Published in final edited form as:

Syst Biol (Stevenage). 2006 January ; 153(1): 4–12.

A Template-Driven Gene Selection Procedure *

Nicholas Knowlton¹, Igor Dozmorov¹, Kimberly D. Kyker², Ricardo Saban³, Craig Cadwell¹, Michael B. Centola¹, and Robert E. Hurst^{2,4,5}

¹ Microarray Core Facility, Oklahoma Medical Research Foundation, Oklahoma City, OK 73104

² Department of Urology,

³ Department of Physiology,

⁴ Department of Biochemistry and Molecular Biology,

⁵ Department of Occupational and Environmental Health, Oklahoma University Health Sciences Centre, Oklahoma City, OK 73104

Abstract

The hierarchical clustering and statistical techniques usually used to analyze microarray data do not inherently represent the underlying biology. Herein we present a hybrid approach involving characteristics of both supervised and unsupervised learning. This approach is based on template matching in which the interaction of the variables of inherent malignancy and the ability to express the malignant phenotype are modelled. Immortalized normal urothelial cells and bladder cancer cells of different malignancy were grown in conventional two-dimensional tissue culture and in three dimensions on extracellular matrices that were either permissive or restrictive for expression of the malignant phenotype. The transcriptome represents the effects of two variables--inherent malignancy and the modulatory effect of extracellular matrix. By assigning values to each of the biological variables of inherent malignancy and the ability to express the malignant phenotype, a template was constructed that encapsulated the interaction between them. Gene expression correlating both positively and negatively with the template were observed, but when iterative correlations were carried out, the different models for the template converged to the same actual template. A subset of 21 genes was identified that correlated with two *a priori* models or an optimized model above the 95% confidence limits identified in a bootstrap resampling with 5,000 permutations of the data set. The correlation coefficients of expression of several genes were > 0.8. Analysis of upstream transcriptional regulatory elements (TREs) confirmed these genes were not a randomly selected set of genes. Several TREs were identified as significantly over-expressed in the sample of 20 genes for which TREs were identified, and the high correlations of several genes were consistent with transcriptional co-regulation. We suggest the template method can be used to identify a unique set of genes for further investigation.

Keywords

Bladder cancer; phenotype; transcriptomics; extracellular matrix; malignancy

*This work was supported in part by NIH grants CA75322 (REH), DK 069808 (REH) and P20 RR1557, P20 RR17703, and P20 RR16478 (MBC) and by a grant from Cook Biotech (REH). Running title: Template of Gene Expression

Address correspondence to: Robert E. Hurst, PhD, Department of Urology, Oklahoma University Health Sciences Center, 940 S.L. Young Blvd, Oklahoma City, OK 73104, TEL: 405-271-3930, FAX: 405-271-3289, Email: Robert-hurst@ouhsc.edu.

The gene expression data are available at the GEO website under accession number GSE796: <http://www.ncbi.nlm.nih.gov/geo/>

Abbreviations

SIS-small intestine submucosa; ECM-extracellular matrix; TCC-transitional cell carcinoma

1. Introduction

The development of microarray techniques that survey a significant fraction of the transcriptome has, in theory, allowed the process of tumour formation to be examined at the system level. However, microarrays are severely underdetermined, and the complex interactions among cellular signalling pathways combined with the noise inherent in the experimental determination of expression means not only is a unique solution to a microarray experiment improbable, there may not even be an “optimal” one [1].

One of the more effective approaches to reducing the effect of noise is to systematically perturb the biology underlying gene expression and identify genes that are thereby modulated. The interaction between cancer cells and the ECM can modulate malignant behaviour both *in vitro* [2;3] and *in vivo* [4]. This modulation is highly relevant to human cancer because metastatic cells often remain dormant for years before emerging as tumours [5], and malignant cells can often masquerade as normal cells before emerging as a recurrence [6]. Understanding the mechanisms for this modulation of the malignant phenotype by ECM may present important clues for cancer treatment or management.

In this paper, five bladder cancer cell lines differing in inherent malignancy (three low grade and two high grade) and one immortalized, but nontumorigenic bladder epithelial cell line were grown on two different ECM preparations (Matrigel and SISgel) and on plastic. On the cancer-modified ECM, Matrigel, the malignant phenotype of the cells studied herein is expressed fully, whereas on SISgel, which is prepared from normal submucosa, the cells display a more normalized, layered phenotype in which invasion is suppressed and the cell layer shows evidence of differentiation [2]. On conventional tissue culture on plastic, the modulating effects of the matrix are absent. This afforded us a means whereby both inherent malignancy and the effect of ECM can be systematically varied to identify genes that modulate the malignant phenotype. The expression levels of 1167 well-annotated genes selected for their relevance to cancer biology in general were determined on a Nylon array to identify such genes.

