, tumor cells may acquire immune tolerance mechanisms by generating complex immunosuppressive networks at the tumor site (24, 25) involving interleukin (IL)-10 and transforming growth factor β (TGFβ; refs. 26, 27) as well as T-cell–specific coinhibitory molecules (CTLA-4 and PD-1; refs. 28, 29).

In this work, we attempted to describe in a comprehensive manner the immune reaction in primary colorectal tumors of patients with high or low densities of infiltrating T cells. Furthermore, we compared the immune microenvironment in patients presenting with invaded lymph nodes and/or distant metastases [META+ patients: Union Internationale Contre le

Note: Supplementary data for this article are available at Cancer Research Online (http://cancerres.aacrjournals.org/).

Requests for reprints: Jérôme Galon, AVENIR Team 15, Cordeliers Research Centre, INSERM U872, 15 rue de L'Ecole de Médecine, 75006 Paris, France. Phone: 33-1-5310-0404; Fax: 33-1-4051-0420; E-mail: jerome.galon@crc.jussieu.fr.

doi:10.1158/0008-5472.CAN-08-2654

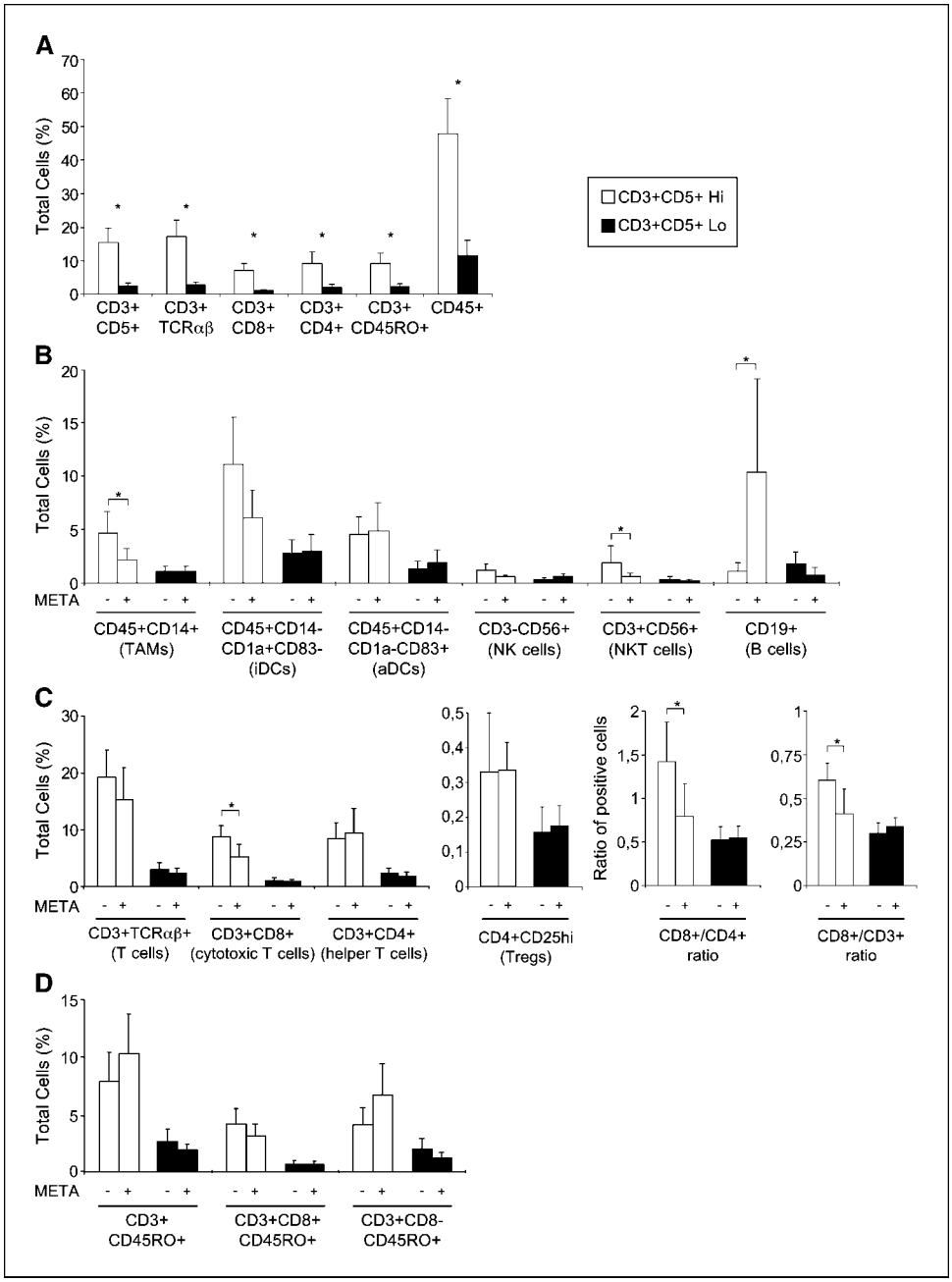


Figure 1. Immune cell populations within primary colorectal tumors. Patients (n = 39) were classified according to the mean percentage of $CD3^+CD5^+$ cells among total cells within tumors (*white columns,* $CD3^+CD5^+$Hi; *black columns,* $CD3^+CD5^+$Lo) and the metastatic status (META− Hi, n = 6; META− Lo, n = 10; META+ Hi, n = 7; META+ Lo, n = 16). Cell populations were represented as the mean percentage of positive cells; *bars,* SE. *, $P < 0.05$, Mann-Whitney test. *A,* T cells ($CD3^+CD5^+$, $CD3^+TCR\alpha\beta^+$), cytotoxic T cells ($CD3^+CD8^+$), helper T cells ($CD3^+CD4^+$), memory T cells ($CD3^+CD45RO^+$), and lymphoid cells ($CD45^+$). *B,* tumor-associated macrophages (*TAMs*; $CD45^+CD14^+$), immature dendritic cells (*iDCs*; $CD45^+CD1a^+CD14^-CD83^-$), activated dendritic cells (*aDCs*; $CD45^+CD1a^-CD14^-CD83^+$), NK cells ($CD3^-CD56^+$), NKT cells ($CD3^+CD56^+$), and B cells ($CD19^+$). *C, left,* T cells ($CD3^+TCR\alpha\beta^+$), cytotoxic T cells ($CD3^+CD8^+$), helper T cells ($CD3^+CD4^+$), and regulatory T cells (*Tregs*: $CD4^+CD25^{hi}$). *Right,* ratios of $CD8^+/CD4^+$ and $CD8^+/CD3^+$ cell subpopulations. *D,* memory T cells ($CD3^+CD45RO^+$), cytotoxic memory T cells ($CD3^+CD8^+CD45RO^+$), and helper memory T cells ($CD3^+CD4^+CD45RO^+$).

