

RESEARCH ARTICLE

Ubiquitous cancer genes: Multipurpose molecules for protein micro-arrays

Brigitte Altenberg¹, Christine Gemuend¹ and Karl Otto Greulich²

¹ European Molecular Biology Laboratory, Bioinformatics Group, Heidelberg, Germany

² Department of Single Cell and Single Molecule Techniques, Leibniz Institute of Age Research/Fritz Lipmann Institute, Jena, Germany

Multipurpose genes in the human genome which are over-expressed in a large variety of different cancers have been identified. Forty-two of the 19,016 human genes annotated to date (0.2%) are ubiquitously over-expressed in half or more of the 36 investigated human cancers. Of these genes, 15 are involved in protein biosynthesis and folding, six of them in glycolysis. A group of 13 solid tumours over-express almost all (39–42 of 42) ubiquitous cancer genes, suggesting a common mechanism underlying these cancers. Others, such as endocrine cancers, have only a few over-expressed ubiquitous cancer genes. The proteins for which these genes code or the corresponding antibodies are candidates for small protein microarrays aiming at maximum information with only a limited number of proteins. Since the over-expression pattern varies from cancer to cancer, distinction between different cancer classes is possible using one single set of protein or antibody molecules.

Received: March 14, 2005

Accepted: May 26, 2005

Keywords:

Cancer genes / Data mining / Human genome / Protein arrays

1 Introduction

Multipurpose DNA microarrays often carry more than 10 000 different molecules, of which the user typically exploits only a limited subclass. For protein microarrays such an approach would be wasteful, since attaching proteins or antibodies to chip surfaces is much more complex than attaching cDNA or oligonucleotides. In particular, protein arrays for marker-free readout using intrinsic fluorescence [1] require optimal orientation of the protein molecules on the surface, *i.e.* presently, the size of such chips is limited to a few molecule types. For the assembly of multipurpose protein arrays a strategy using multipurpose proteins is promising: protein microarrays for cancer research would benefit

from such a strategy, if proteins could be found which are overexpressed in a large variety of cancers, ideally in different combination.

A recent *in silico* investigation identified 291 genes in the human genome which are modified in cancer by mutation [2]. Changes in gene expression have been studied by analysis of 1975 published microarrays spanning 22 tumour types in order to find common points and variation of gene expression between different types of tumours [3]. In such gene studies, gene expression is used as a surrogate for transcriptional regulation [4]. Genes on these 1975 arrays have been assigned to different “modules” functional for a number of cancers. So far, no systematic analysis on all human genes in essentially all human cancers is available.

Studies on the over-expression of all (approx. 20 000) human genes in all (approx. 50) human cancer classes would require around one million data points, not including a certain degree of redundancy, to obtain information on their statistical quality. Corresponding microarrays are still expensive. However, at present, an *in silico* analysis using

Correspondence: Brigitte Altenberg, European Molecular Biology Laboratory, Bioinformatics Group, Meyerhofstr. 1, 69120 Heidelberg, Germany

E-mail: altenberg@embl.de

Fax: +49-6221-387517

microarray data published in the literature to date can give first insights into these ubiquitous cancer genes. Such a meta-analysis does not only give access to a huge set of data, it also averages out flaws and errors in single cDNA or oligonucleotide array studies, which are still error prone or incompatible with each other [5]. Also, the messages from SAGE data and EST data agree by only around 50% [6].

In the present work we used the dbEST database provided by the NIH. It collects microarray data from the literature and makes available such pooled data in a statistically pre-evaluated form. We checked all the annotated human genes to date for their over-expression in 36 different human cancers.

2 Materials and methods

2.1 The dbEST database and its use

Expression data of single genes in the dbEST database provided on the NIH cGAP page ("Virtual Northern" function) can be obtained *via* <http://cgap.nci.nih.gov/Genes/GeneFinder>. The search routine for checking all annotated human genes can be accessed *via* <http://www.embl-heidelberg.de/~altenber/gemuend/Program>.

The database used in the present investigation is dbEST provided by the NIH. It is part of the gene info page from the NCI Cancer Genome Anatomy Project (CGAP) [7–12]. This page allows searches in databases that contain cDNA and EST data [7, 12] and has access to approximately four million ESTs and genes. The database collects gene expression data from other libraries and unifies them. In one of its sub-functions, "Virtual Northern", gene expression data for 51 normal tissues and their corresponding cancerous counterparts are summarised. For specific genes, expression in these tissues is provided in a statistically pre-evaluated manner. Table 1 gives an example for the eukaryotic translation elongation factor 1 (EEF1A1). We have selected this gene because data are available for most tissues and cancers.

