Regulatory network motifs and hotspots of cancer genes in a mammalian cellular signalling network

A. Awan, H. Bari, F. Yan, S. Moksong, S. Yang, S. Chowdhury, Q. Cui, Z. Yu, E.O. Purisima and E. Wang

Abstract: Mutations or overexpression of signalling genes can result in cancer development and metastasis. In this study, we manually assembled a human cellular signalling network and developed a robust bioinformatics strategy for extracting cancer-associated single nucleotide polymorphisms (SNPs) using expressed sequence tags (ESTs). We then investigated the relationships of cancer-associated genes [cancer-associated SNP genes, known as cancer genes (CG) and cell mobility genes (CMGs)] in a signalling network context. Through a graph-theory-based analysis, we found that CGs are significantly enriched in network hub proteins and cancer-associated genes are significantly enriched or depleted in some particular network motif types. Furthermore, we identified a substantial number of hotspots, the three- and four-node network motifs in which all nodes are either CGs or CMGs. More importantly, we uncovered that CGs are enriched in the convergent target nodes of most network motifs, although CMGs are enriched in the source nodes of most motifs. These results have implications for the foundations of the regulatory mechanisms of cancer development and metastasis.

1 Introduction

Cancer cells are characterised by uncontrolled cell growth, invasion of surrounding tissues and finally metastasis to distant regions of the human body. Accumulation of genetic mutations in part triggers tumour development and progression. Gene mutation or deregulation also promotes cell mobility that is highly correlated with tissue invasion and distant metastasis. A set of gene mutations or overexpressions are closely linked to patient clinical outcomes, suggesting that these genes could be cancer biomarkers for diagnostics.

Cells use sophisticated communication between proteins to perform a series of tasks such as growth, maintenance of cell survival, proliferation and development. Signalling pathways, which are used to transmit biological signals, perform the communication between proteins. Signalling pathways are crucial in maintaining cellular homeostasis and determine cell behaviour. Thus, alterations of expression of the genes in cellular signalling pathways could lead to tumour development or promote cell migration. Indeed, alterations to genes that encode signalling proteins are commonly observed in many types of cancers [1–3]. Therefore recent systematic screenings of mutations have focused on gene families involved in signalling pathways, such as kinases and phosphatases in breast and other cancers [4, 5]. These efforts have identified mutations in a variety of genes, including PIK3CA, one of the most commonly mutated oncogenes in human cancers [6–9]. Systematic identification of gene mutations that are involved in signalling pathways and associated with cancer progression and cell mobility has been proven to be useful in finding cancer biomarkers and therapeutic targets [1, 10–12]. With the development of automatic DNA sequencing technology, large-scale genome sequencing projects have generated a vast amount of DNA sequence information. Expressed sequence tag (EST) collections represent partial descriptions of transcribed portions of genomes. So far, more than two million high-quality ESTs from human cancer tissues have been posted in the cancer genome anatomy project (CGAP, http://cgap.nci.nih.gov/) at National Cancer Institute. Bioinformatics analysis of ESTs from normal and cancerous tissues could identify genetic variations associated to cancer. Single nucleotide polymorphisms (SNPs) are the most common genetic variations in the human genome. More and more experimental evidence shows that some SNPs are closely linked to cancer and treated as genotypic markers [13]. Therefore developing a robust bioinformatics method to identify cancer-associated SNPs and studying them in a cellular context such as cellular signalling would help not only in pinpointing cancer biomarkers but also in providing new insights into molecular mechanisms of carcinogenic and metastatic processes.

To elucidate the underlying molecular mechanisms of how signalling gene mutations or overexpression act on tumour development and metastasis, it is necessary to dissect signalling events that are related to the cancer-associated genes. Traditionally scientists treat cellular signalling events in view of biological pathways, study one pathway at a time and then try to gather information from a few pathways together to understand what is going on inside cells. However the proteins, which make up one individual pathway, rarely operate in isolation but ‘cross-talk’ with another pathway’s proteins to process signal

doi:10.1049/iet-syb:20060068

Paper first received 25th September 2006 and in revised form 26th January 2007

A. Awan, H. Bari, S. Moksong, S. Chowdhury, Q. Cui, Z. Yu, E.O. Purisima and E. Wang are with the Biotechnology Research Institute, National Research Council Canada, Montreal, Quebec, Canada H4P 2R2