To analyze the resulting complex data set, we developed a novel template approach that describes the interaction of these two biological variables. The template was developed iteratively from a conceptual model of gene expression in which relevant genes would be expected to increase expression with both increasing malignancy and permissiveness for malignant growth. This template was compared to the expression levels of the 87 genes that were expressed more than 3 s.d. above background. The template model discovered a pattern of interesting gene expression that correlated with the interaction of the modulating effect of ECM on expression of the malignant phenotype. We suggest this template approach may prove useful to finding genes that describe other systems in which two biological variables affect behaviour.

2. Materials and Methods

2.1. Cell Culture

SV-HUC-1, TCCSUP, RT4 and J82 cells were obtained from the American Type Culture Collection, Bethesda, MD, which provided information allowing the cells to be ranked by malignancy of the tumour of origin. The 253 J and 253 JB-V cells were provided by Dr. Colin Dinney [7]. The former is derived from a metastatic lymph node tumour, while the latter is a

highly metastatic variant cloned in Dr. Dinney's laboratory after 5 passages of 253 J cells in the bladder walls of nude mice. Although metastatic, the tumour morphology is papillary but invasive. Details of cell culture on Matrigel and SISgel have been reported previously [2;3]. The ranking according to malignancy from lowest to highest is: SV-HUC-1 (non-malignant but immortalized), RT4 (low grade), 253 J (moderate grade) 253 JB-V (moderate grade), J82 (high grade), TCCSUP (high grade). Excepting the non-malignant SV-HUC-1 cells and possibly the TCCSUP cells (see below), all the cancer cell lines are transitional cell carcinomas (TCC).

2.2. Array Protocol

RNA was isolated from the cells growing in gel using the RNeasy kit (Qiagen) by adding 300 μ l lysis buffer to the culture well and pipeting up and down to lyse the cells and dissolve the gel. The RNA was isolated from the lysate using a QIAshredder spin column to complete homogenization followed by proteinase K digestion, washing, DNase I treatment and elution from RNeasy spin column. Yield of total RNA and the purity were assessed by spectrophotometry. The yield was approximately 2.4 – 6 μ g RNA per culture in 30 μ l. First strand synthesis was carried out at 42° C for 1 h. with reagents supplied by Clontech as part of the SMART II kit, except that Superscript II (Invitrogen) reverse transcriptase was substituted. The SMART II oligonucleotide was included to capture full length cDNA at the 5' end and to permit a proportional 2,000 – 5,000 fold amplification [8]. Labelled probe was produced from an aliquot of the amplified probe with Klenow fragment and 32 P-labeled nucleotide. After probe purification, 15×10^6 cpm of probe was hybridized in duplicate to the Clontech Human Cancer 1.2 array on nylon following the protocol provided by the company. The Clontech Human Cancer 1.2 array is a focused array consisting of 1186 well-annotated cancer-associated genes. The results were visualized by exposing a phosphorimager screen overnight and then reading with the Packard Cyclone phosphorimager and Optiquant software.

2.3. Data Normalization

Data were normalized as described in detail elsewhere [9]. In brief, the normalization method relies on a number of low expression genes to provide an estimate of non-specific binding. This information is then used to perform a Z transformation on the data. Once normalized the data are “unbiased” through robust linear regression to allow comparisons across different membranes. Fitted data are then used to find a homoscedastic group of gene variances that will be used as an internal standard of measurement noise [10]. Genes statistically likely to be expressed (expression levels 3 standard deviations above the mean of the background) were identified, anti-log transformed, and used for subsequent analysis. This filtering step minimizes false positives, though at the cost of the low-expression genes, which are measured with low precision.

Primary template modelling and optimization—Because microarray data are severely underdetermined, many solutions to a given dataset are possible. Our interest was to identify genes whose expression followed a particular type of behaviour, namely that they were modulated both by intrinsic malignancy and by the ECM. We first developed a two-dimensional template of expected gene expression starting from the concept that as cells became more malignant or the ECM more permissive, expression of key genes related to malignancy and the ECM would increase or decrease in expression. We therefore assigned values of 0, 1 and 2 to malignancy for the normal, low-grade and high-grade cells, and 0, 1 and 2 to permissiveness of the ECM on plastic, SISgel and Matrigel respectively. In order to provide a starting place (i.e. the “seed”) for deriving the optimal template, we assumed malignancy and the ECM could interact in two simple ways. In the Linear Model, changes in gene expression derived from inherent malignancy and those from the effect of the ECM were additive. In the Super-Linear Model, a multiplicative interaction was assumed except that where zero resulted;

a value of 1 was substituted. The templates are described in Table 1. The model template (T) was iteratively optimized by finding the maximal Pearson correlation coefficient (ρ) between each gene profile and template. The optimal template was created iteratively by increasing the stringency with every iteration to ensure convergence of the algorithm.