Cancer (UICC) tumor-node-metastasis (TNM) stages III–IV] or without such metastases (META− patients: UICC-TNM stages I–II; ref. 30). We analyzed immune cell phenotypic clusters, or "phenoclusters" (31), obtained by grouping markers according to similar levels of expression. This allowed us to uncover functional marker patterns of efficient and coordinated antitumoral immune responses that represent powerful prognostic criteria for colorectal cancer clinical outcome. At the cellular level, a high degree of functional coordination between intratumoral immune cells could be observed at the primary tumor sites of both META− and META+ colorectal cancer patients. At the molecular level, the coexpression of genes related to the Th1 immune response [IFN-regulatory factor 1 (*IRF1*)] and cytotoxicity [granulysin (*GNLY*)] had strong prognostic values. Finally, we studied several tumor-promoting mechanisms including immunosuppression, angiogenesis, tumor survival, and local and metastatic invasion. Analysis of *in situ* gene expression of protumoral markers in combination with immune parameters revealed that angiogenesis (*VEGF*) was associated with increased risks of colorectal cancer relapse in patients nonetheless presenting evidence of strong intratumoral immune responses.

Materials and Methods

All details about Materials and Methods are available online.

Patients and database. Patients with colorectal cancer (n = 566) who underwent a primary resection at the Laennec/HEGP Hospital between 1986 and 2004 were randomly selected. Time to recurrence or disease-free time was defined as the time period from the date of surgery to confirmed tumor relapse date for relapsed patients and from the date of surgery to the date of last follow-up for disease-free patients.

Large-scale flow cytometric analysis. Cells were extracted from 39 fresh tumors, resuspended in PBS/0.5% bovine serum albumin and incubated for 30 min at 4°C with antibodies against immune cell markers for large-scale phenotypic analysis of T cells and with relevant isotype controls. Analyses were done with a four-color fluorescence-activated cell sorter (FACSCalibur, Becton Dickinson) and CellQuest software (Becton Dickinson). Analyzed markers are presented in Supplementary Fig. S3. Complete-linkage hierarchical clustering was applied and the results were displayed with the use of the Genesis program (32, 33). Correlation matrices were constructed by calculation of Pearson correlation coefficients for all marker combinations, followed by unsupervised hierarchical clustering.

Real-time reverse transcription-PCR assay. Tissue samples were snap-frozen. Total RNA was extracted by homogenization with RNeasy isolation kit (Qiagen). The integrity and the quantity of the RNA were evaluated on Bioanalyzer-2100 (Agilent Technologies). Samples (n = 103) were assessed for gene expression analysis of the following 17 genes (see details about gene expression and name in Supplementary data): *CD3ζ, CD4, CD8α, TBX21, IRF1, IFNγ, GNLY, GZMB, GATA3, FOXP3, CEACAM1, CEA, EBAG9, BIRC5, IL-10, TGFβ*, and *VEGF*. Quantitative real-time TaqMan PCR was done using Low-Density-Arrays and the 7900 robotic real-time PCR system (Applied Biosystems). 18S primers and probes were used as internal controls.

Construction of tissue microarrays. Using a tissue microarray instrument (Beecher Instruments, Alphelys), we removed two representative areas of the tumor (center and invasive margin from paraffin-embedded tissue blocks). Tissue microarrays were cut into 5-μm sections for immunohistochemical staining.

Immunohistochemistry. After antigen retrieval and quenching of endogenous peroxidase activity, sections were incubated for 60 min at room temperature with monoclonal antibodies against CD3 (SP7), CD8 (4B11), CD1a (O10), Ki67 (SP6; Neomarkers), CD68 (PGM1; DAKO), FoxP3 (ab20034; Abcam), and M30 cytoDEATH (Alexis Biochemicals). The Envision+ system (enzyme-conjugated polymer backbone coupled to secondary antibodies) and 3,3′-diaminobenzidine chromogen were applied (DAKO). Tissue sections were counterstained with Harris's hematoxylin. Isotype-matched mouse monoclonal antibodies were used as negative controls.

Statistical analysis. Kaplan-Meier curves were used to assess the influence of immune and tumoral parameters on disease-free survival. The significance of these parameters was assessed by univariate analysis with the use of the log-rank test. To identify markers with significant different levels of expression among tissues, Wilcoxon-Mann-Whitney and t tests (ANOVA) were used. $P < 0.05$ was considered to indicate statistical significance. All analyses were done with the use of statistical software programs R and StatView.

Results

Intratumoral distribution of immune cell populations. We first investigated the immune cellular profiles of patients by flow cytometry with 39 freshly resected primary tumors. Percentages of positive cells for distinct markers were calculated among all cells (tumor and immune cells), thus reflecting the density of cells within tumors. Intratumoral T-cell densities were evaluated according to the mean percentage of double-positive cells for the T-cell–specific markers CD3 and CD5 among all samples. Patients presenting a percentage of $CD3^+CD5^+$ T cells superior to this mean (6.7% of all cells) were named "Hi patients" and otherwise "Lo patients." As a control, the percentages of $CD3^+CD5^+$ and $CD3^+TCR\alpha\beta^+$ T cells were represented (Fig. 1*A*). In Hi patients, there were significant higher densities of T cells of both cytotoxic ($CD3^+CD8^+$) and helper ($CD3^+CD4^+$) phenotypes and of memory

