

Expression profiles of variation integration genes in bladder urothelial carcinoma

J.M. Wang¹, Y.Q. Wang¹, Z.L. Gao¹, J.T. Wu¹, B.K. Shi² and C.C. Yu¹

¹Department of Urology, Yantai Yuhuangding Hospital, Yantai City, Shandong Province, China

²Department of Urology, Qilu Hospital, Jinan City, Shandong Province, China

Corresponding authors: B.K. Shi / J.T. Wu

E-mail: shiyucuikang356@hotmail.com / yddurology@163.com

Genet. Mol. Res. 13 (2): 3486-3494 (2014)

Received January 23, 2013

Accepted December 17, 2013

Published April 30, 2014

DOI <http://dx.doi.org/10.4238/2014.April.30.9>

ABSTRACT. Bladder cancer is a common cancer worldwide and its incidence continues to increase. There are approximately 261,000 cases of bladder cancer resulting in 115,000 deaths annually. This study aimed to integrate bladder cancer genome copy number variation information and bladder cancer gene transcription level expression data to construct a causal-target module network of the range of bladder cancer-related genomes. Here, we explored the control mechanism underlying bladder cancer phenotype expression regulation by the major bladder cancer genes. We selected 22 modules as the initial module network to expand the search to screen more networks. After bootstrapping 100 times, we obtained 16 key regulators. These 16 key candidate regulatory genes were further expanded to identify the expression changes of 11,676 genes in 275 modules, which may all have the same regulation. In conclusion, a series of modules associated with the terms ‘cancer’ or ‘bladder’ were considered to constitute a potential network.

Key words: Bladder cancer; Module network; Control mechanism; Genes

INTRODUCTION

Bladder cancer is the fifth most common cancer and it caused more than 115,000 deaths worldwide in 2008 (Jemal et al., 2010). Although the majority of patients present with superficial bladder tumors, 20-40% either present with or develop invasive disease (Spruck et al., 1994). Another 10% of patients are diagnosed with squamous cell carcinoma, adenocarcinoma, sarcoma, small cell carcinoma, and secondary deposits from cancers elsewhere in the body.

There is currently considerable information available on the genetic alterations that contribute to the development of bladder cancer; however, to date, the specific genes associated with bladder cancer remain unknown. Some of the alterations identified in bladder cancer have clear associations with outcome; for example, mutational inactivation of the cell cycle regulator proteins p53 and the retinoblastoma (RB) protein (Habuchi et al., 1993; Brennan et al., 1995). However, on their own, these single markers have insufficient predictive power to be applied in the management of individual patients (Brennan et al., 1995).

In general, the major cause of cancer is related to genome variations, and such variants can often affect the expression of regulatory genes, which will hereafter be referred to as modulators. These modulators will cause cancer by further affecting the expression of downstream genes. Therefore, the spectrum of cancer samples and their ultimate phenotypic changes can be identified at the genome level. This study aimed to integrate bladder cancer genome copy number variation information and bladder cancer gene transcription level expression data to construct a causal-target module network of the range of bladder cancer-related genomes.

MATERIAL AND METHODS

Data resource

In order to identify the possible mechanism of bladder urothelial carcinoma (BLCA) and mutants of the genome, we collected data from The Cancer Genome Atlas (TCGA), which is a cancer research organization comprised of many institutions that publishes several types of cancer-related data, including genome, epigenetic modification, and expression profile data of mutations associated with more than 20 different types of cancers (Hede, 2008; Cancer Genome Atlas Research Network et al., 2008). The TCGA currently contains data of 59 genome copy variations and 56 gene expression profiles from bladder cancer samples. Since the experimental method required identifying data from corresponding samples for reliable results, we therefore selected 55 of the 59 genomic copy number variations that corresponded to 55 samples in the spectral expression database; these matched 55 samples served as the primary dataset for analysis.

Finally, we filtered copy variation information and gene expression information of the 55 samples, and integrated the read per kilobase of exon per million fragments Mapped (RPKM) value of every RNA sequence (RNA-seq) to a gene as the expressed array of samples.

