*Nucleic Acids Research, 2006, Vol. 34, Database issue* ***D607–D612***
*doi:10.1093/nar/gkj082*

# HPtaa database-potential target genes for clinical diagnosis and immunotherapy of human carcinoma

**Xiaosong Wang[1,3], Haitao Zhao[4], Qingwen Xu[1], Weibo Jin[2], Changning Liu[2], Huagang Zhang[1], Zhibin Huang[1], Xinyu Zhang[6], Yu Zhang[1], Dianqi Xin[3], Andrew J. G. Simpson[5], Lloyd J. Old[5], Yanqun Na[3], Yi Zhao[2] and Weifeng Chen[1,*]**

[1]Department of Immunology, School of Basic Medical Sciences, Peking University Health Science Center, Beijing 100083, China, [2]Bioinformatics Research Group, Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Beijing 100080, China, [3]Department of Urology, First Hospital of Peking University, Institute of Urology, Peking University, Beijing 100034, China, [4]Department of General Surgery, Peking Union Medical College Hospital, Chinese Academy of Medical Sciences, Beijing 100032, China, [5]Ludwig Institute for Cancer Research, New York Branch at Memorial Sloan–Kettering Cancer Center, New York, NY 10021, USA and [6]Department of Biological Science and Biotechnology, Ministry of Education Key Laboratory of Bioinformatics, Tsinghua University, Beijing 100084, China

Received August 15, 2005; Revised September 20, 2005; Accepted October 11, 2005

## ABSTRACT

**Tumor-associated antigens (TAAs) have been the most actively employed targets in the clinical diagnosis and treatment of human carcinoma, such as PSA in the diagnosis of prostate cancer and NY-ESO-1 in the immunotherapy of melanoma and other cancers. However, identification of TAAs has often been hampered by the complicated and laborsome laboratory procedures. In order to accelerate the process of tumor antigen discovery, and thereby improve diagnosis and treatment of human carcinoma, we have made an effort to establish a publicly available Human Potential Tumor Associated Antigen database (HPtaa) with potential TAAs identified by *in silico* computing (http://www.hptaa.org). Tumor specificity was chosen as the core of tumor antigen evaluation, together with other relevant clues. Various platforms of gene expression, including microarray, expressed sequence tag and SAGE data, were processed and integrated by several penalty algorithms. A total of 3518 potential TAAs have been included in the database, which is freely available to academic users. As far as we know, this database is the first one addressing human potential TAAs, and the first one integrating various kinds of expression platforms for one purpose.**

## INTRODUCTION

Tumor-associated antigens (TAAs) have been the most actively employed targets in the clinical diagnosis and treatment of human carcinoma. TAAs are encoded by normal or mutated genes in the human genome whose products can elicit humoral or cellular anti-tumor immunity. They can be classified as tissue restrictive and non-tissue restrictive antigens, according to their expression pattern in normal tissues (NTs). Tissue restrictive TAAs, including cancer-testis antigens (CT antigens), differentiation antigens and oncofetal antigens, have deeply affected the clinical oncology. For example, PSA as a differentiation antigen is indispensable in diagnosis and prognosis evaluation of prostate cancer (1), AFP as an oncofetal antigen has been widely used in the diagnosis of hepatocellular carcinoma (2), and NY-ESO-1 as a cancer-testis antigen has been shown to induce broad integrated immune responses in melanoma patients (3). As a result, identification of clinical applicable TAAs is of great importance to cancer immunologists and clinicians.

Traditionally, TAAs are identified through T cell epitope cloning, serological analysis of cDNA expression libraries, subtraction hybridization and differential display analysis (4). Laboratory procedures, although successful, are extremely laborious. Recently, immunoinformatics has emerged as an efficient way for the identification of TAAs. These *in silico* methods were generally based on the fact that tumor-specific expression patterns usually reflect heterogeneity of the gene products, which, given that protein expression correlates with

*To whom correspondence should be addressed. Tel: +86 10 8280 2593; Fax: +86 10 8280 1436; Email: wfchen@public.bta.net.cn
Correspondence may also be addressed to Yi Zhao. Tel: +86 10 6256 5533, ext. 5717; Fax: +86 10 6256 7724; Email: biozy@ict.ac.cn

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors

mRNA expression, is at the core of immunogenicity. Thus, successful identification of novel TAAs through expression database mining has not been reported occasionally (5–8).