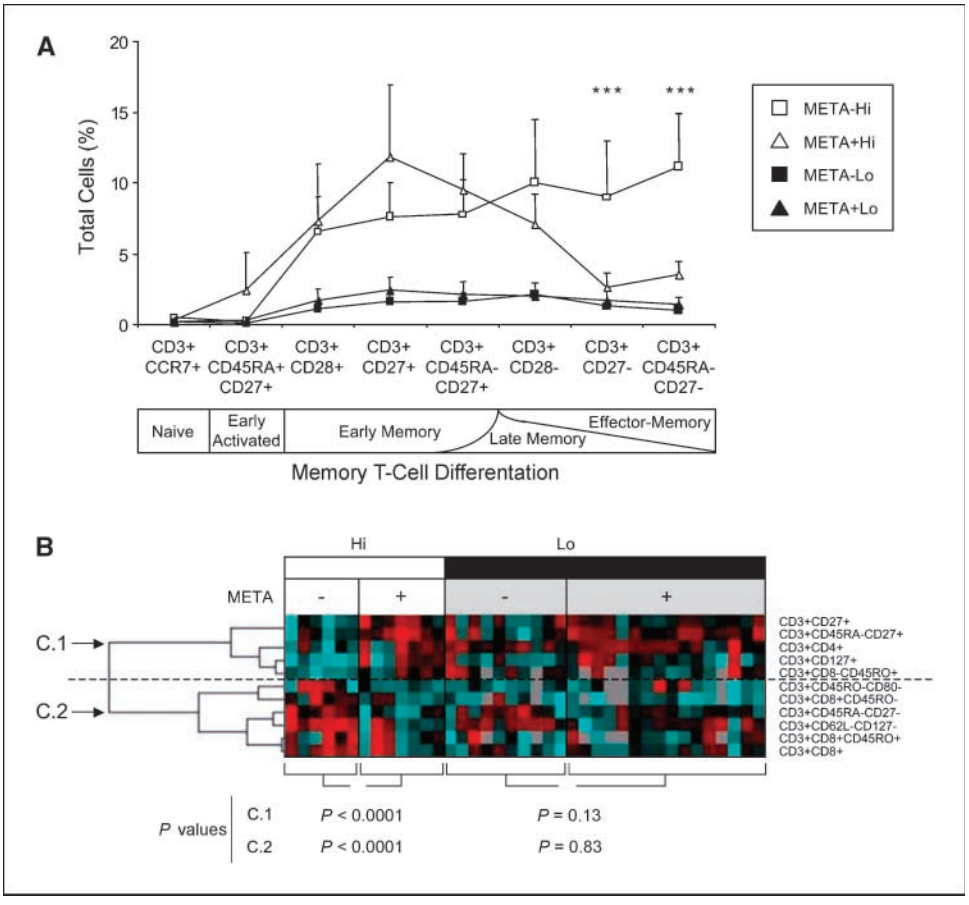


Figure 2. T-cell populations within primary colorectal tumors. *A,* T-cell memory differentiation markers (*white squares,* META− Hi; *white triangles,* META+ Hi; *black squares,* META− Lo; *black triangles,* META+ Lo). Cell populations were represented as the mean percentage of positive cells; *bars,* SE. *P* values (Mann-Whitney test) were presented in Supplementary Fig. S1 (*, $P < 0.05$). *B,* hierarchical clustering of 11 marker combinations among $CD3^+$ cells with significant differential expression among the four groups of patients ($P < 0.05$). Combinations of surface markers were plotted from the minimal (*blue*) to the maximal (*red*) level of expression. *Gray,* not determined.

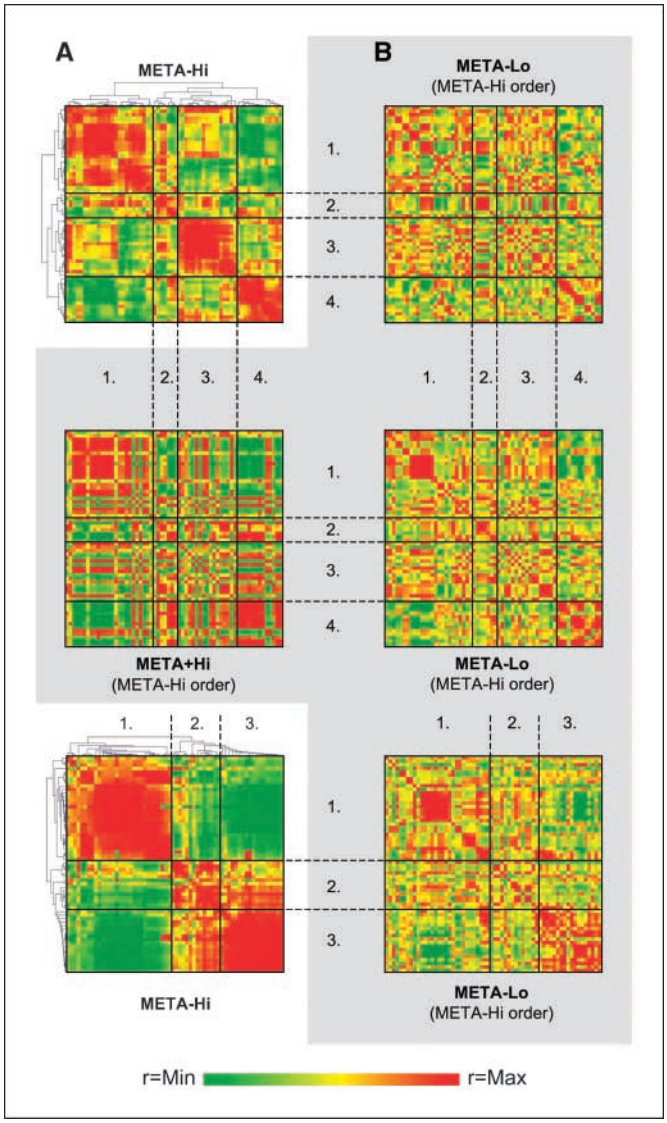


Figure 3. Correlation matrices of flow cytometry data. *P* values and Pearson correlation coefficients (*r*) were calculated between 62 marker combinations that were specific for T cells (markers in combination with CD3) and for major immune cell populations ("total" prefix), presented in Supplementary Fig. S2. *r* values were plotted from *r* = min (*green*) to *r* = max (*red*) in matrix representation, followed by unsupervised hierarchical clustering. Clustered markers were presented in Supplementary Fig. S3. Correlation matrices were independently clustered or arrayed according to the clustering of other correlation matrices (*gray area*). *A, top,* META− Hi patients; *center* and *bottom,* META+ Hi patients. *B, top,* META− Lo patients; *center* and *bottom,* META+ Lo patients.

phenotype ($CD3^+CD45RO^+$) compared with Lo patients (Fig. 1*A*). The distribution of global lymphoid cell populations ($CD45^+$) was consistent with those of T cells (Fig. 1*A*).

We compared the distributions of the major intratumoral immune cell populations according to the metastatic status of the patients: META− patients, no metastases (stages I–II); META+ patients, metastases in lymph node (stage III) and/or distant organ (stage IV). In Lo patients, no differences were found between META− and META+ patients in the distribution of tumor-associated macrophages, immature dendritic cells, activated dendritic cells, natural killer (NK) cells, NKT cells, or B cells. In contrast, in Hi patients, significantly lower percentages of tumor-associated macrophages and NKT cells were observed in META+ Hi patients compared with META− Hi patients. Conversely, B-cell density was significantly higher in META+ Hi patients compared with META− Hi patients (Fig. 1*B*). META+ Hi patients had significantly decreased densities of cytotoxic T cells ($CD3^+CD8^+$) compared with META− Hi patients, whereas no significant differences were observed for helper T cells ($CD3^+CD4^+$) or regulatory T cells ($CD4^+CD25^{hi}$; Fig. 1*C, left*). Finally, $CD8^+/CD4^+$ and $CD8^+/CD3^+$ cell ratios were significantly higher in META− Hi patients compared with META+ Hi and META+ Lo patients (Fig. 1*C, right*).

Memory T-cell differentiation. No differences in the distribution of $CD45RO^+$ memory T-cell subpopulations were observed among the patient groups (Fig. 1*D*). However, we more precisely assessed the density of T cells along memory differentiation steps based on the differential expression of CCR7, CD45RA, CD27, and CD28 markers by $CD3^+$ T cells (Fig. 2*A*). Very few naive ($CCR7^+$) T cells were detected within primary tumors. In META− Lo (*black squares*) and META+ Lo patients (*black triangles*), despite low densities of T cells, similar levels of memory T-cell subpopulations from early ($CD28^+$) to late ($CD45RA^-CD27^-$) memory were observed. META− Hi patients (*white squares*) presented high densities of all memory T-cell subpopulations. In contrast, META+ Hi patients (*white triangles*) presented a significant decrease in the densities of T cells at late stages of memory differentiation ($CD27^-$, $CD45RA^-$) with percentages comparable to Lo patients and significantly inferior to META− Hi patients.