F. Yan is with Institute of Hematology, Chinese Academy of Medical Sciences, Tianjin 300020, China

S. Yang is with School of Chemical Engineering, Tianjin University, Tianjin 300072, China

E. Wang is with Center for Bioinformatics, McGill University, Montreal, Quebec, Canada H3A 2B4

A. Awan, H. Bari and F. Yan contributed equally to this work.

E-mail: edwin.wang@cnrc-nrc.gc.ca

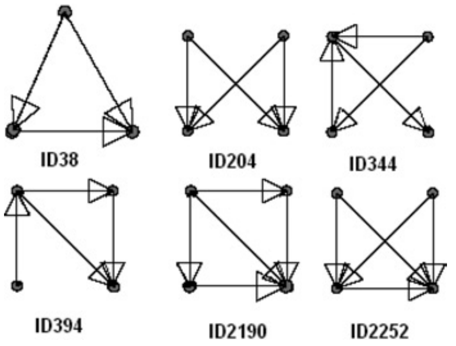

Fig. 1 *Signalling network motifs for cancer-associated genes*

information. A network-level view of signalling events emerges as an important concept. In this study, we first developed a robust bioinformatics strategy to find cancer-associated SNPs by extracting human ESTs of normal and cancer tissues. At the same time, we manually assembled a human cellular signalling network. We then mapped the integrated cancer-associated genes, which include the SNP genes we identified, known as cancer genes (CGs) and cancer cell mobility genes (CMGs), onto the signalling network to study their relationships in a signalling network context.

2 Materials and methods

2.1 Datasets used in this study

Human ESTs of normal (1.89 million) and cancer (2.24 million) tissues were downloaded from NCBI dbEST (http://www.ncbi.nlm.nih.gov/dbEST) and CGAP, respectively. As of May 2005, CGAP had 1870 and 3298 normal and cancerous EST libraries, respectively (supplementary Table 1, supplementary materials are at http://www.bri.nrc.ca/wang/snp1.html). Protein and mRNA sequences of human genome were downloaded from ftp://ftp.ncbi.nlm.nih.gov/genomes/H_sapiens/protein/ and ftp://ftp.ncbi.nlm.nih.gov/genomes/H_sapiens/RNA/, respectively. We took tumour CMGs from a high-throughput, small RNA-interfering screening of a few cancer cell lines including ovarian carcinoma cell line, SKOV-3 and breast cancer cell line, MDA-231 [14]. The screening identified 532 potential tumour CMGs and a few of these genes were further validated using other experimental analyses such as RT-PCR, additional RNA-interfering and cell invasion assays. We collected known CGs from NCBI Online Mendelian Inheritance in Man database (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=OMIM).

2.2 Signalling network construction and network motif detection

To construct the human cellular signalling network, we manually curated signalling pathways from literature. The signalling data source for our pathways is the BioCarta database (http://www.biocarta.com/genes/allpathways.asp), which, so far, is the most comprehensive database for human cellular signalling pathways. Our curated pathway database recorded gene names and functions, cellular locations of each gene and relationships between genes such as activation, inhibition, translocation, enzyme digestion, gene transcription and translation, signal stimulation and so on. To ensure the accuracy and the consistency of the database, each referenced pathway was cross-checked by different researchers and finally all the documented pathways were checked by one researcher. In total, 164 signalling pathways were documented (supplementary Table 2). Furthermore, we merged the curated data with another literature-mined human cellular signalling network [15]. As a result, the merged network contains nearly 1100 proteins (SupplementaryNetworkFile). To construct a signalling network, we considered relationships of proteins as links (activation or inactivation as directed links and physical interactions in protein complexes as neutral links) and proteins as nodes. To detect and extract network motifs, we used mfinder [16]. To obtain statistically significant inference of distributions of the cancer-associated genes in network motifs, re-sampling statistical procedures were used. Briefly, we randomly assigned the same number of the cancer-associated genes as they are in the real network, recalculated the distributions of the cancer-associated genes and compared them to the real distributions of the cancer-associated genes of the network. We repeated the simulation 5000 times and then calculated *P* values. A detailed description of the network re-sampling procedures was described previously [17].

2.3 SNP data mining strategy

To assign ESTs to human genes, we used ESTs to perform non-gap blast on human mRNA and protein sequences using BLASTN and BLASTX programs [18]. E-score cutoff was 1×10^{-20}. In each blast, the matched ESTs to genes and proteins were obtained. If an EST has the best match to a certain gene and also to the gene's coding protein, we assigned the EST to the gene. Otherwise we discarded the EST. We picked up the ESTs that were aligned and assigned to the genes in the network. We observed that some sequencing errors occurred within 100–150 bps of the end-sequence region of the ESTs; thus, we removed 200 bps from the end-sequence regions of ESTs. After cutting off 200 bps from the end-sequence region of an

Table 1: Enrichments of cancer-associated genes in network motifs[a]