Analysis methods

The JISTIS tool (Sanchez-Garcia et al., 2010) was used to statistically analyze genome copy number variation regions to find potentially affected genes. JISTIC is a software

tool based on the GISTIC algorithm (Beroukhim et al., 2007), which determines by calculating a G-Score to statistically identify variant regions of the genome that are associated with cancer more often than would be expected by chance. For all samples, each copy number variation was assigned a G-Score using Equation 1:

$$G^{AMP}(m) = \sum_i CN(m, i) \times I(CN(m, i) > \Theta^{AMP}) \quad (\text{Equation 1})$$

Wherein $I(x)$ represents the indicator function and $CN(m, i)$ represents the m th number of copies in the variant of sample i . The GISTIC algorithm compares the G-Score of each copy number variation in each sample with the expected copy number variation value obtained by perturbation and applies the false discovery rate (FDR) correction for multiple comparisons to obtain a q value; q values less than 0.25 were considered as potential genomic copy number variation regions.

The TCGA database provides identification of the variation region of each bladder cancer sample based on Affymetrix 6.0 chip array results. We input the segment files of each sample into JISTIC to obtain the deletion region of the candidate variants, and then identified the closest non-overlapping gene from the regions as a candidate regulator gene set, including the occurrence of copy number amplifications and reductions of the gene set.

Construction of causal-target module network with the Copy Number and Expression In Cancer (CONEXIC) algorithm

CONEXIC (Akavia et al., 2010) was used to integrate the genome variation data with the gene expression data in order to find the dominant variant gene and its dominant downstream regulatory network based on Bayesian analysis. The variation in regulatory modules of gene expression from multiple samples can be largely explained by using a combination of score-based search methods to identify a group of regulators. This method was used to build a network of causal-target modules through single modulator and network searches, applying bootstrapping to determine confidence levels of the analysis. First, CONEXIC was used to find a set of co-expressed genes, and then the most likely regulator of these co-expressed genes were identified. The procedure mainly involved the following two steps: 1) Single Modulator: We obtained a module formed by preliminarily identified co-expressed genes, and then the differentially expressed target genes and the most likely affected regulators were tested to identify the regulators with the highest confidence level based on bootstrapping. 2) Module Networks: We identified networks based on the Module Networks algorithm (Segal et al., 2003). This step takes a long time, and identifies the optimal regulation network by improving those already generated in step (1) through the cycle network integration module.

Functional analysis of the causal-target module network with Literature Vector Analysis (LitVAn)

LitVAn is a method used to detect gene modules based on a literature search. LitVAn operates on the same main principle as gene set enrichment analysis, which is to find the biological function term with the highest degree of enrichment. This method is mainly based on

the TF-IDF value, wherein TF is the term frequency in the target document, while the inverse document frequency (IDF; Equation 2) identifies the term that is the most common in the literature based on a word or rare keyword search. Higher TF-IDF values indicate that the associated group of keywords tends to appear together (Salton et al., 1988). We calculated the functional TFs among a wide range of studies related to these genes to obtain the TF-IDF value.

$$IDF(t, D) = \log \frac{|D|}{|\{d \in D : t \in d\}|} \quad (\text{Equation 2})$$

wherein D represents the data from the literature and t represents the number of occurrences of the word or keyword.

The overall experimental procedure is shown in Figure 1.