Eleven marker combinations expressed among $CD3^+$ T cells were found significantly differentially expressed between META− Hi and META+ Hi patients. After hierarchical clustering of these markers, two major clusters (C.1 and C.2) were found (Fig. 2*B*). In C.1, $CD4^+$ T-cell subpopulation markers ($CD3^+CD4^+$) and related memory markers ($CD3^+CD8^-CD45RO^+$) grouped with early memory T-cell markers ($CD3^+CD127^+$, $CD3^+CD27^+$, $CD3^+CD45RA^-CD27^+$). In C.2, $CD8^+$ T-cell subpopulation markers ($CD3^+CD8^+$) and related memory/effector T-cell markers ($CD8^+CD45RO^{+/-}$) grouped with effector memory T-cell markers ($CD3^+CD45RA^-CD27^-$) and final effector T-cell markers ($CD3^+CD45RO^-$, $CD3^+CD62L^-CD127^-$). Whereas no distinct pattern was observed in Lo patients, META− Hi patients presented a significant increase of CD8/effector memory T-cell subpopulations (*red squares*; C.2) compared with META+ Hi patients that had a majority of CD4/early memory T cells (C.1). These observations suggested that complete memory T-cell differentiation was associated with a higher proportion of cytotoxic T cells within highly infiltrated tumors and preferentially occurred in META− Hi patients compared with META+ Hi patients.

Association between CD8 T cells and complete memory T-cell differentiation. Evaluation of intratumoral immune coordination was assessed by analyzing the correlations between 62 combinations of cell surface markers of total intratumoral immune cell populations and T-cell subpopulations. For each patient group, pairwise comparisons of the markers were done by measuring Pearson correlation coefficients (*r*) and related *P* values (Supplementary Fig. S2). The relationships implied by these correlations were visualized by using unsupervised hierarchical clustering of *r* values (Fig. 3). The clustered markers were presented in Supplementary Fig. S3. Comparison of META− Hi patients (Fig. 3*A, top*) with other patients was assessed by the construction of META+ Hi (Fig. 3*A, center*), META− Lo (Fig. 3*B, top*), and META+ Lo (Fig. 3*B, center*) correlation matrices arrayed according to META− Hi matrix unsupervised clustering.

META− Hi patients displayed a correlation matrix with four major clusters (Fig. 3*A, top*). Cluster 1 contained markers of total T cells, CD4 T cells, B cells, NK cells, and activated dendritic cells, as well as CD4 T-cell subpopulation markers ($CD4^+$ or $CD8^-$ in combination with $CD25^{+/-}$, $CD26^{+/-}$, $CD103^{+/-}$, $CCR7^-$, $CD45RO^-$) and early memory T-cell markers ($CD28^+$, $CD27^+$, $CD45RA^-CD27^+$). Significant positive correlations were found for the CD4 marker with CD27 and $CD45RA^-CD27^+$ markers ($P = 0.03$ for both correlations). Cluster 2 contained naive $CCR7^+$ T cells. Total CD8 T-cell marker from this cluster positively correlated with total T cells in cluster 1 ($P = 0.03$). In cluster 4, CD8 T-cell subpopulation markers ($CD8^+$ or $CD4^-$ in combination with $CD25^{+/-}$, $CD26^-$, $CD103^+$, $CCR7^-$, $CD45RO^+$) were grouped with markers of late-stage memory T-cell differentiation ($CD45RA^-CD27^-CD127^-CD62L^-$). Significant positive correlations were found between CD8 T-cell subpopulation markers and effector memory T-cell markers ($CD45RA^-CD27^-/CD4^-CD103^+$; $P = 0.02$) and final effector T-cell markers ($CD127^-CD62L^-/CD8^+CCR7^-$; $P = 0.02$). As a control, correlation between CD4 and CD8 T-cell subpopulation markers was negative ($CD3^+CD4^+/CD3^+CD8^+$; $r = -0.871$, $P = 0.02$). Interestingly, cluster 4 (CD8/effector memory T cells) and cluster 1 (CD4/early memory T cells) were, globally, strongly inversely correlated. In cluster 3, early differentiated ($CD45RA^+CD27^+$) and activated ($CD25^+$, $CD26^+$, $CD69^+$) T-cell markers, as well as the global memory T-cell marker ($CD45RO^+$), were grouped with markers of innate immune cell populations: tumor-associated macrophages, immature dendritic cells, and NKT cells. Strong positive correlations were found between all markers of early T-cell activation ($CD45RA^+CD27^+/CD25^+/CD26^+/CD69^+$; $P < 0.05$ for all combinations) and the markers of tumor-associated macrophages and immature dendritic cells ($P = 0.01$). These two functional groups of markers were positively correlated (tumor-associated macrophages/$CD25^+/CD26^+/CD69^+$ and immature dendritic cells/$CD25^+/CD26^+$; $P < 0.05$ for all combinations). Thus, cluster 3 may illustrate the role of the innate immune compartment for T-cell priming, activation, and memory differentiation required for efficient adaptive immune responses.

In contrast, cluster 3 was entirely disrupted in META+ Hi patients (Fig. 3*A, center*). Indeed, the META+ Hi clustered correlation matrix (Fig. 3*A, bottom*) displayed two inversely correlated groups of clusters (cluster 1 versus clusters 2 and 3). As observed in the META− Hi matrix, cluster 1 in META+ Hi matrix contained the markers of CD4 and early memory T-cell subpopulations markers ($CD4^+/CD28^+$; $P = 0.008$), as well as markers of total T cells, CD4 T cells, and B cells. In clusters 2 and 3, CD8 memory T-cell marker ($CD8^+CD45RO^+$) was significantly positively correlated with both effector memory ($CD45RA^-CD27^-$) and final effector ($CD127^-CD62L^-$) T-cell markers ($P = 0.04$ and $P = 0.001$, respectively). Markers of NK and NKT cells, dendritic cells (immature and activated), and tumor-associated macrophages were also grouped in clusters 2 and 3 (Fig. 3*A, bottom*).

