

Michael G. Walker
Wayne Volkmuth

Walker Bioscience,
Sunnyvale, and
Incyte Genomics Inc.,
Palo Alto, CA, USA

Cell adhesion and matrix remodeling genes identified by co-expression analysis

Cell adhesion and matrix remodeling are elements in many diseases, ranging from atherosclerosis and fibrosis to metastatic cancer. However, many genes that participate in these processes have not yet been identified. To find such genes, we looked for previously uncharacterized genes that are co-expressed with known cell adhesion and matrix remodeling genes. The known genes in this study included MMP2, TIMP3, BM-40, chondroitin, connective tissue growth factor, fibromodulin, IGFBP5, laminin, MGP, myosin light chain kinase, several collagens, and other matrix and adhesion proteins. We found eight previously uncharacterized genes, here named MXRA1 through MXRA8, that were strongly co-expressed with these known adhesion and matrix genes. Five of the MXRA genes have a significant similarity to uncharacterized cDNA sequences or predicted proteins listed in the Genbank database, but otherwise show distant or no sequence similarity to genes with known function. Subsequent to our entry of the MXRA gene sequences in the Genbank, three of the eight genes have been independently described by other researchers: MXRA2 is α -parvin, a cell-matrix adhesion protein, MXRA4 is a C1 complement component receptor involved in cell adhesion, and MXRA5 is adican, an adhesion proteoglycan. The analysis described here provides further evidence for the role of these genes in adhesion and matrix remodeling.

Keywords: Co-expression analysis, cell adhesion, matrix

1 Introduction

Cell adhesion and matrix remodeling are key elements of many disease processes, including metastatic cancer, cardiomyopathy, arthritis, angiogenesis, diabetic necrosis, atherosclerosis, fibrosis, and ulceration. While many genes that participate in or regulate these processes are known, many remain to be identified. Identification of currently uncharacterized genes will provide new diagnostic and therapeutic targets and may provide new tools for therapeutic tissue engineering.

2 Materials and Methods

2.1 cDNA library preparation

We examined gene expression in 1375 human cDNA libraries. These libraries were prepared from human tissue samples taken from surgery, biopsy, or cell lines from diverse anatomic and pathologic states. Approximately 5000 cDNA's from each library were sequenced by gel electrophoresis, assembled, and aligned against known genes. All genes detected in at least five of the 1375 libraries were included in the analysis.

Correspondence: Michael G. Walker, Walker Bioscience, 1475 Flamingo Way, Sunnyvale, CA, USA 94807-3405. Phone: +1-408-736-5506, Fax: +1-650-556-1132, e-mail: mwalker@stanfordalumni.org

Table 1. Co-expression of two hypothetical genes A and B as counts of the number of libraries in which the genes are present or absent.

Number of libraries	Gene A present	Gene A absent	Total
Gene B present	8	2	10
Gene B absent	2	18	20
Total	10	20	30

2.2 Computational analysis

For the purpose of this analysis, we encode each gene as being either present or absent in a cDNA library. If mRNA from a gene was detected in the sample, we encode it as a "1". If mRNA from a gene was not detected in the sample, we encode it as a "0". Table 1 shows an example of the occurrences of two 2 hypothetical genes, A and B, in 30 cDNA libraries using this encoding. We use a Fisher Exact test [1] to determine the probability that the co-expression of two genes, such as that in Table 1, occurs by chance. For the data in Table 1, the probability that the two genes are co-expressed by chance is $p = 0.0003$. In subsequent tables, we transform p -values into the negative log of the p -value. Thus, $p = 0.0003$ would be represented as $-\log(0.0003) = 3.5$. An approximation to the Fisher Exact test, the chi-square test, is commonly used to analyze such contingency tables. The chi-squared approximation is only accurate

Table 2. Co-expression of the known genes. Entries in this table are the probability that the co-expression of each pair of genes is due to chance (expressed as the negative log of the *p*-value from the Fisher Exact test).