Motif ID	38	204	344	394	2190	2252
CG	23.6%	11.3%	36.7%	26.3%	33.6%	26.8%
	(153/647)	(170/1505)[b]	(1092/2977)	(735/2795)[b]	(44/131)	(66/246)
	0.57	2×10^{-4}	2×10^{-4}	2×10^{-4}	0.25	0.09
CMG	27.3%	46.9%	33.2%	35.7%	35.1%	34.9%
	(177/647)	(707/1505)	(989/2977)	(997/2795)	(46/131)	(86/246)
	8.5×10^{-4}	2×10^{-4}	2×10^{-4}	2×10^{-4}	0.05	0.01

[a] For each gene type, the rates of motifs having cancer-associated genes are presented in the first row whereas the corresponding *P* values are in the second row

[b] Indicates depletion rather than enrichment

Table 2: Distribution of cancer-associated genes on node positions of network motifs[a]

Motif ID		38	204	344	394	2190	2252
CG	P1	33.3	**36.2**	**35.3**	28.6	25.0	29.7
	P2	29.8	22.1	24.4	21.2	**30.0**	15.8
	P3	**36.8**	24.2	20.3	23.4	15.0	7.9
	P4	–	17.4	20.0	26.8	**30.0**	**46.5**
CMG	P1	37.1	31.9	**33.7**	25.7	17.5	24.8
	P2	**31.0**	**44.9**	22.5	24.4	**31.6**	**30.5**
	P3	**31.9**	12.5	19.1	23.1	22.8	**31.4**
	P4	–	10.7	24.7	26.7	28.1	13.3

[a] P1, P2, P3 and P4 represent node position of motifs. CG and CMG represent cancer genes and cell mobility genes, respectively. The numbers represent the frequencies of CG or CMG on each node position

EST, we scanned the EST and its alignments to find genetic variants. We assumed that mutations are not often clustered in a short region, so we set a 25 bp window to avoid sequencing errors. We defined a single mutation such that it is the only mutation and at the middle position of a 25 bp length window. We counted single mutations, which occurred in at least 30 libraries. To associate SNPs with cancer, we used Fisher's exact test for the significance of occurrence of an SNP in cancerous and normal tissues. To control false positives of multiple tests, false discovery rate was used. We used the standalone pMut program [19] to test whether the identified SNPs affect the protein's function and are relevant to diseases. To further support the prediction, we carried out molecular modelling of the proteins to visualise the locations of the mutations in the three-dimensional structures of the proteins (see supplementary modelling). Crystal structures of the proteins were used when available; otherwise, homology models were built. For example, histone deacetylase 2 (HDAC2) has no crystal structure available; a homology model was built using the available crystal structure of HDAC8 (pdb code 1w22) as a template for the analysis (see supplementary modelling). The structures were examined to see if the mutations were expected to affect the biochemical function of the protein. We should note that molecular modelling is a prediction approach, which has limitations in generating false positives.

3 Results

3.1 Mining of cancer-associated SNPs using ESTs

The availability of a large number of cancer and normal tissue ESTs provides an opportunity for screening genetic variations and identifying genes associated with cancer through bioinformatics analysis. To detect SNPs, we collected 2.24 million cancer tissue ESTs and 1.9 million normal tissue ESTs. We assigned ESTs to human genes by BLASTX and BLASTN. Because we focused on cellular signalling genes, we only took the ESTs, which had been assigned to the genes in the signalling network. We assigned 629 signalling genes to 48 993 cancer ESTs and 723 signalling genes to 33 285 normal tissue ESTs. Both EST pools represent almost 40 human tissues and cancerous ESTs, which represent most of the cancer cell types (supplementary Table 3). Direct link of genes to cancer could test the association between potential functional variants and cancer phenotypes. This involves the examination of non-synonymous SNPs (nsSNPs) that result in an amino acid change. Most of the functional variants of the genes related to diseases occur within coding regions. We identified 44 nsSNPs in the coding regions of 26 genes that are associated with cancer by applying statistical analysis of SNPs in cancer and normal tissues ($P < 0.05$). The assumption is that cancer-associated SNPs are over-represented in cancerous libraries over normal tissue libraries. To further characterise putative functional variants of the identified SNPs, we evaluated the impact of SNPs on protein structure and function using both automatic and manual procedures. To automatically evaluate a SNP's effect on protein function, we used pMut program which was developed to associate human diseases with genetic variation by scanning single-point amino acidic mutations. The program allows fast pinpointing of disease-associated mutations with an accuracy of nearly 80%. Among the 44 SNPs, we identified 21 SNPs on 14 genes that affect protein function and link to cancer (supplementary Table 4). To further confirm pMut predictions, we manually examined the SNPs by structural study of available crystal structures and generating homology models of the proteins. For example, SNPs in HDAC2 and NFκB might cause structural changes affecting biochemical function or protein stability (supplementary modelling).