The algorithm can be defined as follows:

1. Generate an a priori seed that will define the initial gene correlation pattern expected.
2. Compute a Pearson's Product-Moment Correlation (ρ) for every gene against the *a priori* template.

$$\rho = \frac{n(\sum ST) - (\sum S)(\sum T)}{\sqrt{[n\sum S^2 - (\sum S)^2][n\sum T^2 - (\sum T)^2]}}$$

where sums are taken over all n values in each profile.

1. Using only genes significantly correlated with the template, here $|\rho| > 0.6$, average across all expression values in a given experimental unit.
2. These averages create the new expression template
3. Increase statistical significance of ρ by 0.005
4. Compute Pearson's Product-Moment Correlation (ρ) for every gene against the new template
5. Repeat steps (3–6) until $p > 0.99$ or fewer than 4 genes correlate with the template

The linear and super-linear initial “seeds” were given as starting points for the algorithm, and all resulted in the same final template. In order to estimate the significance of correlations of expression within the template without assuming an underlying distribution, the data columns were randomly rearranged (bootstrapped) 5,000 times [11]. All seeds converged on the same final template. The distribution of correlations in the bootstrap analysis is shown in Fig. 1. We selected genes that correlated above the 95% confidence interval.

2.4. Promoter Analysis

The gene list identified from the template analysis was analyzed for transcriptional regulatory elements (TREs) using the program PAINT [12]. This program identifies consensus promoter sequences within a 2000 bp upstream segment of the gene start sequence and from this provides a list of transcription factors. A probability of the observed pattern is determined by reference against a comparison database, in this case determined from a 21K representation of the human genome. The expression of the transcription factors identified as possibly being relevant was confirmed by analysis on a separate long oligonucleotide array containing the transcription factor probes.

3. Results

3.1. Data Quality and Reproducibility

Fig. 2 shows a representative duplicate pair of analyses plotted after filtering out the genes expressed below 3 SD above background. Although accurate statistical measurements of mean expression and detection of differentially expressed genes by t-tests are not feasible with only duplicate measurements, the data are nonetheless reproducible, and less than 0.1% of genes vary by more than 4-fold among duplicates.

3.2. Template Model Analysis

Table 1 shows the optimized template derived from the optimization studies beginning with either a linear or super-linear seed. Genes were found that fit either seed, but both seeds converged on the same optimal template when the best fit was sought. The optimized template more closely resembled the super-linear model than the linear one. Table 2 lists gene transcripts that correlate above the 95% confidence limit determined from the bootstrapping analysis. This limit was at $\rho \geq 0.59$ and $\rho \leq -0.50$ (Fig. 1). Both positive and negative correlations with the model were observed. Positive correlations are likely to represent genes that are associated with malignancy whereas those with negative ones are expected to be associated with a more normal pattern of expression that are lost as malignancy progresses. The TCCSUP line showed the poorest correlation with the models.

3.3. K-Means Clustering

When K-means clustering is applied to our data set of expressed genes, Table 2 shows that the genes identified by our template model do not fall into a unique cluster for either $k = 4$ or $k = 5$.

3.4. Correlation of Expression

Fig. 3 shows the correlations of all the gene expression values with each other. Obviously, the positively-correlating and negatively-correlating gene transcripts are dissimilar as groups, as might be expected from the template models. The positively-correlating genes show some interesting clustering of expression that suggests that several of these may be regulated in common. The two interleukins were very closely correlated (0.99), and several others showed correlations with each other greater than 0.8. The negatively-correlating genes do not show such high correlations either with each other or with the positively-correlating genes.

3.5. TRE Analysis

PAINT analysis identified a number of consensus promoter sequences in the upstream portions of the genes listed in Table 2. The results are illustrated in Fig. 4 by whether the transcripts correlated positively or negatively with the template. Transcription factors with $p < 0.1$ generally are considered as worthy of further consideration [12]. Several sequences with highly significant over-representation, notably Pax-4, FoxD3, Oct-1, and Nkx2-5 binding sites were identified. Those showing a strong differential between the positively and negatively correlating genes could be considered as driving either malignancy or normalization, respectively. All of the transcription factors that were identified as possibly being relevant in Table 3 were found to be expressed at least 2 sd above background at least under some conditions. The Pax-4 and FoxD3 transcription factors were expressed under all conditions between 10 and 1000 times above background. The Oct-1 and Nkx2-5 transcription factors generally were not present in cells grown on plastic. Oct-1 was expressed more highly on SISgel than on Matrigel at levels, ranging from 2–20 times above background, whereas Nkx2-5 transcription factor showed no consistent differential expression. The promoter sequences found in common are consistent with some of the high correlations seen in Fig. 3. For example, ErbB3, HLAC, IL1 β , RPS9 and CDKN1A all share a Nkx2-5 sequence, whereas CDKN1A, TMP21, HLAC, Casp1 and IL1 β all share an Oct-1 sequence. Fig. 5 shows the network of connections among putative transcription factors and template-derived genes coded according to the mathematical sign of the correlation with the template. The associations among correlation coefficients and TREs is more evident in this figure and demonstrates that the template method identifies genes whose expression is co-ordinately regulated.