	α -2mac	α -act	BM-40	Clr	calponin	caveolin 1	chondroitin	coll III	coll I	coll VI a 1	coll VI a 3	ctgf	desmin	fibromodulin	fibulin-1D	filamin	gelsolin	IGFBP5	LTBP4	lumican	MGP	MLCK	MMP2	SM22	SMMHC	TIMP3
α 2m	24	41	52	34	22	54	46	41	32	30	26	42	26	43	28	42	34	15	33	42	50	24	52	37	40	
act	24		28	25	28	24		21	22	33	25	25	25	19	27	48	35	22	21	18	22	25	26	44	30	31
BM40	41	28		38	22	40	68	53	61	35	39	41	25	22	41	27	30	40	14	47	46	33	45	54	16	53
Clr	52	25	38		23	20	55	46	46	46	42	27	33	31	49	35	30	31	16	33	30	29	37	45	30	32
cal	34	28	22	23		18	37	25	30	32	27	14	50	22	41	37	36	28	31	19	24	61	22	50	70	28
cav	22	24	40	20	18		38	27	37	24	31	45	25	21	30	21	27	32	15	26	25	24	36	31	17	35
cspg	54		68	55	37	38		76	71	40	53	34	35	40	54	26	28	42	16	69	48	42	44	62	26	59
cIII	46	21	53	46	25	27	76		91	57	69	31	23	35	47	31	22	30	13	57	22	34	56	41	21	31
cI	41	22	61	46	30	37	71	91		52	61	33	25	30	54	37	27	45	20	57	27	33	55	56	24	42
cVIa1	32	33	35	46	32	24	40	57	52		49	22	36	29	46	50	34	34	30	41	19	29	47	47	32	31
cVIa3	30	25	39	42	27	31	53	69	61	49		42	25	35	44	34	28	36	17	52	17	39	56	34	24	44
ctgf	26	25	41	27	14	45	34	31	33	22	42		22	25	30	22	34	27		29	34	21	34	24		40
des	42	25	25	33	50	25	35	23	25	36	25	22		19	33	42	39	32	29	18	30	30	24	47	42	27
fibro	26	19	22	31	22	21	40	35	30	29	35	25	19		34	25	26	23	21	28	24	27	26	28	27	31
fibu	43	27	41	49	41	30	54	47	54	46	44	30	33	34		38	42	40	21	38	27	39	51	48	31	47
fila	28	48	27	35	37	21	26	31	37	50	34	22	42	25	38		35	26	36	19		33	36	41	32	28
gels	42	35	30	30	36	27	28	22	27	34	28	34	39	26	42	35		34	23	24	33	27	25	48	35	37
igfbp	34	22	40	31	28	32	42	30	45	34	36	27	32	23	40	26	34		21	31	22	31	29	40	26	35
Ltbp4	15	21	14	16	31	15	16	13	20	30	17		29	21	21	36	23	21			16	24	16	26	36	21
lumic	33	18	47	33	19	26	69	57	57	41	52	29	18	28	38	19	24	31			26	27	41	31	17	40
MGP	42	22	46	30	24	25	48	22	27	19	17	34	30	24	27		33	22	16	26		21	15	49	19	27
mlck	50	25	33	29	61	24	42	34	33	29	39	21	30	27	39	33	27	31	24	27	21		26	54	58	34
mmp	24	26	45	37	22	36	44	56	55	47	56	34	24	26	51	36	25	29	16	41	15	26		39	19	35
SM22	52	44	54	45	50	31	62	41	56	47	34	24	47	28	48	41	48	40	26	31	49	54	39		42	36
smhc	37	30	16	30	70	17	26	21	24	32	24		42	27	31	32	35	26	36	17	19	58	19	42		15
timp3	40	31	53	32	28	35	59	31	42	31	44	40	27	31	47	28	37	35	21	40	27	34	35	36	15	

when the expected number of counts in any cell in the table is sufficiently large (greater than 10). For many datasets (including ours) the expected number in at least one cell is often less than 10, so we prefer the more accurate Fisher Exact test.

Because we examine all pairwise relationships among genes, rather than specifying a hypothesis for association between a specific pair of genes, we risk finding spurious associations simply by performing so many comparisons. One method to control for this risk is to divide the *p*-value specified for significance (say, $p = 0.05$) by the number of comparisons done for each gene. Usually, we only regard pairs of genes with co-expression *p*-values less than $1.0e-10$ to be worth consideration, and then only if we observe

multiple co-expressed genes with a known common biological function.

2.3 Accession numbers

Gene name	Genbank accession number
MXRA1	AW888226
MXRA2	AW888219
MXRA3	AW888220
MXRA4	AW888224
MXRA5	AW888221
MXRA6	AW888222
MXRA7	AW888225
MXRA8	AW888223

Table 3. Co-expression of the known genes with the MXRA genes. Entries in this table are the probability that the co-expression of each pair of genes is due to chance (expressed as the negative log of the p-value from the Fisher Exact test).