Among the identified 14 genes which have cancer-associated SNPs, four of them have been found to bear cancer-related mutations: the transmembrane protein tyrosine kinase ERBB2, HDAC2, histone acetyltransferase (HAT) P300/CBP, the NFκB/Rel family of transcript factor RelA and the α subunit of the stimulatory G protein ($G_{\alpha}S$) are related with different types of cancers. HDACs and HATs are enzymes that catalyse the deacetylation and acetylation of lysine residues located in the N-terminal tails of histones and non-histone proteins. Emerging evidence demonstrates that perturbation of this balance is often observed in human cancers, and inhibition of HDACs is considered to be among the most promising novel therapeutic strategies against cancer. The role of P300 as a tumour suppressor was first demonstrated as it was identified as an adenoviral E1A-binding protein. In breast and colon cancers, P300 expression is extremely low [20, 21]. The discovery of SNPs of these proteins in this study indicates that extracting from EST datasets is a powerful tool for finding gene mutations in cancer cells.

3.2 Distribution of cancer-associated genes in the network

To obtain insights into the molecular mechanisms of how gene mutations or deregulations act on tumour development in a cellular signalling network context, we studied the relationships of cancer-associated genes in a signalling network. To do so, we first manually curated human cellular signalling information from literature and then merged the data with another literature-mined human signalling network. Most of these pathways represent central signalling events in cells. Therefore the network could be seen as a general signal information centre in cells. The network is presented as a graph with directed and neutral links, in which, nodes represent proteins, directed links represent activating and inhibitory relations and neutral links represent only physical interactions between proteins. To study the relationships of cancer-associated genes on the cellular signalling network, we first combined the known CGs and the cancer SNP genes we identified into a set called CGs. We defined the CGs and the 532 genome-wide RNAi screened cancer CMGs as cancer-associated genes and then mapped these genes onto the network. Ninety-five CGs and 87 CMGs were mapped onto the network. We first asked if the CGs and the CMGs are network hub proteins which have many more links than other proteins in the network. We ranked network proteins based on their link numbers and then defined the hub proteins as the top 15% of highly linked proteins. We found that 22% ($P = 0.02$) and 17% ($P = 0.23$) of hub proteins are CGs and CMGs, respectively. These results suggest that CGs but not CMGs are enriched in hub proteins. Hub proteins are the functionally important nodes shared by many signalling pathways. Therefore mutations or deregulations of these hub genes may lead to cancer. To discover the distribution of cancer-associated genes in the network, we divided the network proteins into three groups based on the cellular location of the proteins and signal information flow: ligand-receptor, intracellular components and nuclear proteins. We calculated the fractions of the CGs and the CMGs in each region. We found that downstream network regions are significantly enriched with CGs ($P < 2 \times 10^{-4}$): 7.9%, 9.2% and 18.1% in network ligand-receptor, intracellular components and nucleus, respectively, in contrast to 8.6%, the average rate of the CGs of the network proteins. This fact suggests that CGs are more enriched in network downstream proteins. On the other hand, CMGs have no significant enrichment in any region.

