Differential Coexpression Analysis For The Identification Of Dysregulated Genes In Prostate Cancer

Obi L Griffith and Steven JM Jones

British Columbia Cancer Agency
Vancouver, British Columbia, Canada

Canada's Michael Smith Genome Sciences Centre
www.bcgsc.ca

1. Abstract

Objective & Design
DNA regulatory sequences determine the level, location and timing of gene expression. In some cases, the onset of cancer is related to changes in these regulatory sequences. In a typical example a gene (such as bcl2) is translocated from one chromosome to another. In its new position the gene is under the wrong regulation and produces a protein that prevents cell death and ultimately leads to cancer. Numerous studies have used the correlation of gene expression (coexpression) to identify potentially coregulated genes. However, few have looked to see if loss of coexpression (differential coexpression) can identify potentially deregulated genes. We present an approach that calculates the change in coexpression for each individual gene and ranks them according to which is most likely to be deregulated.

Materials & Methods
A published prostate cancer dataset was chosen with 52 tumour and 50 normal samples analyzed on the Affymetrix HG-U95Av2 chip [2]. We consider each normal microarray sample as a 'snapshot' of normal expression. A Pearson correlation coefficient (PCC) is calculated between every possible gene pair. Gene pairs with similar expression levels across the 50 normal samples will have a PCC ≈ 1 and are said to be 'coexpressed'. Generally, all genes have at least some highly coexpressed gene pairs. For each gene in normal data, the n closest gene neighbours (most coexpressed) are defined and the mean PCC calculated. Then, the mean PCC for these same gene neighbours is determined for the tumour dataset. A differentially coexpressed gene is defined as a gene with a large change in coexpression (difference in mean PCC for n nearest neighbours) from normal to cancer. A random permutation test was performed to assess the significance of the differentially coexpressed genes observed.

Results
Of the top 25 most differentially coexpressed microarray probes, 2 could not be mapped, 5 map to genes of unknown function (LOC153561, MGC5576, DNAJA2, C7orf24, SNX4), 6 to genes previously implicated in cancer (SLC26A3, G1P2, RAD23B, TFDP2, ABL1, GSPT1), 5 to genes implicated specifically in prostate cancer (AMACR, SOX9, DLGAP2, RAD51C, HGF) and 2 to genes that interact with PSA and play a role in sexual reproduction (SEMG1, SEMG2). In most cases, these genes would not have been identified by standard differential expression methods.

Conclusion
Differential coexpression analysis represents a useful and complementary method to traditional differential expression methods for identifying potentially relevant cancer genes. These genes may represent novel prostate cancer genes and would make good candidates for regulatory mutation analysis.

2. Background

Figure 1. Differential expression versus differential coexpression

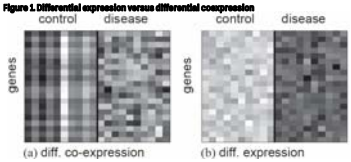

Figure 2. Differential coexpression

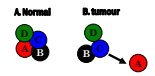

Fig 1: Standard microarray heat maps illustrate the difference between diff. expression and diff. coexpression: **(a)** the average expression level does not change but a consistent pattern of expression (coexpression) is lost. **(b)** No change in expression pattern is apparent but the average expression level changes.

Fig 2: Consider four genes in 'coexpression space'. **(A)** In normal tissue, genes A,B,C and D are tightly coexpressed. **(B)** In tumour tissue, gene A is no longer coexpressed with B,C and D. We say that gene A is differentially coexpressed.

3. Method

Figure 3. Differential coexpression method

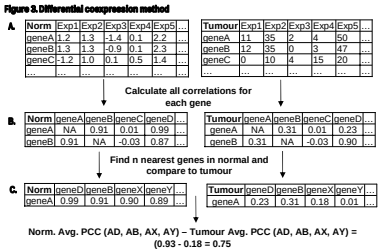

Fig 3: (A) An R program[3] was created to read in expression files (genes x exp) for normal and tumour datasets. **(B)** For each gene, all gene-gene Pearson correlations were calculated. **(C)** The correlations were sorted and the n highest correlations (most coexpressed) determined. **(D)** The difference in mean correlation for the gene is determined for neighbourhood size n (in this case the difference is for n=4 where gene A's nearest four neighbours are D, B, X and Y). The level of correlation between these genes is much lower in the tumour tissue and thus this gene is said to be 'differentially coexpressed'. Finally, genes are ranked by the magnitude of differential coexpression.

4. Differential coexpression in prostate cancer

Figure 4. Distribution of mean Pearson for all genes

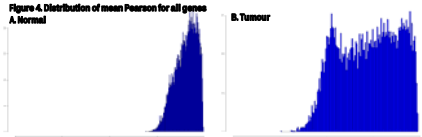

Figure 5. Difference in Mean Pearson for all genes (neighbourhood = 10)

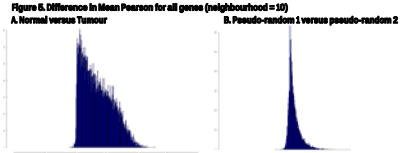

Fig 4: (A) In the normal dataset, most genes have an average correlation of 0.5 to 1 for their 10 most correlated genes (n=10). **(B)** If we consider the same gene sets in tumour tissue, some genes retain a high level of coexpression while many others lose their coexpression. Note: x-axis scale from -1 to 1.