It has been conventionally considered that different expression platforms cannot be integrated together because of the difficulties of normalization. Based on the fact that individual series of expression data can be used separately in the case of tumor antigen identification, we believe that all kinds of expression platforms can be integrated by gathering all the individual results. Our own experience shows that platform integration greatly increases the efficiency of TAA identification.

In order to accelerate the process of tumor antigen discovery, and thus improve diagnosis and treatment of human carcinoma, we decided to establish a publicly available database for potential TAAs (pTAA) identified by *in silico* computing, named Human Potential Tumor Associated Antigen database (HPtaa). As mentioned above, tumor-specific expression pattern not only correlates with immunogenicity, but also is the prerequisite for clinical application. Thus, we chose tumor-specific expression as the core of tumor antigen evaluation. Other relevant clues were also considered; including coding capacity, chromosomal location, subcellular location and the knowledge of gene function. As far as we know, this database is the first one addressing human potential TAAs, and the first one integrating various kinds of expression platforms for one purpose.

## DATABASE CONSTRUCTION

### Data source

The HPtaa database integrates various expression platforms, including carefully chosen publicly available microarray expression data, GEO SAGE data, expressed sequence tag (EST) expression data together with other relevant databases required for TAA discovery, such as CGAP (9), CCDS (http://www.ncbi.nlm.nih.gov/projects/CCDS/), OMIM (10), Uniprot (11) and the Gene Ontology database (12). Microarray datasets were divided into normal tissue series and cancer series. Normal tissue series include five famous datasets: GNF (13,14), UCLA (http://microarray.genetics.ucla.edu/geneexp/public/), GENENOTE (15), GeMDBJ (https://www.gemdbj.jp/dgdb/) and GEOJP (http://www.genome.rcast.u-tokyo.ac.jp/normal/). The cancer series include 45 datasets from 12 series, covering 14 major cancer types (16–27). The EST (28) and SAGE data (29) covers 9 additional cancer types, resulting in HPtaa covering a total of 23 cancer types.

### Data processing

Each microarray dataset was processed individually to avoid the problem of normalization. For datasets of NTs, we used known cancer-testis antigens (30) as a training set to generate a detection call matrix, and then tissue restriction score (TRS) was computed for each probe. The tissue restriction threshold for each dataset was determined according to the TRS interval containing 90% CT antigens, and then the TRS of all the normal tissue datasets were assembled by Unigene ID. The tissue restriction penalty (TRP) was computed according to all the probes' TRS of each Unigene and their confidence evidenced by source sequences of probes and the amount of samples in the corresponding dataset. See Supplementary Data for details of the algorithms.

Differential expression analysis and significance tests were carried out separately for each cancer microarray datasets and each cancer type of the EST and SAGE expression data. For each significantly expressed probe, the cancer/normal ratio was computed and assembled by Unigene ID. An overexpression penalty (OP) for each cancer type was computed according to the overexpression ratios and their confidence accessed by source sequences of probes. The tumor specificity penalty (TSP) was then computed as $TSP = TRP \times OP$. This algorithm is designed according to the assumption that tumor specificity increases in proportion to OP and TRP. (Figure 1).

### Database content

All genes with $TSP > 115$ were considered as tumor specific and were included in the HPtaa database. This cutoff was set to obtain an optimal balance between database content and identification of known tumor antigens. The CGAP, GO, CCDS, UniProt and OMIM databases were thereafter integrated to annotate each pTAA, and all the original expression data were picturized and linked to the corresponding pTAAs.

### Statistics

The HPtaa database contains 3518 potential TAAs for up to 23 human cancer types. To test the quality of the database, we checked how many known tumor antigens it contains. We found that 41 known CT antigens (50% of all known CT antigens) (30), 6 known differentiation antigens (33% of all known differentiation antigens) (31) and 2 known oncofetal antigens (100%, CEA and AFP) were successfully screened out (see Supplementary Data for detailed information). Interestingly, most of the CT antigens screened out with current algorithms generally have a high overexpression rate compared with those not found. This shows that with our statistical significance test, genes stably upregulated in cancerous tissues are more likely to be picked out, which are also more valuable than those occasionally overexpressed. Totally 3163 known genes and 355 uncharacterized genes were included in the database, among which 1804 genes have publication reports, 2172 genes have CCDS annotation. The database contains 237 membrane proteins, 172 secretory proteins and 127 genes mapped to the X chromosome. (See Data Retrieval below for the significance of these properties.)