3.3 Regulatory network motifs of cancer-associated genes

One way to study a complex system is to break down the system into sub-systems that are independently functional units. Biological networks can be decomposed into statistically over-represented subgraphs, which appear recurrently in networks and are called network motifs [22]. A network motif is a group of interacting components capable of signal processing and also known as regulatory loops in biology. Network motifs have been shown to have distinct regulatory functions and are robust to resistant internal noise. Integration of commonly accessible data types such as protein interaction, gene expression profiles and gene orthologues onto networks has revealed insights into network motif usages in different cellular conditions [23–25]. We have integrated a dataset of genome-wide mRNA decay rates onto gene regulatory network motifs and revealed the design principles of gene regulatory network motifs [17]. Furthermore, the integrative analysis of interactions between microRNAs and a human cellular signalling network revealed the microRNA regulation principles of the signalling network [26]. Therefore integration of cancer-associated genes onto signalling network motifs would help to understand the regulatory mechanisms of how cancer-associated genes work on cancer development and metastasis. To this end, we first identified all the three- and four-node motifs in the network. We are interested in cellular regulation of cancer-associated genes. Therefore we only picked up the motifs in which all the links are directed. Using this criterion, we found three- and four-node statistically significant motifs with the following motif IDs (mIDs): 38, 204, 344, 394, 2190 and 2252 (Fig. 1). We identified all the members of each motif type and mapped cancer-associated genes to them. We defined a motif rate as the number of motifs having the CGs or the CMGs of the motif type divided by the total number of the motifs of that type. We found that CMGs and CGs are significantly enriched in some particular motif types (Table 1), suggesting that perturbation of motif genes has more chance to lead to cancer and metastasis. Notably, CGs are not significantly enriched in mIDs 204 and 394 motifs, suggesting that these motifs may buffer gene mutations that prevent cancer development. These results also hint that carefully studying the relationships of cancer-associated genes on network motifs will lead to uncover the regulatory mechanisms of cancer-associated genes. Therefore we further examined the distribution of cancer-associated genes on node positions for each motif type (Table 2). CMGs are enriched in source nodes in most of the motif types, whereas CGs are enriched in the convergent nodes which are the target nodes receiving signals from two or more source nodes in most of motif types except the two less CG enriched motif types (Table 2). These results indicate different regulatory mechanisms between cancer development and metastasis. Therefore we inquired whether the CGs and the CMGs share some regulatory network motifs. If a motif contains both CGs and CMGs, we counted this motif as shared motifs. We found that only a few shared motifs, indicating that CGs and CMGs avoid sharing motifs. This result is consistent with our observation that CGs and CMGs use distinct motifs and regulatory mechanisms. We further speculated about whether some cancer-associated genes are clustered in the network and become hotspots. If all the nodes of a motif are the CGs or the CMGs, we called this motif as a CG or CMG hotspot, which indicates the vital role of this motif in cancer development or metastasis. We identified 11 three-node and 9 four-node motifs for CGs and 2 three-node motifs and 10 four-node motifs for CMGs. Statistical analyses showed that all these hotspots are not expected by chance ($P < 2 \times 10^{-4}$). These results suggest that some network regions or regulatory network motifs are critical to induce cancer or metastasis and these genes may work together to govern cell behaviours. These hotspots are potentially biomarker clusters or drug target clusters for curing cancer.

4 Discussion

Cells use signalling networks to communicate between and within cells to control many cellular processes. Biochemical signalling events, such as phosphorylation, acetylation, ubiquitylation, proteolytic cleavage and so on, are known to have mechanisms of activating or inactivating signalling proteins. The relationships among signalling proteins are thought to determine cell behaviour; therefore mutations or overexpression of signalling genes will affect signalling relationships of proteins [1, 3]. Mapping the cancer-associated genes onto a signalling network could uncover mechanisms of initiation, proliferation, survival, mobility and invasion of cancer cells. In this study, we mapped the cancer-associated genes onto the signalling network and found that CGs are enriched in hub proteins and cancer-associated genes are enriched or less enriched in some particular network motifs; furthermore, CGs and CMGs are enriched in the target and source nodes, respectively. In addition, we manually curated a human cellular signalling network, which, thus far, is the largest constructed cellular signalling network, and developed a strategy to extract cancer-associated SNPs from ESTs of normal and cancer tissues.

4.1 Mining of cancer-associated SNPs

Genome sequence data including cancerous ESTs increase as novel and cheaper DNA sequencing techniques are rapidly developing. We developed a more robust method to extract cancer-associated SNPs using ESTs. Compared to other reports [27], we paid more attention on controlling false positives and sequencing errors. We assigned the ESTs to genes by performing BLASTX and BLASTN to not only gene sequences but also the protein sequences. If an EST matches both a gene and its protein sequences, we assigned that EST to the gene. This could reduce the chance of wrong gene assignment of ESTs. ESTs are known as one-pass, partial sequences of cDNAs; therefore more sequencing errors appear in the end-sequencing regions. To control sequencing errors, we cut off 200 bps from the end-sequencing region of ESTs; furthermore, we defined a single mutation such that it is the only mutation and at the middle position of a 25 bp length window. We also used automatic (pMut program) and protein molecular modelling techniques to examine the potential impacts of SNPs on protein structure and function. By doing so, we could remove almost half of the insignificant SNPs that could not relate to cancer. Literature validation of the identified cancer-associated SNPs showed that almost 30% of known CG mutations are included in our list. For example, among the cancer-associated SNP genes we discovered, four of them have been found to bear cancer-related mutations: ERBB2, HDAC2, P300/CBP and RelA. Our method helps reducing false positives; however, it also loses true cancer-associated SNPs. Furthermore, by combining SNP discovery, protein structural studies and molecular modelling would help finding out cancer-associated SNPs. Nevertheless, our major goal here is to find cancer-associated SNP genes and integrate them with other types of data onto a signalling network.