4. Discussion

A major challenge with microarray data is delineating those changes in expression that are associated with scientifically or medically relevant differences in biological behaviour from those that are inconsequential to the biology [13]. At present, statistical techniques in which there are no *a priori* hypotheses concerning the biology are used most frequently to analyze microarray experiments [14]. This kind of analysis is referred to as “unsupervised learning” and can actually add complexity [15], and neither hierarchical clustering nor self-organizing maps provides a p-value with the analysis. Even with the criterion that the analyses must pass some rigorous test of significance, any set of microarray data still has multiple solutions, none of which may even be optimal in the sense of being inherently superior to other analyses [1;16]. Moreover, genes identified by clustering analyses of microarray data are rarely the same ones identified as being significant from clinical, in situ, molecular, single-nucleotide polymorphism (SNP) association, knockout and drug perturbation data [16]. Although clustering and statistical techniques can be quite powerful, it is clear that additional improvements in analysis of microarrays will prove useful.

“Supervised learning” seeks to include knowledge of the biology driving the changes in gene expression in the analysis of the experimental results. In this communication we report a novel hybrid template approach incorporating both supervised and unsupervised characteristics. The method is based on sound statistical criteria and particularly seems useful for extracting a useful solution when two, interacting biological variables are at work. Such an analytical approach can be useful to analysis of experiments in which systematic manipulation of relevant biological variables elicits a patterned, system-level response that facilitates separating biologically relevant responses from random or irrelevant ones.

Given the above, is therefore not surprising that the genes in Table 2 show little in common with clinical bladder cancer samples analyzed with the Affymetrix HuGeneFl array containing approximately 5,600 genes [17;18], or bladder cancer cell lines grown on plastic using the IMAGE cDNA set of 8,976 ESTs and genes of known function [19]. Interestingly, these other two studies based on different microarray platforms showed only 3 genes in common with each other, in spite of using very similar analytical techniques. Yet, each study certainly is valid within the context of the analysis, and each combination of arrays and analyses selected a distinct and different pattern of expression for reasons discussed more fully by Ein-Dor and colleagues [1]. The microarray truly might be considered a scientific tool for the post-modernist age in which no “narrative” is inherently privileged, and “meaning” depends on context.

The proposed clustering method can be applied almost universally to any system where there are known or suspected biological interactions. Additionally, the method is fairly robust to initial seed variations. For example, if this method was applied to identify genes modulated by circadian rhythm, the only biologic knowledge needed is that genes of interest oscillate up and down within a 24 hour timeframe, i.e. initial seed 0, 1, 0, -1, 0. An optimal template, based upon these assumptions, will be generated along with a list of genes that follow the predefined pattern. This template method finds a unique set of genes that other unsupervised methods can overlook (Table 2). The difference arises due to K-means algorithms separating genes into groups based on overall expression level. Our method relies on changes in expression, not their level, to determine cluster membership.

Although establishing a template of expected values and identifying genes that match an expected template is not itself novel, previous approaches have either sought to classify results into different “bins,” such as identifying genes that varied with the grade of tumour [17], or to match expression to linear, binary templates of “high expression” or “low expression”[20]. The current approach allowed us to model quantitatively the interaction between two complex

biological variables and to refine the model on the basis of observed data. Both *a priori* models converged to the same optimized model, demonstrating that *a priori* predictions could be used to find a pattern of gene expression that was inherent in the data but which more or less followed theoretical patterns of gene expression. In this case a subset of genes was identified that increased or decreased systematically as the malignant potential of the cells increased and also as permissiveness of growth conditions increase. That the template was approximately geometric rather than linear implies that the effects of mutations, gene silencing and other primary modulators of expression and function are approximately multiplicative as cancers progress.

Although validation often is equated to showing equivalence of expression of a few genes by an independent method (e.g. PCR or Northern blot), this is only the first level of validation [21;22]. Far more interesting and complex, however, is the concept of biological validation [21]. Examination of the gene list showed several genes have been associated clinically with bladder cancer. The template gene with the highest positive correlation with the template, ErbB3, is a member of the EGF receptor family and has reported to bear a strong association with bladder cancer progression [23;24]. IL-10 expression has been identified as a suppressor of the immune response *in vivo* that is associated with progression [25;26]. IL-1 β has been reported in elevated amounts in the urines of bladder cancer patients [27]. Caspase 1 (interleukin 1 converting enzyme), which cleaves the inactive precursor to the active cytokine, also showed a positive association with the template. Apparently contrary associations were seen for expression of HLA-C, which is reported to decrease with progression [28] and with CDKN1A, which is a cell-cycle regulator that produces growth arrest in bladder cancer cells [29]. However, it is not known whether other defects render the implied regulation ineffective. Without knowing the fraction of genes on the array that are associated with bladder cancer, the significance of the observation that several of the positively-correlating subset has been associated clinically with bladder cancer cannot be tested rigorously. Nonetheless, this observation does suggest the template approach identifies a different set of genes than does the typical cluster analysis.