The rows in Table 1 contain data for the different cancers. The left column lists the cancer class. Columns two and three represent the number of hybridisation signals divided by the number of data points available in total for normal tissue and for the corresponding cancer tissue. The right-most column gives the *p* value for the ratio of cancer and normal tissue, a measure of the statistical quality of the data (see also below). In order to understand the data provided in this table, one may imagine a virtual chip carrying 2 154 515 test sequences in all 51 tissues (rows) listed in the table. In 10 645 cases, hybridisation of the EEF1A1 sequences against this chip was observed, *i.e.* the expression level was $10\,645/2\,154\,515 = 0.0049$. The result for corresponding cancer tissues was $17\,929/1\,974\,339 = 0.0091$, *i.e.* averaged over all tissues, this gene was over-expressed in cancer. In various publications [8–11] it has been shown that this is an adequate strategy for quantifying gene expression. For any

Table 1. EST data for EEF1A1 (Hs 439552) given by the "Virtual Northern" function (for details, see text)

Tissue	Normal	Cancer	<i>p</i> value
All tissues	10645/2154515	17929/1974339	
Adipose	106/9952	1/722	0,01
Adrenal cortex		53/6221	
Adrenal medulla		0/297	
Bone	44/7108	329/40512	0,04
Bone marrow	84/14454	212/20961	0
Brain	1281/209529	592/151390	0
Cartilage	86/10470	166/35784	0
Cerebellum	16/4309	0/0	
Cerebrum	211/67392		
Cervix	0/1020	726/39560	0
Colon	41/17386	940/144087	0
Ear	64/11926		
Embryonic tissue			
Endocrine	18/6739	17/2876	0,02
Oesophagus	0/84	8/2591	0,42
Eye	299/74541	755/43968	0
Gastrointestinal tract	7/655	147/12496	0,38
Genitourinary	4/1362	268/28145	0
Germ cell		42/46905	
Head and neck	365/42961	147/57415	0
Heart	237/53290		
Kidney	111/58083	356/71217	0
Limb			
Liver	217/56963	993/71251	0
Lung	365/97376	954/164020	0
Lymph node	410/78636	1398/46552	0
Lymphoreticular	192/30672	627/75782	0
Mammary gland	196/39575	849/79986	0
Muscle	40/67686	336/35975	0
Nervous system	113/11552	609/56801	0,18
Ovary	44/8892	508/80203	0,05
Pancreas	83/7119	517/71373	0
Pancreatic islet	749/82694	0/0	
Peripheral nervous system	211/23114	0/670	0,01
Pineal gland	17/6137		
Pituitary gland	202/12974	4/753	0,01
Placenta	623/183001	350/38484	0
Pooled tissue	451/294697	343/29023	0
Prostate	174/59569	547/58716	0
Retina	206/44604		
Salivary gland	5/189	247/16856	NaN
Skin	174/41702	690/120243	0
Soft tissue	2/275	92/6135	0,17
Spleen	49/15719		
Stem cell			
Stomach	223/17960	1205/112880	0,02
Synovium	0/247	35/1340	0,01
Testis	131/90345	884/36779	0
Thymus	4/4103	0/161	0,45
Thyroid	21/4279	23/2717	NaN
Uncharacterised tissue	1846/161901	321/37638	NaN
Uterus	97/31298	1638/124854	NaN
Vascular	270/25159		NaN
White blood cells			NaN
Whole body	556/64816		NaN

culated the expression value in the cancer tissue and the expression value in normal tissue. If the expression in cancer tissue is larger than in normal tissue and if in addition the given *p* value is smaller than 0.05, we recorded this gene. After collecting these data we could easily evaluate which gene is over-expressed in cancer in more than a given number of tissues. We set this threshold to 18 (50%). The program can be accessed *via* <http://www.embl-heidelberg.de/~altenber/gemuend/Program>.

3 Results

Table 2 lists the 42 of 19 016 genes (rows) that have been found over-expressed in half or more of the 36 investigated cancers (columns). The search included all 51 cancers listed in Table 1 (see Section 2) and all 19 016 annotated genes, but only the genes that are over-expressed in 18 (50%) or more of these cancers have been included in Table 2. We define these genes as “ubiquitously over-expressed” (or “ubiquitous”), but we are aware that this definition may finally need some readjustment, when more data become available in the literature. If, in Table 2, a field is blank, this may have one of several meanings: Either, the gene is indeed not over-expressed, or there was not a sufficient amount of experiments available in literature. In other words, empty cells in Table 2 may be filled in the future, whereas the probability that an already filled position will have to be removed is small.