4.2 Network motifs of cancer-associated genes

Cellular signal information flow initiates from extracellular space, a ligand binds to a cellular membrane receptor to start the signal, which is then transmitted by intracellular signalling components in cytosol and finally reaches the signalling components in the nucleus. In the process of signal transduction, mutated genes may result in tumourgenesis and increased cell mobility and invasion. We found that CGs are enriched in hub proteins which are the information processing centres for different signalling pathways. A few examples of such cancer hub genes can be found in the network: P53, PIK3CA, Ras, who have many regulatory partners in the network and have potentials in integrating multiple upstream signals and diverge many downstream signals [28–30]. This result suggests that mutation or deregulation of hub proteins in signalling networks could lead cells to a wrong state and promote cancer development. Furthermore, we found that CGs are enriched in downstream regions of the signalling network, especially in the nucleus. This finding supports the notion that downstream network components determine cell behaviour and evoke biological responses whereas upstream network components maintain homeostasis. Previously we showed that microRNA, a small, non-coding RNA also predominately regulates downstream components of the human singlling network [26]. A substantial amount of microRNAs has been reported to be associated with cancer [31]. Taken together, one of the mechanisms of cancer development and progression might be associated with microRNA's regulation of signalling network downstream proteins.

Errors in signal transduction lead to wrong development and behavioural decisions and sometimes result in uncontrolled growth or cancer. Signalling gene mutation or overexpression often results in signal transduction errors. To understand how mutations and overexpression of cancer-associated genes induce cancer and metastasis in complex cellular signalling networks, it is useful to identify the simplest units of commonly used network architecture. These simple units, or network motifs, such as switches [32], gates [33], positive or negative feedback loops [34] provide specific regulatory capacities and decode signal strength and process information. Both theoretical and experimental studies have shown that network motifs bear particular kinetic properties that determine the temporal program of gene expression [35]. These motifs can be self-assembled into networks that help explaining how a complex regulatory network program is regulated [17]. Therefore the frequencies and types of network motifs with which cells use reveal the regulatory strategies that are selected in different cellular conditions [17, 36]. For example, FFLs are buffers that respond only to persistent input signals [37] and are suited for endogenous conditions, although the motifs whose key regulator's transcripts have fast decay rates are preferentially used for exogenous conditions [17]. Therefore one starting point in the study of cancer signalling networks might be to characterise how cancer-associated genes are distributed in the regulatory network motifs of the signalling network. Our results showed that cancer-associated genes are enriched in some particular network motif types. This fact suggests that regulatory network motifs are critical for cancer development and metastasis. On the other hand, we found that CGs are not significantly enriched in two motif types, suggesting that these motifs provide a buffer mechanism for gene mutations, alternatively, suggesting that for some motif types having only one gene mutation is not sufficient to induce cancer. Indeed, we found that 11 and 2 three-node motifs (hotspots) in which all nodes are CGs and CMGs, respectively. We also identified nine and ten four-node motif hotspots of CGs and CMGs, respectively. These results suggest that some regulatory network motifs and network regions are important for cancer and metastasis development. The hotspots are also potentially biomarker clusters or anticancer drug target clusters. We further examined the frequencies of cancer-associated genes on node positions of each motif. Interestingly, we found that CGs are enriched on the target nodes of most motifs, especially, the convergent target nodes that receive signal information consolidated from two or more source nodes. This character hints that the convergent nodes of the CG-enriched motifs are critical nodes that might be sufficient to activate other network nodes and then induce cancer development. In the CG-enriched motifs, source nodes activate the same signalling target node. It may suggest that the source nodes could trigger the critical nodes (the convergent target nodes) for cancer development. Signalling networks govern homeostasis or promotion of cellular state changes. In signalling networks, multiple information flows could be convergent to produce a limited set of phenotypic responses [38]. The convergence provides redundant cellular functions and robustness. Critical signalling nodes fall into two categories in the network: those that preserve homeostasis during perturbation and those that evoke phenotypic changes. Taken together, the convergent nodes in the CG-enriched motifs could be the key regulators for preserving homeostasis. Therefore perturbation of these nodes would lead to losing cellular homeostasis and inducing cancer. On the other hand, the source nodes of the CMG-enriched motifs are the critical nodes for evoking phenotypic changes. These data suggest that regulatory mechanisms for cancer development and metastasis are different.

In conclusion, we developed an approach to study the relationships of these cancer-associated genes in a signalling network context. We found that CGs are enriched in hub proteins, and that cancer-associated genes are significantly enriched or depleted in some particular network motif types. More importantly, we uncovered that CGs are enriched in the convergent target nodes of most motifs, although CMGs are enriched in the source nodes of motifs. These results have implications for understanding the regulatory mechanisms of cancer development and metastasis.

5 Acknowledgments

We thank H. Hogue for setting up NCBI BLAST on computer cluster environment. This work is partially supported by Genome Health Initiative, Canada. Supplementary materials are accessible at http://www.bri.nrc.ca/wang/snp1.html.