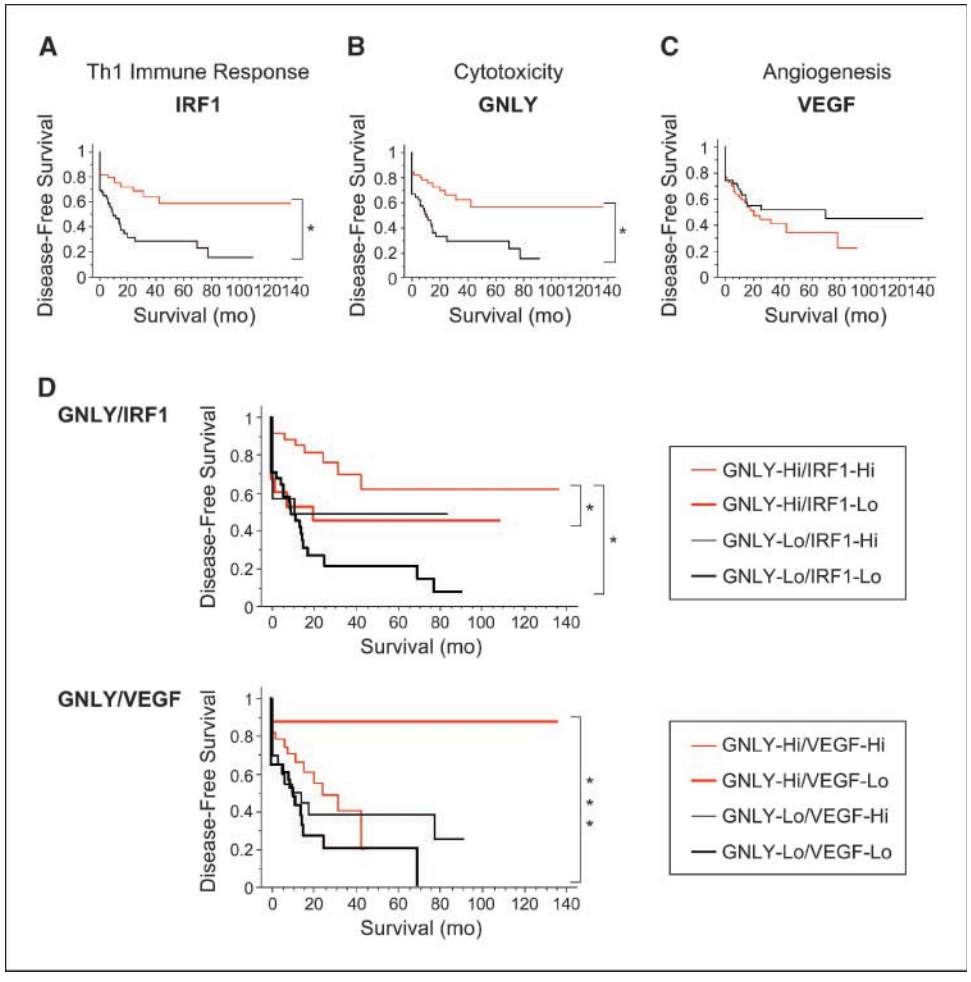


Figure 4. Disease-free survival of colorectal cancer patients according to expression of genes. *A* to *C,* disease-free survival of 103 patients according to high (*red lines*) or low (*black lines*) mRNA expression levels of IRF1 (*A*), GNLY (*B*), and *VEGF* (*C*) genes. *D,* disease-free survival of patients according to the expression levels of the *GNLY* gene in combination with *IRF1* (*top*) and *VEGF* (*bottom*) genes (*thin red lines,* GNLY-Hi/IRF1-Hi, GNLY-Hi/VEGF-Hi; *bold red lines,* GNLY-Hi/IRF1-Lo, GNLY-Hi/VEGF-Lo; *thin black lines,* GNLY-Lo/IRF1-Hi, GNLY-Lo/VEGF-Hi; *bold black lines,* GNLY-Lo/IRF1-Lo, GNLY-Lo/VEGF-Lo). The cutoff value for the expression of each gene was defined at the median of the cohort. *, $P < 0.05$, log-rank test.

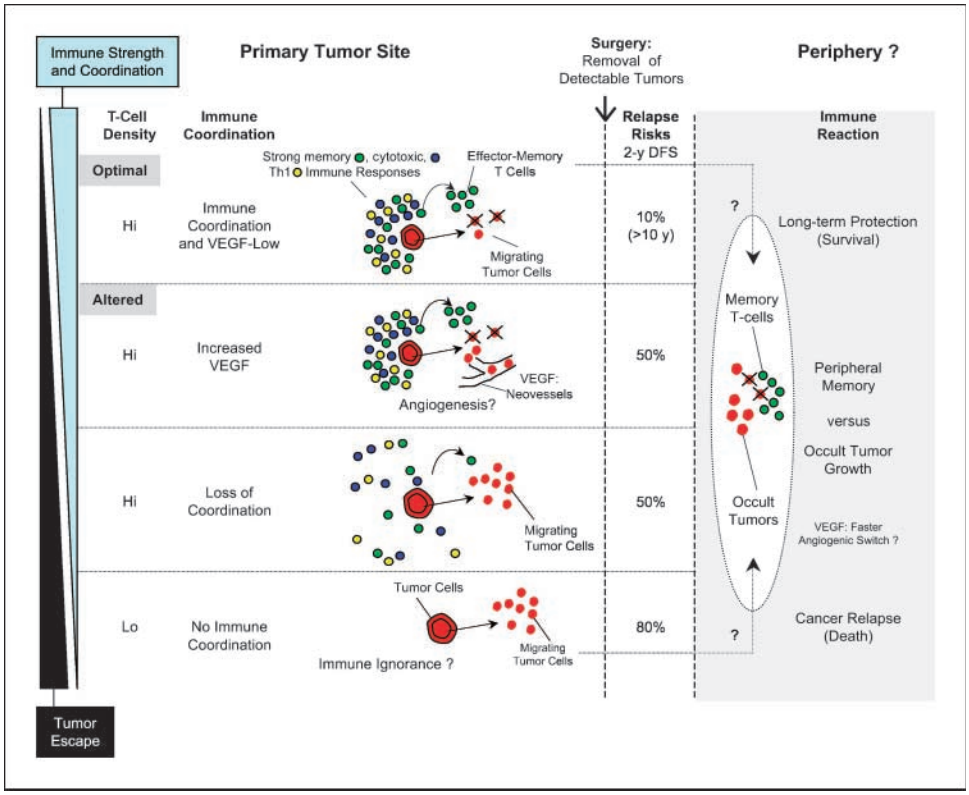


Figure 5. Proposed model: control of colorectal cancer outcome by the immune system. Before surgery, immune strength and coordination are in balance with mechanisms of tumor escape (tumor immunogenicity, inflammation, and angiogenesis) to control metastatic invasion from the primary tumor site. Four major immune coordination profiles within colorectal cancer primary tumors are found: (*a*) Strong and coordinated adaptive immune responses mediated by cytotoxic (*blue cells*; *GNLY*) and Th1 (*yellow cells*; *IRF1*) effector memory T cells (*green cells*) may contribute to the elimination of migrating tumor cells (*red cells*). (*b*) Angiogenic (*VEGF*) and inflammatory processes may facilitate metastatic invasion and (*c*) noncoordinate immune responses. (*d*) Weak (*Lo*) immune reactions (immune ignorance?). After surgical removal of clinically detectable tumors, the parameters defining this balance are significantly associated with the risks of cancer relapse [2-y disease-free survival (*DFS*)]. It could be postulated that the amount of invading occult tumors and the amount of circulating memory T cells, generated within distinct primary tumor microenvironment, are in balance to control cancer re-emergence in the periphery (after surgery).

Compared with Hi patients, META− Lo (Fig. 3*B, top*) and META+ Lo (Fig. 3*B, center* and *bottom*) patients had very distinct correlation profiles with a majority of noncorrelated markers (*yellow*). Furthermore, except for the only significant positive correlation between final effector and $CD8^+$ T-cell subpopulations ($CD8^+/CD127^-CD62L^-$; $P = 0.03$) in the META+ Lo matrix, all patterns of significant positive correlations observed in Hi patients were lost in Lo patients (Supplementary Fig. S2).

This analytic approach allowed us to visualize the absence of immune coordination in patients with low intratumoral T-cell densities, whereas patients with high intratumoral T-cell densities presented correlation patterns consistent with continual recruitment and proliferation of activated CD8 T cells associated with complete memory T-cell differentiation at the primary tumor site (CD8 T-cell/effector memory T-cell correlations). This profile of efficient immune reaction is in balance (negative correlation) with patterns that could illustrate altered immune responses (CD4/early memory T-cell/B-cell correlations). Because in Hi patients the presence of metastases was associated with (*a*) a significant decrease of CD8 and late memory T cells and innate cells and (*b*) a significant increase of B cells (Figs. 1 and 2), our data suggest altered immune reactions in META+ Hi patients.