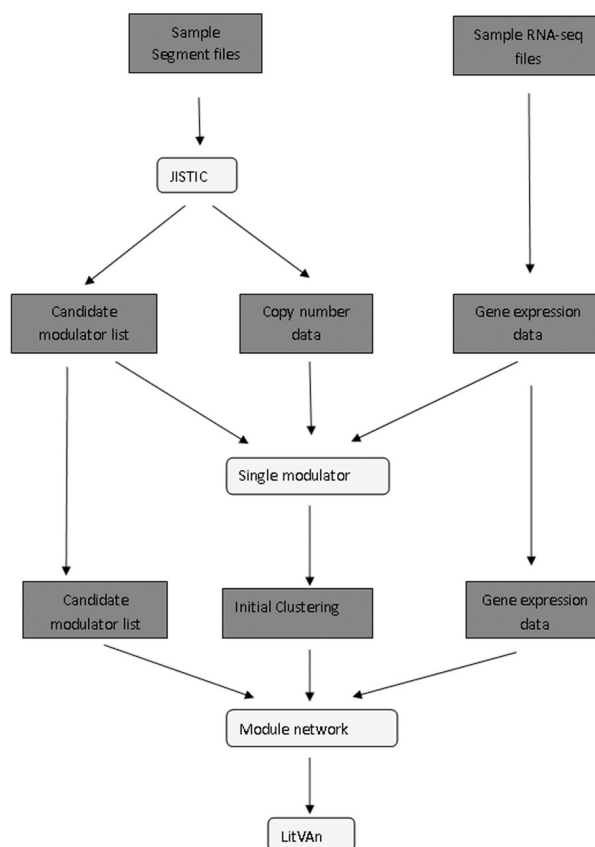


Figure 1. Overall experimental procedure. The white modules represents the methods and the gray modules represent the results.

RESULTS

JISTIC analysis of genome copy number variation regions

We input the copy variation data files and the location profile from the Affymetrix SNP-6.0 array into JISTIC and obtained 919 variation regions, including 709 affected candidate genes. After FDR correction, 205 candidate genes showed increased copy numbers and 214 candidate genes showed copy number deletions.

Signal modulators

The candidate copy number variation genes selected as candidate regulators, as well as the 55 samples of copy number variations in the expression profiling matrix were input to the CONEXIC Single Modulator module for analysis. After bootstrapping 100 times (each bootstrapping was carried out 10,000 times per mutation), the potential candidate regulators in different samples were further screened. Preliminary analysis revealed 22 modules whose candidate gene copy number variation could explain expression changes in 12,611 genes with high confidence (>90%).

Module network

The 22 modules selected for the initial module network were used to expand the search to screen for more networks. After bootstrapping 100 times, we obtained 16 key regulators, which are shown in Table 1.

Table 1. 16 CNV genes.

CNV-AMP genes	CNV-DEL genes
81607	23682
169166	1806
256435	653316
9319	219972
286	58480
34	781
5468	189
	304
	834

These 16 key genes identified as candidate regulatory genes were further expanded using the 22 modules identified in the first step (Single Modulator), and then screened (Module Network). The final results showed that the 16 key regulatory genes controlled the expression changes of 11,676 genes in 275 modules, indicating that all of these modules may have the same regulation mechanism. Although various regulatory mechanisms were identified, determining the specific function of these modules as well as their regulation mechanisms requires controlled biological experiments in organisms.

LitVAn functional analysis

The 11,676 genes identified in the 275 modules were subjected to functional analysis

by using the LitVAn literature search based on keyword co-occurrence frequency in publications related to these genes to calculate the TF-IDF value. A series of modules associated with the words 'cancer' or 'bladder' were considered to represent a potential network (Figure 2, Tables 2, 3, 4).

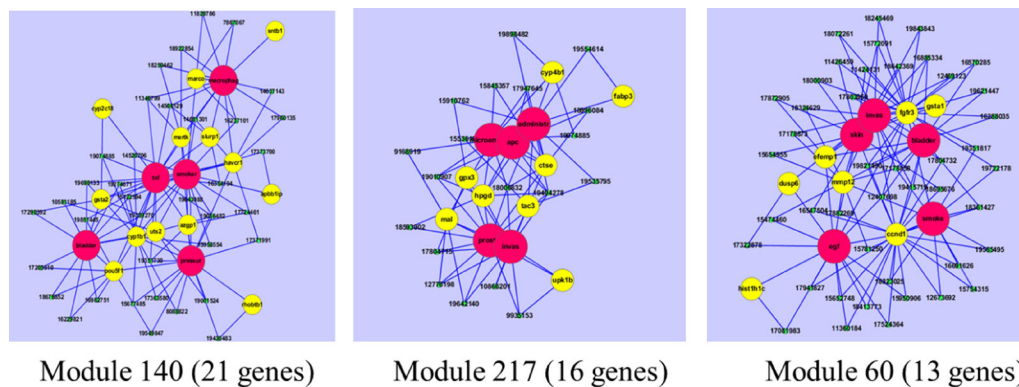


Figure 2. Gene regulation of key modules.

Table 2. Module 140 (21 genes).

Term	TF*IDF
Smoker	83.5857214407
Macrophage	82.3193406866
tnf	80.918584375
Pressure	76.5369441426
Bladder	67.6205431966
Category	62.5586649346
Discovery	62.3398500029
Caucasian	62.0480050393
Tumorigenesis	59.8631371386
Covariates	59.5879644243

TF = target document frequency; IDF = inverse document frequency; tnf = tumor necrosis factor.