## DATA RETRIEVAL

The database provides an easy-to-use query interface. Users can query interesting genes against HPtaa with a basic search, or query for pTAAs with defined features through an advanced search. The cancer type choice allows users to choose pTAAs of their cancer types of interest. Chromosome choice allows users to choose whether the pTAAs should locate on the X or on the Y chromosome, where CT antigens aggregate. The coding capacity choice allows users to define the coding capacity of a pTAA, as coding genes are more likely to be TAAs. It should be noted that novel genes often have undetermined coding capacity. Subcellular location choice allows

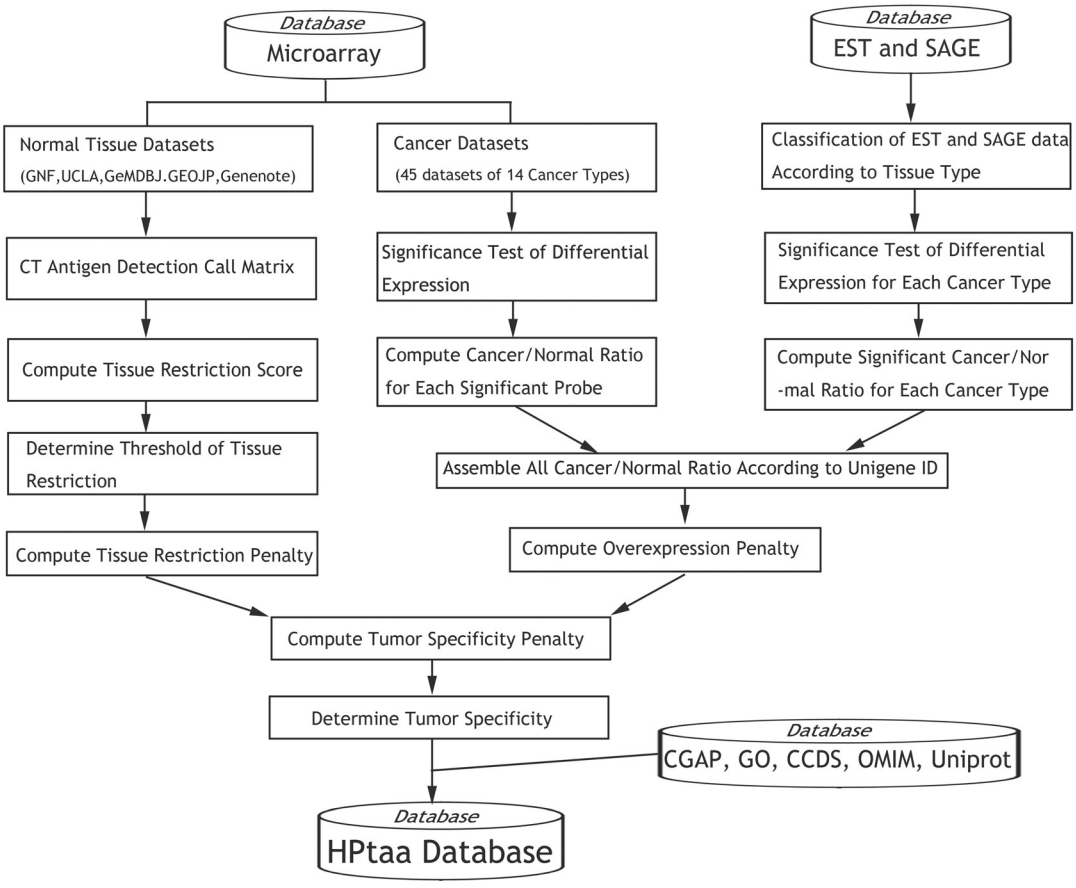


**Figure 1.** The flow chart of data procession of HPtaa database.

users to choose membrane pTAAs or secretory pTAAs. Membrane proteins are more valuable in the clinical treatment of carcinoma, while secretory proteins are of more interest to diagnostics. The mRNA choice allows users to choose pTAA with an mRNA sequence, which is easy to identify. OMIM choice allows users to choose whether pTAAs have publication supported functional annotations. Genes with no OMIM ID usually have no cancer-related reports. 'ESTs from NT' choice allows users to choose the number of ESTs from non-germinal and non-fetal NTs clustered to each pTAA.