The negatively-correlating genes are, for the most part, associated with normal cellular processes that would be expected to be lost as the cells express the malignant phenotype more strongly. DAP3 is associated with anoikis (cell death following detachment from substrate) [30], but most cancer cells are not subject to this mechanism. Thus, loss of this gene would be expected to be associated with progression.

The data themselves offer another interesting validating clue. Even though the original tumour from which TCCSUP was derived was classified as a TCC, TCCSUP cells form squamous cell carcinomas in xenografts [31]. The TCCSUP line failed to fit the template as well as the other lines, which suggests some common pathways shared by all TCCs are not relevant to the TCCSUP line. Thus, it may prove possible to derive signatures of specific cancer types by this template process.

Examination of the correlation coefficients of the expression levels shows that the expression of the genes associated with the optimal template are mostly above 0.8, whereas those that are associated only with the *a priori* templates are lower. This suggests that the optimal template has identified a set of co-regulated genes that may have an important biological function in establishing the phenotype. This hypothesis is supported by the identification of highly over-represented promoter sequences that are consistent with the correlations among gene expression levels. This finding also presents a second and independent level of biological confirmation of the validity of the template approach. Figs. 3 and 4 demonstrate that the significant genes identified in Table 2 are not a random collection of genes. A few promoter sequences are highly over-represented, and the hierarchical clustering distinguishes the

positively- and negatively-correlated genes. Many of the high correlations seen in Fig. 3 are consistent with transcriptional regulation by the 4 transcription factors with the lowest p-values as being responsible. Interestingly, the promoter sequences associated with genes identified only by the optimal template model seem to be somewhat different from those associated with genes identified solely by the two *a priori* models. This is seen most clearly in Fig. 5, where those associated uniquely with the optimal template seem to contain mostly Oct-1, Nkx2-5 and Pax-4 sequences, whereas those associated with the *a priori* models involve a number of other sequences such as ISRE, SOX-9, AREB-6, FOX, J2, FOXD3 and cMyb.

Some investigators have suggested that the function of microarrays is to identify pathways that are dysregulated, and that differentially expressed genes are just one means to achieve this [32]. The negatively-correlating gene LRRC16 contains a RAR β promoter sequence. Loss of RAR β due to methylation silencing has been associated with progression of many cancers including bladder cancer [33]. Expression of hRAR β itself was low and too noisy for firm conclusions, but the negative correlation of LRRC17 is consistent with the progression-related loss of hRAR β regulation. LRRC17 (P37NB), is a gene of unknown function previously associated with neuroblastoma cells expressing a differentiated phenotype [34] and may play a differentiation-related role in urothelial cells as well.

The template matching approach here presented should prove to be a useful tool for analyzing gene array data sets and appeared to identify genes that relate to the biology underlying malignant transformation. The template matching approach includes statistical criteria for validating gene selections through the bootstrap simulation. Further research at the biological level clearly will be needed to validate the genes implicated with this method. The findings suggest that it identifies co-ordinately regulated genes as well as genes also identified from clinical studies. This approach can supplement conventional techniques in analyzing transcriptome data for genes that relate to the mechanism of cancer and to identify markers or targets for therapy.

Acknowledgements

The authors thank Dr. Colin Dinney of the MD Anderson Cancer Center for his gift of 253 J and 253 JB-V cells, Cook Biotech (W. Lafayette, IN) for providing SISgel, and Jean Coffman for her excellent technical assistance in growing cells.

References

1. Ein-Dor L, Kela I, Getz G, Givol D, Domany E. Outcome signature genes in breast cancer: is there a unique set? *Bioinformatics* 2005 Jan 20;21(2):171–8. [PubMed: 15308542]
2. Hurst RE, Kyker KD, Bonner RB, Bowditch RG, Hemstreet GP. Matrix-Dependent Plasticity of the Malignant Phenotype of Bladder Cancer Cells. *Anticancer Res* 2003;23(4):3119–28. [PubMed: 12926044]
3. Kyker KD, Culkin DJ, Hurst RE. A model for 3-dimensional growth of bladder cancers to study mechanisms of phenotypic expression. *Urologic Oncology* 2003;21(4):255–61. [PubMed: 12954494]
4. Bonkhoff H. Analytical molecular pathology of epithelial-stromal interactions in the normal and neoplastic prostate. *Anal Quant Cytol Histol* 1998 Oct;20(5):437–42. [PubMed: 9801762]
5. Kauffman EC, Robinson VL, Stadler WM, Sokoloff MH, Rinker-Schaeffer CW. Metastasis suppression: the evolving role of metastasis suppressor genes for regulating cancer cell growth at the secondary site. *J Urol* 2003 Mar;169(3):1122–33. [PubMed: 12576866]
6. Koch WM, Boyle JO, Mao L, Hakim J, Hruban RH, Sidransky D. p53 gene mutations as markers of tumor spread in synchronous oral cancers. *Arch Otolaryngol Head Neck Surg* 1994;120(9):943–7.
7. Dinney CP, Fishbeck R, Singh RK, Eve B, Pathak S, Brown N, et al. Isolation and characterization of metastatic variants from human transitional cell carcinoma passaged by orthotopic implantation in athymic nude mice. *J Urol* 1995 Oct;154(4):1532–8. [PubMed: 7658585]