Fig 5: (A) The distribution of differences in PCC shows that many genes change only marginally (their nearest neighbours from normal remain highly coexpressed in tumour tissue) while others change a great deal. (B) Artificial datasets generated by random sampling of the data and analyzed in the same fashion show much less extreme changes in coexpression. Note: x-axis scale from -1 to 2.

5. Effect of neighbourhood size

Figure 6. Effect of neighbourhood size

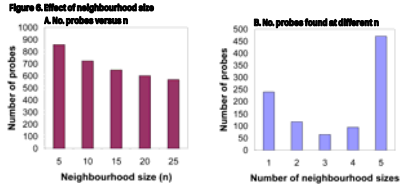

Fig 6: (A) As n increases, less genes meet the criteria for diff. coexpression. **(B)** Most genes meet the criteria for diff. coexpression at multiple values of n.

6. An example of differential coexpression (AMACR)

Figure 7. Expression of AMACR compared to 10 nearest neighbours

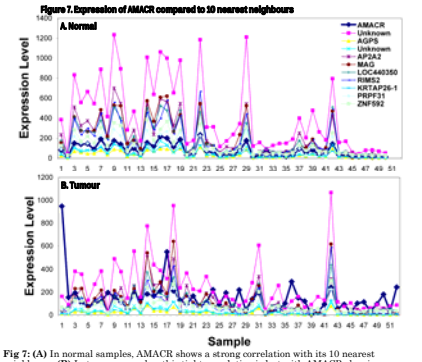

Fig 7: (A) In normal samples, AMACR shows a strong correlation with its 10 nearest neighbours. **(B)** In tumour samples, this tight correlation is lost with AMACR showing a more random pattern of expression compared to the other 10 genes.

7. Differentially coexpressed genes in prostate cancer

Table 1. Partial list of most differentially coexpressed genes in prostate cancer

Symbol	Comments
SLC26A3	Protein downregulated in colon adenoma and adenocarcinoma
CELSR1	Developmentally regulated, neural-specific gene which plays an unspecified role in early embryogenesis
AMACR	**Proven biomarker that can help distinguish cancer from benign cells, with high sensitivity and specificity for prostate carcinoma.**
PEX5	peroxisomal biogenesis factor
G1P2	Induced by camptothecin and Retinoic Acid in human tumour cells
SOX9	Overexpression results in suppression of growth and tumourigenicity in the prostate tumour cell line M12
ATP6V1E1	ATPase
LOC153561	function unknown
SEMG1	Interacts with PSA
MGC5576	function unknown
RAD51C	mRNA and protein levels elevated ~2- to 5-fold in malignant prostate cell lines
RAD23B	Highly expressed in the human testis and in ejaculated spermatozoa. Down-regulated during early apoptosis in human hepatoma cells exposed to Paeoniae
SEMG2	Interacts with PSA
SNX4	not well characterized (only 4 pubmed)
DLGAP2	putative tumour suppressor gene. Chromosomal region (8p23.2) frequently deleted in prostate cancer.
TFDP2	Differential expression shown in some cancer cell lines
HGF	c-Met/HGF receptor have roles in prostate neoplasm progression
PEX10	peroxisomal biogenesis factor
ABL1	Important in leukemia - Bcr-Abl translocation
GSPT1	Overexpressed in gastric cancer
DNAJA2	function unknown
C7orf24	function unknown
GRM5	glutamate receptor, metabotropic

Cancer
Prostate Cancer

Figure 8. Statistically over-represented GO terms

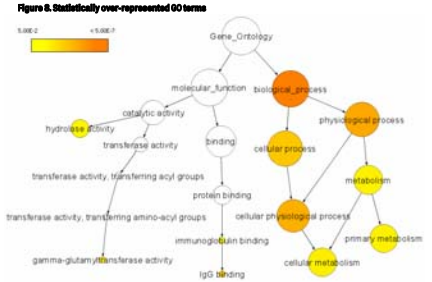

Table 1: An examination of the 25 most differentially coexpressed probes in prostate tumour identified 2 that could not be mapped to genes, 5 map to genes of unknown function, 6 to genes previously implicated in cancer, 5 to genes implicated specifically in prostate cancer and 2 to genes that interact with PSA and play a role in sexual reproduction.

Fig 8: A Gene Ontology analysis using BINGO[4]/Cytoscape[5] of all genes that meet our criteria for differential expression reveals significant statistical over-representation of several categories from both the biological process and molecular function ontologies (corrected p-value is indicated by color as per legend).

8. Conclusions and Future work

Conclusions:
>Differential coexpression analysis represents a useful and complementary method to traditional differential expression methods for identifying potentially relevant cancer genes.
>Such genes may represent novel prostate cancer genes and would make good candidates for regulatory mutation analysis.

Future Work:
>Develop statistical methods further.
>Investigate possible mechanisms for deregulation
>>Known translocations in region of gene
>>Mutations/SNPs in regulatory elements
>Apply method to additional prostate cancer datasets and other cancer types

9. Acknowledgments

funding | Natural Sciences and Engineering Council of Canada (OG); Michael Smith Foundation for Health Research (OG and SJ); Canadian Institutes of Health Research (OG); BC Cancer Foundation.

references | [1] Kostka and Spang. 2004; [2] Singh et al. 2002; [3] Ihaka and Gentleman. 1996, http://www.r-project.org/; [4] Maere et al. 2005, http://www.psb.ugent.be/cbd/papers/BiNGO/; [5] Shannon et al. 2003, http://www.cytoscape.org/