The result page of a database search contains three important parameters for evaluating a pTAA, i.e. the TRP, OP and TSP, as outlined above. When trying to identify highly tumor-specific genes, the three values should be considered together. TRP defines the degree of restrictive expression of a given gene across human NTs and its confidence. The higher the TRP, the more restrictive is the expression of a given gene across NTs. OP defines whether the expression of a given gene is significantly upregulated in cancerous tissues compared with corresponding NTs. The value of OP does not merely reflect the differential expression ratio, but combines the ratio with other clues indicating overexpression. The higher the OP value, the higher is the likelihood of overexpression. TSP gives an overall view of the tumor specificity. The higher the TSP, the higher is the degree of tumor-specific expression.

Users will find that for a given pTAA/gene the OP and TSP values varies between different cancer types. The reason behind this is that individual researchers will usually need tumor-specific genes that are overexpressed in the particular cancer type they study. Cancer type specific OP and TSP values may accommodate for this requirement.

## DISCUSSION

### How to make your choice

The HPtaa database aims directly at clinical diagnosis and treatment of human carcinoma, and users should thus choose pTAAs according to their purpose. If a user wants to find tumor markers for the cancer types he/she studies, the secretory pTAAs with the highest cancer type specific TSP and OP values should be favored irrespective of the TRP value. The rationality of this lies in the fact that tumor markers usually have less tissue-restrictive expression, and the expression in cancerous tissue needs to be extremely high to favor about detection. We recommend users to examine the figure of differential expression ratio to evaluate the details and degree of overexpression (Figure 2).

If users want to find pTAAs with therapeutic value, the pTAAs with highest TRP should be selected, as higher TRP values are likely to imply lesser side effects. We recommend users to examine in detail the figure of normal tissue expression in the detail page, as pTAA with extremely low detection value across NTs may best serve the therapeutic purposes. By