8. Seth D, Gorrell MD, McGuinness PH, Leo MA, Lieber CS, McCaughan GW, et al. SMART amplification maintains representation of relative gene expression: quantitative validation by real time PCR and application to studies of alcoholic liver disease in primates. *J Biochem Biophys Methods* 2003 Jan 31;55(1):53–66. [PubMed: 12559588]
9. Dozmorov I, Knowlton N, Tang Y, Shields A, Pathipvanich P, Jarvis JN, et al. Hypervariable genes--experimental error or hidden dynamics. *Nucleic Acids Res* 2004 Oct;32(19):e147. [PubMed: 15514108]
10. Dozmorov I, Knowlton N, Tang Y, Centola M. Statistical monitoring of weak spots for improvement of normalization and ratio estimates in microarrays. *BMC Bioinformatics* 2004 May 5;5(1):53. [PubMed: 15128432]
11. Efron B, Gong G. A Leisurely Look at the Bootstrap, the Jackknife, and Cross-Validation 1. *American Statistician* 1983;37(1):36–48.
12. Vadigepalli R, Chakravarthula p, Zak DE, Schwaber JS, Gonye GE. PAINT: A Promoter Analysis and INteraction Network Generation Tool for Gene Regulatory Network Identification. *OMICS: A Journal of Integrative Biology* 2003;7(3):235–52. [PubMed: 14583114]
13. Hampton GM, Frierson HF. Classifying human cancer by analysis of gene expression. *Trends Mol Med* 2003 Jan;9(1):5–10. [PubMed: 12524204]
14. Mateos A, Dopazo J, Jansen R, Tu Y, Gerstein M, Stolovitzky G. Systematic learning of gene functional classes from DNA array expression data by using multilayer perceptrons. *Genome Res* 2002 Nov;12(11):1703–15. [PubMed: 12421757]
15. Brown MP, Grundy WN, Lin D, Cristianini N, Sugnet CW, Furey TS, et al. Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proc Natl Acad Sci U S A* 2000 Jan 4;97(1):262–7. [PubMed: 10618406]
16. Miklos GL, Maleszka R. Microarray reality checks in the context of a complex disease. *Nat Biotechnol* 2004 May;22(5):615–21. [PubMed: 15122300]
17. Thykjaer T, Workman C, Kruhoffer M, Demtroder K, Wolf H, Andersen LD, et al. Identification of gene expression patterns in superficial and invasive human bladder cancer. *Cancer Res* 2001 Mar 15;61(6):2492–9. [PubMed: 11289120]
18. Dyrskjot L, Thykjaer T, Kruhoffer M, Jensen JL, Marcussen N, Hamilton-Dutoit S, et al. Identifying distinct classes of bladder carcinoma using microarrays. *Nat Genet* 2002 Dec 9;33(1):90–6. [PubMed: 12469123]
19. Sanchez-Carbayo M, Socci ND, Charytonowicz E, Lu M, Prystowsky M, Childs G, et al. Molecular profiling of bladder cancer using cDNA microarrays: defining histogenesis and biological phenotypes. *Cancer Res* 2002 Dec 1;62(23):6973–80. [PubMed: 12460915]
20. Pavlidis P, Noble WS. Analysis of strain and regional variation in gene expression in mouse brain. *Genome Biol* 2001 Jan;2(10):research0042.1–research0042.15. [PubMed: 11597334]
21. Chuaqui RF, Bonner RF, Best CJ, Gillespie JW, Flaig MJ, Hewitt SM, et al. Post-analysis follow-up and validation of microarray experiments. *Nat Genet* 2002 Dec;32(Suppl):509–14. [PubMed: 12454646]
22. Tan PK, Downey TJ, Spitznagel EL Jr, Xu P, Fu D, Dimitrov DS, et al. Evaluation of gene expression measurements from commercial microarray platforms. *Nucleic Acids Res* 2003 Oct 1;31(19):5676–84. [PubMed: 14500831]
23. Junttila TT, Laato M, Vahlberg T, Soderstrom KO, Visakorpi T, Isola J, et al. Identification of patients with transitional cell carcinoma of the bladder overexpressing ErbB2, ErbB3, or specific ErbB4 isoforms: real-time reverse transcription-PCR analysis in estimation of ErbB receptor status from cancer patients. *Clin Cancer Res* 2003 Nov 1;9(14):5346–57. [PubMed: 14614020]
24. Chow NH, Chan SH, Tzai TS, Ho CL, Liu HS. Expression profiles of ErbB family receptors and prognosis in primary transitional cell carcinoma of the urinary bladder. *Clin Cancer Res* 2001 Jul;7(7):1957–62. [PubMed: 11448910]
25. Cardillo MR, Sale P, Di SF. Heat shock protein-90, IL-6 and IL-10 in bladder cancer 12. *Anticancer Res* 2000 Nov;20(6B):4579–83. [PubMed: 11205307]
26. Loskog A, Dzojic H, Vikman S, Ninalga C, Essand M, Korsgren O, et al. Adenovirus CD40 ligand gene therapy counteracts immune escape mechanisms in the tumor Microenvironment 3. *J Immunol* 2004 Jun 1;172(11):7200–5. [PubMed: 15153545]