Prognostic value of cellular immune coordination. We next assessed the effect of immune coordination on the proliferation/apoptosis status of primary tumor cells by Ki67/M30 immunohistochemical stainings of cognate tumor samples (Supplementary Fig. S4). No differences were observed among the patient groups, suggesting that the effect on cancer progression of the immune system may be inefficient for the destruction of the primary tumor. To validate the effect of the coordination of *in situ* immune response on colorectal cancer prognosis, we evaluated the density of intratumoral immune T cells in a large cohort of 435 patients. We investigated the CD8/CD3 T-cell density ratio in relation to clinical outcome in TMA experiments. Increased densities of T-cell infiltrates exhibiting high proportions of CD8 cytotoxic T cells within the primary tumor of colorectal cancer patients were associated with a significant protection against tumor recurrence (Supplementary Fig. S5).

To better characterize the mechanisms involved in antitumoral activity at the tumor-host interface, we investigated the effect on clinical outcome of mRNA expression levels of 17 mediators involved in immune or tumoral mechanisms. For each gene, patients were defined as high or low according to median gene expression. Disease-free survival rates were then calculated for each patient group. The prognostic value of the expression levels of genes related to T-cell populations (*CD3ζ, CD4, CD8α*), Th1 adaptive immune responses (*TBX21/T-BET, IRF1, IFNγ*), and cytotoxicity (*GNLY, GZMB*) were assessed. High expression of *TBX21/T-BET, IFNγ, IRF1*, and *GNLY* was associated with significantly improved disease-free survival rates ($P = 0.02$, $P = 0.02$, $P = 0.0003$, and $P = 0.0004$, respectively). Disease-free survival Kaplan-Meier curves according to *GNLY* and *IRF1* gene expression were illustrated (Fig. 4*A* and *B*, respectively). Conversely to immune mediators, the expressions of cancer-promoting genes involved in tumor invasion (*CEACAM1*), metastasis spreading (*EBAG9* and *CEA*), tumor cell antiapoptosis (*BIRC5/Survivin*), immune suppression (*IL-10* and *TGFβ*; data not shown), and angiogenesis (*VEGF*; Fig. 4*C*) had no prognostic values.

Immune coordination at the molecular level was assessed by analyzing combined expression of genes. We found significantly improved disease-free survival rates in patients with high combined gene expressions (Hi/Hi) of marker combinations related to CD4 T cells of Th1 phenotype (CD4/T-BET, CD4/IFNγ, CD4/IRF1 patients) and cytotoxic CD8 T cells (CD8/GNLY)

compared with patients expressing low levels of these genes (Lo/Lo; $P = 0.04$, $P = 0.002$, $P = 0.002$, and $P = 0.004$, respectively; Supplementary Fig. S6). High coexpression of *IRF1* and *GNLY* genes (GNLY-Hi/IRF1-Hi) was essential for beneficial outcome with median disease-free survival >140 months, whereas patients expressing low levels of one of these genes or both (GNLY-Hi/IRF1-Lo, GNLY-Lo/IRF1-Hi, GNLY-Lo/IRF1-Lo) had median disease-free survival <15 months (Fig. 4*D, top*). These observations confirmed the importance of a strong coordination between immune mediators of cytotoxic and Th1 adaptive immune responses for favorable colorectal cancer outcome.

Finally, we assessed the effects of the tumor microenvironment on *in situ* antitumoral immune responses. We analyzed the prognostic value of the expression of protumoral mediators in combination with *GNLY*. Among all tested genes, only *VEGF* showed a profound effect on patient survival when expressed with *GNLY*. Patients expressing high levels of *GNLY* and *VEGF* genes (GNLY-Hi/VEGF-Hi) and patients expressing low levels of *GNLY* (GNLY-Lo/VEGF-Hi and GNLY-Lo/VEGF-Lo) had similar disease-free survival rates that were significantly lower than the disease-free survival rates of GNLY-Hi/VEGF-Lo patients ($P < 0.004$ for all comparisons; Fig. 4*D, bottom*).

Discussion

The mechanisms controlling tumor progression and cancer relapse are not clearly characterized. Here, we investigated the quality of the immune reaction at the primary tumor site during cancer progression (i.e., according to the density of tumor-infiltrating T cells and the metastatic status of the patients). We showed that coordination of the immune response was drastically impaired in patients with low densities of intratumoral T cells compared with patients with high densities of such cells. Phenotypic correlation analyses showed matrices that were highly fragmented with no particular functional relevance in both META− Lo and META+ Lo patients. This suggested the absence of coordinated immune response independently of the metastatic status in Lo patients. Conversely, a high density of tumor-infiltrating T cells (Hi patients) was associated with strong immune coordination. Significant positive correlations between T cells of late memory and cytotoxic phenotypes indicated continual recruitment, activation, and memory differentiation of CD8 T cells at the primary tumor site. In larger cohorts of patients using tissue microarrays, we also showed that high densities of T cells associated with a high CD8/CD3 density ratio correlated with a very good prognosis. In contrast, low adaptive immune coordination was associated with very poor prognosis. Consistently, patients presenting high and coordinated intratumoral expression of the global Th1 immune response marker *IRF1* and the cytotoxicity-specific marker *GNLY* had significantly better survival rates compared with patients expressing heterogeneous or low levels of these genes.

Yet, in Hi patients, the presence of metastases was associated with (*a*) a significant decrease of innate immune cells, (*b*) a significant decrease of CD8 T cells and fully differentiated memory T cells, (*c*) loss of the phenocluster of markers of innate cells and early activated T cells (illustrating innate/adaptive immune compartment interactions), and (*d*) a significant increase of B cells (suggesting immune deviation mechanisms; refs. 34, 35).

According to the coexpression of *IRF1* and *GNLY*, the frequencies of strong immune coordination parameters were reduced in META+ patients (data not shown). META+ patients represented only 48% of GNLY-Hi/IRF1-Hi patients and 65% of GNLY-Lo/IRF1-Lo patients. Overall, these observations represent clues of altered immune responses when metastases are present. However, a significant number of META+ patients displayed a high degree of immune coordination preventing relapse events. This raises two hypotheses: Does the alteration of the immune reaction at the primary tumor site facilitate metastatic invasion? Is the immune system overwhelmed and affected by the presence of metastases? Interestingly, some patients without lymph node and/or distant organs (META− patients) have an absence of immune coordination and low densities of T cells. Thus, local immune escape mechanisms may exist in the primary tumor even before metastatic spread. Because proliferation and apoptosis rates of tumor cells were not significantly different between the patient groups, the outgrowth of the primary tumor may overcome the destruction by the immune system. Whichever hypothesis on the mechanisms of long-term relapse prevention after surgery may be related to the quality of the immune reaction at the primary tumor site even in patients with advanced colorectal cancer.