Table 3. Module 217 (16 genes).

Term	TF*IDF
Prostate	53.9674123853
Invasion	41.8961155994
Microenvironment	37.2441435758
Administration	35.8849004726
Apc	35.7519550316
Metastasis	32.8811633203
Codes	31.712412
Bladder	30.6717930645
Cd11c	30.0737204026
Committe	29.897352

Apc = adenomatous polyposis coil; Cd11c = complement component 3 receptor 4 subunit. For other abbreviations, see legend to Table 2.

Table 4. Module 60 (13 genes).

Term	TF*IDF
Invasion	247.269095614
Bladder	197.451154675
Smoke	167.239032731
Egf	152.783070364
Skin	149.664463831
Collagen	143.725896548
Head	141.58796797
Neck	138.239332667

Egf = epidermal growth factor. For other abbreviations, see legend to Table 2.

DISCUSSION

Bladder cancer is a common cancer worldwide whose incidence continues to increase. It is estimated that there are 261,000 cases of bladder cancer resulting in 115,000 deaths annually worldwide. Although the majority of patients present with superficial bladder tumors, 20 to 40% either present with or develop invasive disease (Lerner et al., 1993). Therefore, the pathological stage is an important survival determinant for patients undergoing cystectomies for bladder cancer.

In this study, we employed a new analysis method to directly determine the relationship between the copy number variation status and abnormal downstream gene expression. To accomplish this, we took advantage of the one-to-one relationship of gene copy numbers and the expression spectrum from samples in the TCGA database, and the stability of the results were verified with repeated sampling inspection. The ultimate goal of this approach was to identify the major genes and regulatory genes contributing to the development of bladder cancer. Further progress in the Human Genome Project, together with the development of new technologies, such as microarray techniques to analyze global changes in gene expression (Kallioniemi et al., 1995; DeRisi et al., 1996) or gene copy number (Bignell et al., 2004), will undoubtedly lead to the identification of the genes targeted by these deletions and amplifications relatively soon.

Currently, predictive power is limited in clinical tests. Multiple genetic and epigenetic events contribute to tumor development, and few tumors are genetically identical. To improve predictive power at the molecular level, the tumor-genotype interaction can be examined in different ways. One example is the potential escape from G1 checkpoint control by mutational activation or inactivation of a gene or a group of genes involved in this control mechanism (Grossman et al., 1998). Through the present analysis, we were able to identify more than one variation region in the bladder cancer genome and the genes affected. By combining this information with the expression profiles of these genes, we identified the affected genes and the regulatory network. The module network results revealed 16 key regulators that controlled 275 modules, and these genes were found to be involved in several different regulatory mechanisms. Although these genes were identified as candidate molecular markers after bootstrapping cancer samples several times, determining the specific functions of these regulators requires further biological experimental verification.

To date, genetic studies have generally attempted to identify the spectrum of genetic changes associated with bladder cancer. Several known oncogenes and genes have been found to be mutated in cancers, including the genes encoding the G1 checkpoint proteins, p53 and

D1 (Sidransky et al., 1991; Spruck et al., 1993; Habuchi et al., 1993; Shibata et al., 1994). Predictions from the model for G1 checkpoint control are that complete dysregulation requires inactivation of RB or INK4A and p14/ARF or TP53, and that the phenotype generated by alterations to either TP53 or RB alone will be less aggressive than when both pathways are altered. The ideal situation would be to use a single marker, possibly a final effector in a pathway, to identify a specific phenotype (Cordon-Cardo et al., 1997). A larger study will be needed to confirm the apparent predictive power of this single marker.

In addition, there is currently no clear understanding of the molecular basis of the differentiation process involved in bladder cancer. Although identifying the mechanism involved in bladder cancer-related cell phenotypes, such as mutants and variation of cell cycle checkpoints, would yield a potentially powerful suite of markers, in the long term, it would be simpler to identify single protein surrogates for these phenotypes that are easily measurable by using immunohistochemistry.