27. Martins SM, Darlin DJ, Lad PM, Zimmern PE. Interleukin-1B: a clinically relevant urinary marker. *J Urol* 1994 May;151(5):1198–201. [PubMed: 8158759]
28. Cordon-Cardo C, Fuks Z, Drobnjak M, Moreno C, Eisenbach L, Feldman M. Expression of HLA-A,B,C antigens on primary and metastatic tumor cell populations of human carcinomas. *Cancer Res* 1991 Dec 1;51(23 Pt 1):6372–80. [PubMed: 1933900]
29. Hall MC, Li Y, Pong RC, Ely B, Sagalowsky AI, Hsieh JT. The growth inhibitory effect of p21 adenovirus on human bladder cancer cells. *J Urol* 2000 Mar;163(3):1033–8. [PubMed: 10688044]
30. Miyazaki T, Shen M, Fujikura D, Tosa N, Kim HR, Kon S, et al. Functional role of death-associated protein 3 (DAP3) in anoikis. *J Biol Chem* 2004 Oct 22;279(43):44667–72. [PubMed: 15302871]
31. Nayak SK, O'Toole C, Price ZH. A cell line from an anaplastic transitional cell carcinoma of human urinary bladder. *Br J Cancer* 1977 Feb;35(2):142–51. [PubMed: 836756]
32. Dohr S, Klingenhoff A, Maier H, Hrabe de AM, Werner T, Schneider R. Linking disease-associated genes to regulatory networks via promoter organization. *Nucleic Acids Res* 2005;33(3):864–72. [PubMed: 15701758]
33. Chan MW, Chan LW, Tang NL, Tong JH, Lo KW, Lee TL, et al. Hypermethylation of multiple genes in tumor tissues and voided urine in urinary bladder cancer patients 3. *Clin Cancer Res* 2002 Feb;8(2):464–70. [PubMed: 11839665]
34. Kim D, LaQuaglia MP, Yang SY. A cDNA encoding a putative 37 kDa leucine-rich repeat (LRR) protein, p37NB, isolated from S-type neuroblastoma cell has a differential tissue distribution. *Biochim Biophys Acta* 1996 Dec 11;1309(3):183–8. [PubMed: 8982252]

Bootstrap Study of Random Correlations

Fig. 1.
Bootstrap distribution of correlations with 95% CI denoted.

Fig. 2. Reproducibility of replicate analyses after normalization and filtering of genes expressed at < 3 sd above background. Data shown are for one pair of replicates and are generally representative.

Fig. 3.
Correlations among expression levels of the 21 genes identified by the template method.

Fig. 4. Regulatory sequence frequency analysis. Sequences that are over-represented significantly are shown in red. Those that are present but not significantly over-expressed or under-expressed are shown in grey. Transcripts that associated positively with the template are shown in green on the right whereas those shown in red were negatively associated with the template. Note the high representation of highly over-represented regulatory sequences among the gene transcripts whose abundance is positively correlated with the template model. No regulatory sequences were identified for IMPD2.

Fig. 5.

Interconnections among expression levels through common upstream regulatory sequences in template-identified genes. Bright Green = positively correlated with Optimal Template Model, Light Green = positively correlated with either Linear or Super-Linear Models. Genes that were correlated with the Optimal Model and one or both of the *a priori* models are presented in bright green. Red = Negatively correlated with any model. Note that the genes whose expression correlated most closely with the Optimal Model expressed upstream sequences for Nkx2-5, Oct-1 and Pax4 (bright yellow). In contrast, SOX9, FOXJ2, FOXD3, and cMYB tended to be associated with genes that fit the *a priori* models.