Among the distinct factors potentially involved in immune escape at the primary site, several mechanisms or cell types may participate, such as immature dendritic cells, regulatory T cells, Th1/Th2 immune response switch, immunosuppression, local metastatic invasion, inflammation, and angiogenesis. In patients with metastases or low intratumoral T-cell densities, there was no increased expression of markers of regulatory T cells, tumor-associated macrophages, and immature dendritic cells by flow cytometry ($CD4^+CD25^{hi}$, $CD45^+CD14^+$, and $CD45^+CD1a^+CD14^-CD83^-$, respectively) and by tissue microarray ($Foxp3^+$, $CD68^+$, and $CD1a^+$, respectively) experiments (data shown). This may suggest the existence of immune ignorance or reduced tumor immunogenicity mechanisms for differential immune cell recruitment among patients. Interestingly, we found that only the proangiogenic factor VEGF had a deleterious effect on relapse prevention mechanisms associated with strong antitumoral immune reaction. *VEGF* expression levels had no prognostic value per se, in agreement with immunohistochemical-based studies (36, 37). Our data indicate that cytotoxic Th1 adaptive immune responses may be necessary, but not sufficient, to prevent tumor recurrence. At the primary tumor site, inflammatory cytokines (such as IL-1A, IL-6, IL-8, oncostatin M, and tumor necrosis factor α) can enhance tumorigenic processes by up-regulating important mediators of angiogenesis, such as VEGF (38). In this context, the effect on colorectal cancer outcome of the balance between *GNLY/IRF1* and *VEGF* expressions may reflect beneficial cytotoxic Th1 adaptive immune responses versus deleterious inflammatory reaction (39–41). However, other roles of angiogenesis may affect cytotoxic Th1 adaptive immune responses. In the primary tumor site, the role of angiogenesis in promoting nutrient supply (21) may not explain the obliteration of the beneficial role of strong immune responses. In contrast, the induction of vascular exit paths for migrating tumor cells (42) could result in increased metastatic dissemination favoring relapse occurrence. In this case, even strong *in situ* cytotoxic Th1 adaptive immune responses may not be sufficient to counteract metastatic invasion. Thus, in the periphery, great number of occult tumor cells may overwhelm immunosurveillance mechanisms during the equilibrium phase. Furthermore, if the migrating tumor cells inherit the strong angiogenic properties of their resident counterparts, occult tumor outgrowth may be further enhanced (43). This idea that angiogenesis

and adaptive immune responses are strongly linked in cancer recurrence should be taken into account when considering therapeutic options.

We were able to describe four major immune coordination profiles within colorectal cancer primary tumors depending on the balance between tumor escape and immune coordination: (*a*) strong and coordinate cytotoxic Th1 immune responses (*GNLY/IRF1*) without or (*b*) with tumor angiogenesis (*VEGF*), (*c*) noncoordinate immune responses, and (*d*) weak (Lo) immune reactions (immune ignorance?). These distinct immune profiles are associated with significant distinct cancer outcome (relapse risks), as summarized in Fig. 5.

It is suspected that metastatic invasion can lead to the dissemination of tumor cells that can remain in an asymptomatic and nondetectable state of dormancy (i.e., not expanding in mass) for long periods of time before cancer re-emergence (44). Control of cancer dormancy involves various mechanisms such as cellular dormancy (G_0-G_1 arrest), angiogenic dormancy, and immunosurveillance (45). Recently, Koebel and colleagues (46) showed that stable lesions of transformed immunogenic cells in mice were controlled by the adaptive immune system of the host in a condition of "equilibrium." In these experiments, loss of either immunocompetence or immunogenicity could lead to tumor outgrowth. We previously showed that the absence of microscopic evidence of early metastatic invasiveness within lymphovascular vessels was associated with high densities of effector memory T cells within primary tumors and that both criteria were powerful indicators of improved prognosis in human colorectal cancer (17). Based on these data, it could be proposed that the immune system exerts its protective role against cancer relapse (*a*) at the primary tumor site by eliminating migrating tumor cells, subsequently reducing the number of disseminated occult tumors, and (*b*) in the periphery by controlling occult tumor evolution from dormancy state to cancer re-emergence (equilibrium phase). Moreover, these two antitumoral functions of the immune system could be tightly associated. As suggested in mice (47), cytotoxic effector memory T cells reacting at the primary tumor site might also, after surgical removal of tumors, be in charge of long-term antitumoral immunity in colorectal cancer.

In conclusion, our study argues for the involvement of immune coordination and late memory and cytotoxic T-cell populations in antitumoral activity against human colorectal cancer (Fig. 5). First, due to their enhanced cytotoxic capabilities, effector memory T cells may be involved in the control of metastatic invasion at the primary tumor site. Second, due to their memory properties, effector memory T cells may provide long-term protection against outgrowth of disseminated occult tumor cells potentially involved in relapse events. Depending on the strength and coordination of antitumoral immune responses elicited in primary tumor microenvironments (level of immunogenicity and angiogenesis), populations of T cells with distinct quantity (number of clones) and quality (memory differentiation state) could be generated. Subsequently, distinct potentials for long-lived antitumoral immunity may be maintained after surgical resection of primary and secondary tumors. Future comparative studies of tumors according to immune parameters and angiogenesis may reveal biological mechanisms involved in emergence and cancer progression. More adapted treatment and therapeutic strategies may ultimately be proposed to cure colorectal cancer.

Disclosure of Potential Conflicts of Interest

No potential conflicts of interest were disclosed.

Acknowledgments

Received 7/10/2008; revised 12/11/2008; accepted 1/2/2009; published OnlineFirst 3/3/09.

Grant support: Association pour la Recherche sur le Cancer (ARC), the National Cancer Institute (INCa), the Canceropole Ile de France, Ville de Paris, Immucan, INSERM, the Austrian Federal Ministry of Science and Research (GEN-AU project Bioinformatics Integration Network) and Academic Cooperation and Mobility Unit (fellowship to P Charoentong), and the European Commission (7FP, Geninca Consortium, grant no. 202230).

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked *advertisement* in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