Placenta, skin, prostate, liver and testis cancers have all 42 ubiquitous genes over-expressed, those of the kidney, uterus, lymph node, mammary gland, brain, lung, muscle and eye over-express between 39 and 41 of these genes. Then a gap occurs, the cancer with the next frequent over-expression is ovary cancer with only 35 over-expressed genes. When these genes are classified according to their biochemical function, it turns out that 15 are involved in protein biosynthesis and folding, and six are involved in glycolysis. The latter finding reflects the well-known fact that many cancers enhance glycolysis (the Warburg effect). With the results of the present investigation and those from the study by Altenberg and Greulich [13], cancers can now be classified according to the mechanisms with which the Warburg effect is achieved, and it helps to settle a controversial discussion on the relevance of glycolysis [14]. The findings regarding protein biosynthesis and folding highlight the important role that this type of biochemical mechanism has in general cancer. Table 3 lists details on these findings.

The fact that 21 of the 42 ubiquitously over-expressed cancer genes are involved in protein biosynthesis, protein folding and glycolysis indicates that these functions may reflect a very general aspect in cancer: cancer as a protein biosynthesis and folding disease as well as a glycolysis disorder. The latter is not new, whereas the protein aspect may be somewhat more surprising, since in the literature cancer

is often treated as a disease of the cell cycle. However, cell cycle genes, with five genes, are not as frequently over-expressed as genes of protein biosynthesis.

Table 3. Ubiquitous cancer genes according to their function

Protein biosynthesis and folding	15
Glycolysis	6
Cell cycle genes, growth and mobility	5
DNA replication and processing	4
Translation/elongation	3
Oncogene	1
Other	8
Total	42

4 Discussion

In Table 2 the expression pattern for each column (= cancer) is different. Thus, using the 42 gene products found in the present work, each cancer class can in principle be identified. Theoretically, an even much smaller subclass of these gene products would be sufficient, *i.e.* the 42 molecules identified in the present work allow for a substantial internal control. It may even be possible to detect individual deviation from the general cancer pattern resulting from the fact that expression in the same type of cancer varies considerably from patient to patient.

The present data on 42 genes in 36 cancers, *i.e.* on more than 1500 data points, each representing a large number of experiments, appear to be sufficiently robust in order to achieve meaningful results on cancers as a whole. In addition, they warrant general statements on the role of these genes in cancer.

When scheduling a protein chip based on these findings, one would probably select ten to 20 of the ubiquitous genes and add some more, for example, oncogene or tumour suppressor gene products which are specific for single or small numbers of cancers. This would allow, with a total of 30–50 molecules, for an assembly of a multipurpose protein micro-assay for cancer diagnosis.

This work was supported by the German Research Ministry, Grant 3 N 8028 (SCREEN) and by the Thüringer Ministerium für Wissenschaft und Kunst. B. A. and C. G. thank Dr. Toby Gibson for enabling them to perform this work at the European Molecular Biology Laboratory.

5 References

- [1] Striebel, H. M., Schellenberg, P., Grigaravicius, P., Greulich, K. O., *Proteomics* 2004, 4, 1703–1711.
- [2] Futreal, P. A., Coin, L., Marshall, M., Down, T., Hubbard, T., Wooster, R., Rahman, N. *et al.*, *Nat. Rev. Cancer* 2004, 4, 177–183.

- [3] Segal, E., Friedmann, N., Koller, D., Regev, A., *Nat. Genetics* 2004, *36*, 1090–1098.
- [4] Stamatoyannopoulos, J. A., *Genomics* 2004, *84*, 449–457.
- [5] Järvinen, A.-K., Hautaniemi, S., Edgren, H., Auvinen, P., Saarela, J., Kallioniemi, O.-P., Monni, O., *Genomics* 2004, *83*, 1164–1168.
- [6] Lu, J., Lal, A., Merriman, B., Nelson, S., Riggins, G., *Genomics*, 2004, *84*, 631–636.
- [7] Lal, A., Lash, A. E., Altschul, S., Velculescu, V., Zhang, L., McLendon, R., Marra, M. A. *et al.*, *Cancer Res.* 1999, *59*, 5403–5407.
- [8] Martin, K. J., Pardee, A. B., *Proc. Natl. Acad. Sci. USA*, 2000, *97*, 3789–3791.
- [9] Strausberg, R. L., Greenhut, S. F., Grouse, L. H., Schaefer, C. F., Buetow, K. H., *Trends Cell. Biol.*, 2001, *11*, S66–S71.
- [10] Strausberg, R. L., *J. Pathol.*, 2001, *195*, 31–40.
- [11] Schaefer, C., Grouse, L., Buetow, K., Strausberg, R. L., *Cancer J.*, 2001, *7*, 52–60.
- [12] Zhang, L., Ma, X. L., Zhang, Q., Ma, C. L., Wang, P. P., Sun, Y. F., Zhao, Y. X. *et al.*, *Gene* 2001, *67*, 193–200.
- [13] Altenberg, B., Greulich, K. O., *Genomics* 2004, *84*, 1014–1020.
- [14] Zu, X. L., Guppy, M., *Bioch. Bioph. Res. Comm.*, 2004, *313*, 459–465.