ACKNOWLEDGMENTS

Research supported by grants from the Department of Science and Technology of Shandong Province (Grant #BS2011YY063), and the Department of Health of Shandong Province (Grant #2011QZ030).

REFERENCES

- Akavia UD, Litvin O, Kim J, Sanchez-Garcia F, et al. (2010). An integrated approach to uncover drivers of cancer. *Cell* 143: 1005-1017.
- Beroukhi R, Getz G, Nghiemphu L, Barretina J, et al. (2007). Assessing the significance of chromosomal aberrations in cancer: methodology and application to glioma. *Proc. Natl. Acad. Sci. U. S. A.* 104: 20007-20012.
- Bignell GR, Huang J, Greshock J, Watt S, et al. (2004). High-resolution analysis of DNA copy number using oligonucleotide microarrays. *Genome Res.* 14: 287-295.
- Brennan JA, Boyle JO, Koch WM, Goodman SN, et al. (1995). Association between cigarette smoking and mutation of the p53 gene in squamous-cell carcinoma of the head and neck. *N. Engl. J. Med.* 332: 712-717.
- Cancer Genome Atlas Research Network (2008). Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* 455: 1061-1068.
- Cordon-Cardo C, Zhang ZF, Dalbagni G, Drobnjak M, et al. (1997). Cooperative effects of p53 and pRB alterations in primary superficial bladder tumors. *Cancer Res.* 57: 1217-1221.
- DeRisi J, Penland L, Brown PO, Bittner ML, et al. (1996). Use of a cDNA microarray to analyze gene expression patterns in human cancer. *Nat. Genet.* 14: 457-460.
- Grossman HB, Liebert M, Antelo M, Dinney CP, et al. (1998). p53 and RB expression predict progression in T1 bladder cancer. *Clin. Cancer Res.* 4: 829-834.
- Habuchi T, Takahashi R, Yamada H, Ogawa O, et al. (1993). Influence of cigarette smoking and schistosomiasis on p53 gene mutation in urothelial cancer. *Cancer Res.* 53: 3795-3799.
- Hede K (2008). Superhighway or blind alley? The cancer genome atlas releases first results. *J. Natl. Cancer Inst.* 100: 1566-1569.
- Jemal A, Siegel R, Xu J and Ward E (2010). Cancer statistics, 2010. *CA Cancer J. Clin.* 60: 277-300.
- Kallioniemi A, Kallioniemi OP, Citro G, Sauter G, et al. (1995). Identification of gains and losses of DNA sequences in primary bladder cancer by comparative genomic hybridization. *Genes Chromosomes Cancer* 12: 213-219.
- Lerner SP, Skinner DG, Lieskovsky G, Boyd SD, et al. (1993). The rationale for en bloc pelvic lymph node dissection for bladder cancer patients with nodal metastases: long-term results. *J. Urol.* 149: 758-764.
- Salton SR, Blum M, Jonassen JA, Clayton RN, et al. (1988). Stimulation of pituitary luteinizing hormone secretion by gonadotropin-releasing hormone is not coupled to beta-luteinizing hormone gene transcription. *Mol. Endocrinol.* 2: 1033-1042.

- Sanchez-Garcia F, Akavia UD, Mozes E and Pe'er D (2010). JISTIC: identification of significant targets in cancer. *BMC Bioinformatics* 11: 189.
- Segal E, Shapira M, Regev A, Pe'er D, et al. (2003). Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nat. Genet.* 34: 166-176.
- Shibata A, Ohneseit PF, Tsai YC, Spruck CH, III, et al. (1994). Mutational spectrum in the p53 gene in bladder tumors from the endemic area of black foot disease in Taiwan. *Carcinogenesis* 15: 1085-1087.
- Sidransky D, Von EA, Tsai YC, Jones P, et al. (1991). Identification of p53 gene mutations in bladder cancers and urine samples. *Science* 252: 706-709.
- Spruck CH, III, Rideout WM, III, Olumi AF, Ohneseit PF, et al. (1993). Distinct pattern of p53 mutations in bladder cancer: relationship to tobacco usage. *Cancer Res.* 53: 1162-1166.
- Spruck CH, III, Ohneseit PF, Gonzalez-Zulueta M, Esrig D, et al. (1994). Two molecular pathways to transitional cell carcinoma of the bladder. *Cancer Res.* 54: 784-788.