References

1. Finn OJ. Cancer immunology. N Engl J Med 2008;358:2704–15.
2. Dighe AS, Richards E, Old LJ, Schreiber RD. Enhanced *in vivo* growth and resistance to rejection of tumor cells expressing dominant negative IFNγ receptors. Immunity 1994;1:447–56.
3. Kaplan DH, Shankaran V, Dighe AS, et al. Demonstration of an interferon γ-dependent tumor surveillance system in immunocompetent mice. Proc Natl Acad Sci U S A 1998;95:7556–61.
4. Smyth MJ, Thia KY, Street SE, MacGregor D, Godfrey DI, Trapani JA. Perforin-mediated cytotoxicity is critical for surveillance of spontaneous lymphoma. J Exp Med 2000;192:755–60.
5. van den Broek ME, Kagi D, Ossendorp F, et al. Decreased tumor surveillance in perforin-deficient mice. J Exp Med 1996;184:1781–90.
6. Shankaran V, Ikeda H, Bruce AT, et al. IFNγ and lymphocytes prevent primary tumour development and shape tumour immunogenicity. Nature 2001;410:1107–11.
7. Zhou G, Lu Z, McCadden JD, Levitsky HI, Marson AL. Reciprocal changes in tumor antigenicity and antigen-specific T cell function during tumor progression. J Exp Med 2004;200:1581–92.
8. Dunn GP, Old LJ, Schreiber RD. The three Es of cancer immunoediting. Annu Rev Immunol 2004;22:329–60.
9. Clemente CG, Mihm MC, Jr., Bufalino R, Zurrida S, Collini P, Cascinelli N. Prognostic value of tumor infiltrating lymphocytes in the vertical growth phase of primary cutaneous melanoma. Cancer 1996;77:1303–10.
10. Sato E, Olson SH, Ahn J, et al. Intraepithelial $CD8^+$ tumor-infiltrating lymphocytes and a high $CD8^+$/regulatory T cell ratio are associated with favorable prognosis in ovarian cancer. Proc Natl Acad Sci U S A 2005;102:18538–43.
11. Zhang L, Conejo-Garcia JR, Katsaros D, et al. Intratumoral T cells, recurrence, and survival in epithelial ovarian cancer. N Engl J Med 2003;348:203–13.
12. Baier PK, Wimmenauer S, Hirsch T, et al. Analysis of the T cell receptor variability of tumor-infiltrating lymphocytes in colorectal carcinomas. Tumour Biol 1998;19:205–12.
13. Dalerba P, Maccalli C, Casati C, Castelli C, Parmiani G. Immunology and immunotherapy of colorectal cancer. Crit Rev Oncol Hematol 2003;46:33–57.
14. Diederichsen AC, Hjelmborg JB, Christensen PB, Zeuthen J, Fenger C. Prognostic value of the $CD4^+$/$CD8^+$ ratio of tumour infiltrating lymphocytes in colorectal cancer and HLA-DR expression on tumour cells. Cancer Immunol Immunother 2003;52:423–8.
15. Naito Y, Saito K, Shiiba K, et al. $CD8^+$ T cells infiltrated within cancer cell nests as a prognostic factor in human colorectal cancer. Cancer Res 1998;58:3491–4.
16. Prall F, Duhrkop T, Weirich V, et al. Prognostic role of $CD8^+$ tumor-infiltrating lymphocytes in stage III colorectal cancer with and without microsatellite instability. Hum Pathol 2004;35:808–16.
17. Pages F, Berger A, Camus M, et al. Effector memory T cells, early metastasis, and survival in colorectal cancer. N Engl J Med 2005;353:2654–66.
18. Galon J, Costes A, Sanchez-Cabo F, et al. Type, density, and location of immune cells within human colorectal tumors predict clinical outcome. Science 2006;313:1960–4.
19. Galon J, Fridman WH, Pages F. The adaptive immunologic microenvironment in colorectal cancer: a novel perspective. Cancer Res 2007;67:1883–6.
20. Hanahan D, Weinberg RA. The hallmarks of cancer. Cell 2000;100:57–70.
21. Bergers G, Benjamin LE. Tumorigenesis and the angiogenic switch. Nat Rev Cancer 2003;3:401–10.
22. Smyth MJ, Godfrey DI, Trapani JA. A fresh look at tumor immunosurveillance and immunotherapy. Nat Immunol 2001;2:293–9.
23. Dunn GP, Koebel CM, Schreiber RD. Interferons, immunity and cancer immunoediting. Nat Rev Immunol 2006;6:836–48.
24. Kim R, Emi M, Tanabe K, Arihiro K. Tumor-driven evolution of immunosuppressive networks during malignant progression. Cancer Res 2006;66:5527–36.
25. Zou W. Immunosuppressive networks in the tumour environment and their therapeutic relevance. Nat Rev Cancer 2005;5:263–74.

26. Khong HT, Restifo NP. Natural selection of tumor variants in the generation of "tumor escape" phenotypes. Nat Immunol 2002;3:999–1005.

27. Mocellin S, Wang E, Marincola FM. Cytokines and immune response in the tumor microenvironment. J Immunother 2001;24:392–407.

28. Egen JG, Kuhns MS, Allison JP. CTLA-4: new insights into its biological function and use in tumor immunotherapy. Nat Immunol 2002;3:611–8.

29. Okazaki T, Honjo T. The PD-1-PD-L pathway in immunological tolerance. Trends Immunol 2006;27:195–201.

30. Sobin LH, Greene FL. Global TNM advisory group. Cancer 2004;100:1106.

31. Boulton SJ, Gartner A, Reboul J, et al. Combined functional genomic maps of the *C. elegans* DNA damage response. Science 2002;295:127–31.

32. Galon J, Franchimont D, Hiroi N, et al. Gene profiling reveals unknown enhancing and suppressive actions of glucocorticoids on immune cells. FASEB J 2002;16:61–71.

33. Sturn A, Quackenbush J, Trajanoski Z. Genesis: cluster analysis of microarray data. Bioinformatics 2002;18:207–8.

34. Shah S, Divekar AA, Hilchey SP, et al. Increased rejection of primary tumors in mice lacking B cells: inhibition of anti-tumor CTL and TH1 cytokine responses by B cells. Int J Cancer 2005;117:574–86.

35. Tan TT, Coussens LM. Humoral immunity, inflammation and cancer. Curr Opin Immunol 2007;19:209–16.

36. Doger FK, Meteoglu I, Tuncyurek P, Okyay P, Cevikel H. Does the EGFR and VEGF expression predict the prognosis in colon cancer? Eur Surg Res 2006;38:540–4.

37. Zheng S, Han MY, Xiao ZX, Peng JP, Dong Q. Clinical significance of vascular endothelial growth factor expression and neovascularization in colorectal carcinoma. World J Gastroenterol 2003;9:1227–30.

38. Angelo LS, Kurzrock R. Vascular endothelial growth factor and its relationship to inflammatory mediators. Clin Cancer Res 2007;13:2825–30.

39. Balkwill F, Coussens LM. Cancer: an inflammatory link. Nature 2004;431:405–6.

40. Moore RJ, Owens DM, Stamp G, et al. Mice deficient in tumor necrosis factor-α are resistant to skin carcinogenesis. Nat Med 1999;5:828–31.

41. Voronov E, Shouval DS, Krelin Y, et al. IL-1 is required for tumor invasiveness and angiogenesis. Proc Natl Acad Sci U S A 2003;100:2645–50.

42. Folkman J. Role of angiogenesis in tumor growth and metastasis. Semin Oncol 2002;29:15–8.

43. Naumov GN, Akslen LA, Folkman J. Role of angiogenesis in human tumor dormancy: animal models of the angiogenic switch. Cell Cycle 2006;5:1779–87.

44. Vessella RL, Pantel K, Mohla S. Tumor cell dormancy: an NCI workshop report. Cancer Biol Ther 2007;6:1496–504.

45. Aguirre-Ghiso JA. Models, mechanisms and clinical evidence for cancer dormancy. Nat Rev Cancer 2007;7:834–46.

46. Koebel CM, Vermi W, Swann JB, et al. Adaptive immunity maintains occult cancer in an equilibrium state. Nature 2007;450:903–7.

47. Xiang R, Lode HN, Gillies SD, Reisfeld RA. T cell memory against colon carcinoma is long-lived in the absence of antigen. J Immunol 1999;163:3676–83.