Gene name	MXRA1	MXRA2	MXRA3	MXRA4	MXRA5	MXRA6	MXRA7	MXRA8
<i>α</i> -2macroglobulin	22	15	28	18	7	35	15	11
<i>α</i> -actinin	22	22	16			24	17	15
BM-40	17	22	21	18	13	27	31	13
C1r	18	23	17	15	8	29		20
calponin	19	21	28			28	20	12
caveolin 1	16	15	18	23	7	18	23	11
chondroitin	19	21	33	19	12	26	23	17
coll III	14	18	23		18	19		22
coll I	18	21	25	15	18	27	16	18
coll VI <i>α</i> 1	16	24	27		11	33	14	28
coll VI <i>α</i> 3	21	25	22	18	19	22		23
ctgf	16	14	17	19	6	17	25	14
desmin	22	20	26	11		31		17
fibromodulin	14	19	16	22	10	18	14	14
fibulin-1D	15	20	25	12	6	27	18	20
filamin	20	21	21	14	7	32		29
gelsolin	18	19	20	20	6	31	22	21
IGFBP5	23	25	25	13	6	24	17	18
LTBP-4	13	13	17	15		20	14	28
lumican	12	25	18	13	13		14	17
MGP	17		16	21		16	28	
MLCK	20	22	21	13	6	33	16	12
MMP2	16	19	26	13	11	18		20
SM22	14	18	28	15	6	32	22	17
SMMHC	15	16	26			33		15
TIMP-3	26	22	26	20	10	38	28	16

3 Results and Discussion

We observed strong co-expression among a set of known matrix remodeling and cell adhesion genes that included *α*-2-macroglobulin, *α*-actinin, BM-40 (Sparc), C1r, calponin, caveolin 1, chondroitin, collagen III, collagen I, collagen VI alpha 1, collagen VI *α* 3, connective tissue growth factor (CTGF), desmin, fibromodulin, fibulin-1D, filamin, gelsolin, Insulin-like growth factor binding protein 5 (IGFBP5), latent transforming growth factor- β binding protein 4 (LTBP-4), lumican, matrix Gla protein (MGP), myosin light chain kinase (MLCK), matrix metalloprotease 2 (MMP2), SM22, smooth muscle myosin heavy chain (SMMHC), and TIMP-3. Table 2 shows the co-expression of these genes with each other. Other matrix and cell adhesion genes, such as thrombospondin, fibrillin, and urokinase-type plasminogen activator, were also co-expressed with these genes, though not as closely.

We observed eight previously uncharacterized genes (here named MXRA1 through MXRA8), that are the most closely co-expressed with these known genes. The co-expression of the eight MXRA genes with the known genes is shown in Table 3. Each of the MXRA genes is co-expressed with at least ten of the known genes with a *p*-value of less than 1.0e-10, and has *p*-values for association comparable to those among the known genes. In this database, genes that

encode, for example, ribosomal proteins, neurotransmitters or other proteins that have no relation to adhesion or remodeling, have *p*-values for co-expression with the adhesion genes of 1.0e-3 or greater.

Five of the MXRA genes have significant similarity to uncharacterized cDNA sequences or open reading frames in Genbank, but otherwise show distant or no sequence similarity to genes with known function. After our entry of the MXRA gene sequences in Genbank, three of the eight genes have been independently described by other researchers. MXRA2 is *α*-parvin, a cell-matrix adhesion protein that co-localizes with actin filaments at membrane ruffles and focal contacts in fibroblasts [2]. MXRA4 is a C1q complement component receptor (C1QR, Genbank accession number XM_012949); C1q stimulates endothelial expression of cell adhesion molecules and promotes cell attachment, among other functions [3, 4]. MXRA5 is adlican, an adhesion proteoglycan that shows elevated expression in cartilage from patients with osteoarthritis (Genbank accession number AF245505). The analysis described here provides further evidence for the role of these genes in adhesion and matrix remodeling.