6 References

1 Bianco, R., Melisi, D., Ciardiello, F., and Tortora, G.: 'Key cancer cell signal transduction pathways as therapeutic targets', *Eur. J. Cancer*, 2006, **42**, (3), pp. 290–294

2 Hanahan, D., and Weinberg, R.A.: 'The hallmarks of cancer', *Cell*, 2000, **100**, (1), pp. 57–70

3 Martin, G.S.: 'Cell signaling and cancer', *Cancer Cell*, 2003, **4**, (3), pp. 167–174

4 Bardelli, A., and Velculescu, V.E.: 'Mutational analysis of gene families in human cancer', *Curr. Opin. Genet. Dev.*, 2005, **15**, (1), pp. 5–12

5 Stephens, P., Edkins, S., Davies, H., Greenman, C., Cox, C., and Hunter, C.: 'A screen of the complete protein kinase gene family identifies diverse patterns of somatic mutations in human breast cancer', *Nat. Genet.*, 2005, **37**, (6), pp. 590–592

6 Bachman, K.E., Argani, P., Samuels, Y., Silliman, N., Ptak, J., and Szabo, S.: 'The PIK3CA gene is mutated with high frequency in human breast cancers', *Cancer Biol. Ther.*, 2004, **3**, (8), pp. 772–775

7 Broderick, D.K., C, Di, Parrett, T.J., Samuels, Y.R., Cummins, J.M., and McLendon, R.E.: 'Mutations of PIK3CA in anaplastic oligodendrogliomas, high-grade astrocytomas, and medulloblastomas', *Cancer Res.*, 2004, **64**, (15), pp. 5048–5050

8 Samuels, Y., and Velculescu, V.E.: 'Oncogenic mutations of PIK3CA in human cancers', *Cell Cycle*, 2004, **3**, (10), pp. 1221–1224

9 Samuels, Y., Wang, Z., Bardelli, A., Silliman, N., Ptak, J., and Szabo, S.: 'High frequency of mutations of the PIK3CA gene in human cancers', *Science*, 2004, **304**, (5670), p. 554

10 Bild, A.H., Yao, G., Chang, J.T., Wang, Q., Potti, A., and Chasse, D.: 'Oncogenic pathway signatures in human cancers as a guide to targeted therapies', *Nature*, 2006, **439**, (7074), pp. 353–357

11 Huang, E., Ishida, S., Pittman, J., Dressman, H., Bild, A., and Kloos, M.: 'Gene expression phenotypic models that predict the activity of oncogenic pathways', *Nat. Genet.*, 2003, **34**, (2), pp. 226–230

12 Downward, J.: 'Cancer biology: signatures guide drug choice', *Nature*, 2006, **439**, (7074), pp. 274–275

13 Bond, G.L., Hu, W., and Levine, A.: 'A single nucleotide polymorphism in the MDM2 gene: from a molecular and cellular explanation to clinical effect', *Cancer Res.*, 2005, **65**, (13), pp. 5481–5484

14 Collins, C.S., Hong, J., Sapinoso, L., Zhou, Y., Liu, Z., and Micklash, K.: 'A small interfering RNA screen for modulators of tumor cell motility identifies MAP4K4 as a promigratory kinase', *Proc. Natl. Acad. Sci. USA*, 2006, **103**, (10), pp. 3775–3780

15 Ma'ayan, A., Jenkins, S.L., Neves, S., Hasseldine, A., Grace, E., Dubin-Thaler, B., Eungdamrong, N.J., Weng, G., Ram, P.T., Rice, J.J., Kershenbaum, A., Stolovitzky, G.A., Blitzer, R.D., and Iyengar, R.: 'Formation of regulatory patterns during signal propagation in a mammalian cellular network', *Science*, 2005, **309**, pp. 1078–1083

16 Kashtan, N., Itzkovitz, S., Milo, R., and Alon, U.: 'Efficient sampling algorithm for estimating subgraph concentrations and detecting network motifs', *Bioinformatics*, 2004, **20**, (11), pp. 1746–1758

17 Wang, E, and Purisima, E: 'Network motifs are enriched with transcription factors whose transcripts have short half-lives', *Trends Genet.*, 2005, **21**, pp. 492–495

18 Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., and Miller, W.: 'Gapped BLAST and PSI-BLAST: a new generation of protein database search programs', *Nucleic Acids Res.*, 1997, **25**, (17), pp. 3389–3402

19 Ferrer-Costa, C., Gelpi, J.L., Zamakola, L., Parraga, I., de lC, X., and Orozco, M.: 'PMUT: a web-based tool for the annotation of pathological mutations on proteins', *Bioinformatics*, 2005, **21**, (14), pp. 3176–3178