Table 1

Derivation of an optimized model describing the interaction of the effects of extracellular matrix/culture conditions and inherent malignancy. Two seed models were tested, a linear model (L) in which the values were additive and a superlinear model (SL) in which the values were multiplied after substituting “1” for the “0” values for all but the SV-HUC cells grown on plastic. Both seed models converged on the same optimized model (O). Values listed in L and SL rows are the expected relative expression levels of genes that fit the template according to the linear and superlinear models respectively. Values listed in the O rows are the actual values for the best fit model..

Extracellular Matrix-Growth Phenotype		SV-HUC-control cells: 0	RT4 Low-grade, papillary tumours: 1	253 J 253 JBV Low-grade, papillary tumours: 1	J82 High-grade, invasive tumours: 2	TCCSUP High-grade, invasive tumours: 2
Plastic (neutral, but no 3-D growth): 0	L	0	1	1	1	2
	SL	0	1	1	1	2
	O	0.1037	1.3134	0.5753	0.5754	1.4604
SISgel (suppresses invasion, but allows 3-D growth): 1	L	1	2	2	2	3
	SL	1	2	2	2	4
	O	0.8206	1.6001	3.28472	3.1161	5.0108
Matrigel (fully permissive of malignant phenotype): 2	L	1	3	3	3	4
	SL	2	4	4	4	8
	O	2.0757	4.0856	3.1400	3.3085	7.5288
						8.0791

Correlation of transcripts matched to the different templates (Table 1) with correlation coefficient, $|r| > 0.6$. The Pearson's Correlation Coefficient (ρ), can take on the values from -1.0 to 1.0 , where -1.0 is a perfect negative (inverse) correlation, 0.0 is no correlation, and 1.0 is a perfect positive correlation. The absolute values of ρ were utilized for analysis so that correlated and anti-correlated relationships between the predicted and experimental values could be identified. The K-Means Cluster columns show that the template clustering shows no relation to K-means clustering. Clustering with other numbers of clusters did not effect any improvement.

Table 2

Acc. No	Gene name	Functional group	Localization	L	Pearson's ρ SL	O	K-Means Cluster K=4	K-Means Cluster K=5
X02492	GIP3; interferon- α -inducible protein (IFI-6-16)	Functionally unclassified	Unclassified	0.703	0.838		4	5
M29366	ERBB3; proto-oncogene; HER3; EGF receptor B3	Oncogenes and tumor suppressors; growth factor & chemokine receptors	plasma membrane		0.792		3	3
J05032	DARS; aspartyl-tRNA synthetase (ASPRS);	Other proteins involved in translation	cytoplasm	0.781			4	5
U12431	EALVL2; HEL-N1	mRNA processing/turnover/transport	cytoplasm	0.776			4	5
M57627	IL10; interleukin 10	Interleukins and interferons	extracellular secreted	0.725	0.707	0.719	2	3
M11886	HLA-C; major histocompatibility complex class I C	Major histocompatibility complex	plasma membrane	0.611	0.669	0.71	3	3
X02152	LDHA; L-lactate dehydrogenase M, A-subunit	Simple carbohydrate metabolism	cytoplasm	0.676			2	2
U60519	CASP10; caspase 10	Caspase apoptosis protein	cytoplasm	0.628			1	1
K02770	IL1B; interleukin 1 beta; catabolin	Interleukins and interferons	extracellular secreted	0.669		0.696	3	3
U14971	RPS9; 40S ribosomal protein S9	Ribosomal proteins	cytoplasm			0.639	3	3
S40706	DDIT3; GADD153	DNA repair	nuclear		0.628		3	4
U13699	CASP1; caspase 1; IL-1-beta converting enzyme (ICE)	Caspase apoptosis protein	cytoplasm	0.667		0.624	3	5
U09579	CDKN1A; cyclin-dependent kinase inhibitor 1A; CIP1; WAF1	CDK inhibitors; kinase activators and inhibitors; oncogenes and tumor suppressors	nucleus			0.605	3	3
U32907	LRRC17; 37 kDa leucine-rich repeat protein; P37NB	functionally unclassified	unclassified			-0.59	3	4
U18321	DAP3; ionizing radiation resistance-conferring protein + death-associated protein 3	other apoptosis-associated proteins	unclassified			-0.6	4	5
M15353	EIF4E; eukaryotic translation initiation factor 4E 25-kDa subunit	translation factor	cytoplasm			-0.6	3	4
X95648	EIF2B1; translation initiation factor	Translation factor	cytoplasm			-0.61	4	5
L33842	IMP2; inosine-5'-monophosphate dehydrogenase 2	Nucleotide metabolism	cytoplasm			-0.63	4	5
L22253	SFRS7; arginine/serine-rich splicing factor 7	mRNA processing/turnover/transport	cytoplasm		-0.64		2	2
D66904	HRMT1L2	Methyl transferase, binds p53	cytoplasm, nucleus	-0.681	-0.67		4	5
X97442	TMP21 transmembrane protein 21	Vesicular trafficking	Golgi complex			-0.74	2	2