Calculation of the *p*-value for association of pairs of genes makes several assumptions that are affected by the methods of library selection and preparation. The calculation as-

sumes that libraries are independent, but this assumption is violated because more than one library may be obtained from a single patient (for example, multiple organs, or matched tumor and non-tumor tissue, or normalized and non-normalized). Library normalization or subtraction will enrich the proportion of genes expressed at low levels; this alteration makes the detection of associations between genes expressed at different levels more difficult. Repeated sequencing of the same tissue from different individuals would tend to increase the risk of spurious associations, if two genes occurred predominantly in that tissue but were otherwise unrelated. In this situation, we perform co-expression analysis on a set of libraries restricted to that tissue type. Libraries with abundant novel genes were in some cases sampled to greater depth, leading to an inconsistent measure of presence or absence. Effects of sampling, temporal or tissue differences in expression, and errors in sequence assembly or annotation may obscure associations. We expect that these effects are most likely to obscure a true relationship. It is unlikely that a spurious relationship would be introduced consistently across such a large set of libraries by chance.

The expression patterns of the MXRA genes indicate that they participate in cell adhesion and matrix remodeling. Bet-

ter understanding of these genes may help us alleviate the effects of diseases involving these processes.

Acknowledgements

We gratefully acknowledge the assistance and support of our colleagues at Incyte Genomics.

References

- [1] Agresti, A. (1990) *Categorical data analysis*. Wiley, New York, NY.
- [2] Olski, T. M., Noegel, A. A., Korenbaum, E. (2001) Parvin, a 42 kDa focal adhesion protein, related to the alpha-actinin superfamily. *J Cell Sci* **114**: 525-38.
- [3] Lozada, C., Levin, R. I., Huie, M., Hirschhorn, R., Naime, D., Whitlow, M., Recht, P. A., Golden, B., Cronstein, B. N. (1995) Identification of C1q as the heat-labile serum cofactor required for immune complexes to stimulate endothelial expression of the adhesion molecules E-selectin and intercellular and vascular cell adhesion molecules 1. *Proc Natl Acad Sci USA* **92**: 8378-82.
- [4] Luddington, S., Qwarnstrom, E. E., Page, R. C., Bordin, S. (1993) Expression and function of gingival fibroblast C1q receptors are upregulated by interleukin-1 beta and transforming growth factor-beta. *J Cell Physiol* **155**: 157-63.

Printed on acid-free paper

© 2002 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim. Printed in the Federal Republic of Germany.

Editor-in-Chief: W. Doerfler, Inst. of Genetics, 50931 Köln, Germany.

Publisher: WILEY-VCH Verlag GmbH & Co. KGaA, Boschstraße 12, 69469 Weinheim, Federal Republic of Germany.

Correspondence concerning advertisements should be addressed to WILEY-VCH Verlag GmbH & Co. KGaA, Boschstraße 12, 69469 Weinheim, Telefax +49-06201/606-550, e-mail: adsales@wiley-vch.de.

Institutional subscription rates 2003 print only or online only/print & online: Europe € 338,00/355,00; Switzerland sFr 568,00/597,00; Outside Europe US \$ 368,00/387,00 incl. postage and handling.

Reduced rates for personal subscribers are available on request. Prices are subject to change.

Rate of publication: 6 times a year.

Cancellation of subscriptions: The publisher must be notified not later than three months before the end of the calendar year.

Printed by betz-druck, Darmstadt.

All rights reserved (including those of translation into other languages). No part of this journal may be reproduced in any form – by photoprint, microfilm, or any other means – nor transmitted or translated into a machine language without the permission in writing of the publisher. – Only single copies of contributions, or parts thereof, may be reproduced for personal use. – This journal was carefully

produced in all its parts. Nevertheless, authors, editors and publisher do not warrant the information contained therein to be free of errors. Readers are advised to keep in mind that statements, data, illustrations, procedural details or other items may inadvertently be inaccurate.

Valid for users in the USA: The appearance of the code at the bottom of the first page of an article in this journal (serial) indicates the copyright owner's consent that copies of the article may be made for personal or internal use, or for the personal or internal use of specific clients. The consent is given on the condition, however, that the copier pay the stated percopy fee through the Copyright Clearance Center, Inc., for copying beyond that permitted by Section 107 or 108 of the U.S. Copyright Law. This consent does not extend to other kinds of copying, such as copying for general distribution, for advertising or promotional purposes, for creating new collective works, or for resale. For copying from back volumes of this journal see Permissions to Photo-Copy: Publisher's Fee List of the CCC.