20 Iyer, N.G., Ozdag, H., and Caldas, C.: 'p300/CBP and cancer', *Oncogene*, 2004, **23**, (24), pp. 4225–4231

21 Iyer, N.G., Chin, S.F., Ozdag, H., Daigo, Y., Hu, D.E., and Cariati, M.: 'p300 regulates p53-dependent apoptosis after DNA damage in colorectal cancer cells by modulation of PUMA/p21 levels', *Proc. Natl. Acad. Sci. USA*, 2004, **101**, (19), pp. 7386–7391

22 Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D., and Alon, U.: 'Network motifs: simple building blocks of complex networks', *Science*, 2002, **298**, (5594), pp. 824–827

23 Han, J.D., Bertin, N., Hao, T., Goldberg, D.S., Berriz, G.F., Zhang, L.V., Dupuy, D., Walhout, A.J.M., Cusick, M.E., Roth, F.P., and Vidal, M.: 'Evidence for dynamically organized modularity in the yeast protein-protein interaction network', *Nature*, 2004, **430**, (6995), pp. 88–93

24 Luscombe, N.M., Madan Babu, M., Yu, H., Snyder, M., Teichmann, S.A., and Gerstein, M.: 'Genomic analysis of regulatory network dynamics reveals large topological changes', *Nature*, 2004, **431**, (7006), pp. 308–312

25 Zhang, L.V., King, O.D., Wong, S.L., Goldberg, D.S., Tong, A.H.Y., Lesage, G., Andrews, B., Bussey, H., Boone, C., and Roth, F.P.: 'Motifs, themes and thematic maps of an integrated *Saccharomyces cerevisiae* interaction network', *J. Biol.*, 2005, **4**, (2), p. 6

26 Cui, Q., Yu, Z., Purisima, E.O., and Wang, E.: 'Principles of microRNA regulation of a human cellular signaling network', *Mol. Syst. Biol.*, 2006, **2**, p. 46

27 Qiu, P., Wang, L., Kostich, M., Ding, W., Simon, J.S., and Greene, J.R.: 'Genome wide in silico SNP-tumor association analysis', *BMC Cancer*, 2004, **4**, p. 4

28 Oikonomou, E., and Pintzas, A.: 'Cancer genetics of sporadic colorectal cancer: BRAF and PI3KCA mutations, their impact on signaling and novel targeted therapies', *Anticancer Res.*, 2006, **26**, (2A), pp. 1077–1084

29 Rodriguez-Viciana, P., Tetsu, O., Oda, K., Okada, J., Rauen, K., and McCormick, F.: 'Cancer targets in the Ras pathway', *Cold Spring Harb. Symp. Quant. Biol.*, 2005, **70**, pp. 461–467

30 Toledo, F., and Wahl, G.M.: 'Regulating the p53 pathway: in vitro hypotheses, in vivo veritas', *Nat. Rev. Cancer*, 2006, **6**, (12), pp. 909–923

31 Calin, G.A., and Croce, C.M.: 'MicroRNA-cancer connection: the beginning of a new tale', *Cancer Res.*, 2006, **66**, (15), pp. 7390–7394

32 Bhalla, U.S., Ram, P.T., and Iyengar, R.: 'MAP kinase phosphatase as a locus of flexibility in a mitogen-activated protein kinase signaling network', *Science*, 2002, **297**, (5583), pp. 1018–1023

33 Blitzer, R.D., Connor, J.H., Brown, G.P., Wong, T., Shenolikar, S., and Iyengar, R.: 'Gating of CaMKII by cAMP-regulated protein phosphatase activity during LTP', *Science*, 1998, **280**, (5371), pp. 1940–1943

34 Angeli, D., Ferrell, Jr. J.E., and Sontag, E.D.: 'Detection of multistability, bifurcations, and hysteresis in a large class of biological positive-feedback systems', *Proc. Natl. Acad. Sci. USA*, 2004, **101**, (7), pp. 1822–1827

35 Mangan, S., Zaslaver, A., and Alon, U.: 'The coherent feedforward loop serves as a sign-sensitive delay element in transcription networks', *J. Mol. Biol.*, 2003, **334**, (2), pp. 197–204

36 Balazsi, G., Barabasi, A.L., and Oltvai, Z.N.: 'Topological units of environmental signal processing in the transcriptional regulatory network of *Escherichia coli*', *Proc. Natl. Acad. Sci. USA*, 2005, **102**, (22), pp. 7841–7846

37 Mangan, S., and Alon, U.: 'Structure and function of the feed-forward loop network motif', *Proc. Natl. Acad. Sci. USA*, 2003, **100**, (21), pp. 11980–11985

38 Prinz, A.A., Bucher, D., and Marder, E.: 'Similar network activity from disparate circuit parameters', *Nat. Neurosci.*, 2004, **7**, (12), pp. 1345–1352