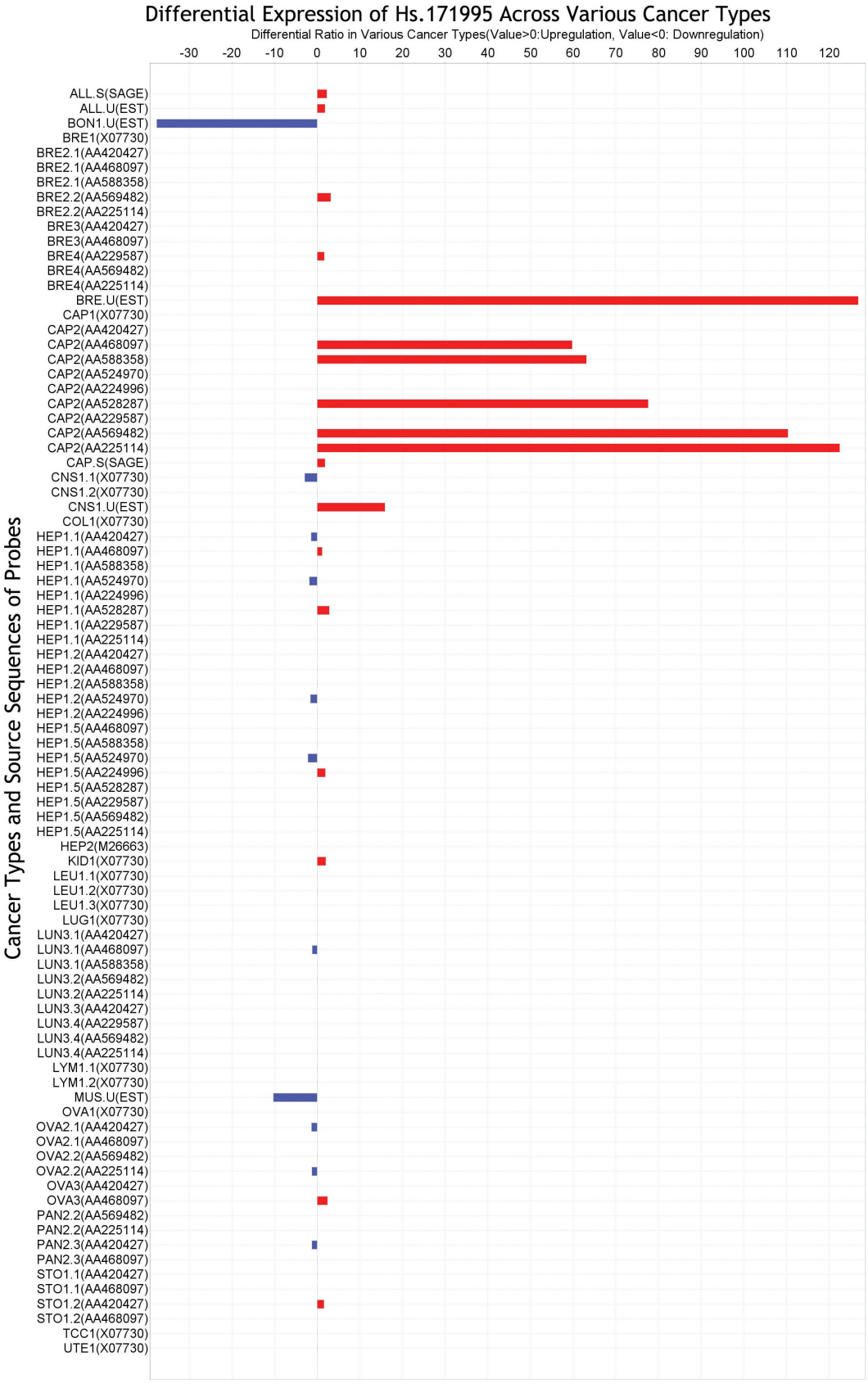


**Figure 2.** Mean differential expression ratio of PSA across various cancer types. When upregulated significantly in cancerous tissues, the value was computed as 'cancer/normal'; when downregulated significantly the value was computed as '– (normal/cancer)'. The *y*-axis shows the names of the cancer datasets and source sequences of the probes in a given dataset. Red color represents upregulation and blue color downregulation.

restricting the number of ESTs from NT, users can further screen out tissue-restrictive genes also evidenced by EST data. With respect to subcellular location, membrane pTAAs are best targets for monoclonal antibody treatment, while intracellular pTAAs constitute a good repertoire of peptide vaccination targets.

### Evaluating potential TAA

The expression patterns of CT antigens were usually evaluated by endpoint RT–PCR. As RT–PCR is generally more sensitive than other methods, tissue restrictive genes in the database may appear less tissue restrictive when analyzed by RT–PCR with 35 cycles. In our experience, the coincidence of HPtaa defined tumor specificity with RT–PCR result should be ~10%. As a result, we recommend real-time PCR or northern blot instead of end point PCR in evaluating the expression difference between cancerous tissues and NTs of human body.

### Functional considerations

As more and more TAAs have been found to be related to carcinogenesis, the functional aspects of tumor antigens have gradually aroused immunologists' attention. As pTAAs are virtually tumor-specific genes, together with the fact that many organ-specific genes are found to be related to the function of the organs they are specifically expressed, it is not surprising to find that these genes also contribute to the proliferation or metastasis of human carcinomas. In evidence of this, users can find many genes known to be related to carcinogenesis in our database. To help with users interested in functional aspects of cancer-specific genes, we provide an annotation of gene ontology and motif for each gene in the detail page.

Users may find that some genes with high-TSP are actually immune system-specific genes. We suspect that the upregulation of these genes may originate from tumor infiltration activity of immune cells. However, as it has been shown that tumor cells overexpress genes encoding antibodies with unknown specificity (32), we cannot exclude the possibility that other unrecognized mechanisms may explain the high-TSP scores of these genes.

## FUTURE DIRECTION

The development of penalty algorithms for the HPtaa database has been guided by practical experience. Further experimental validation will be carried out to evaluate their efficacy, and to facilitate refining of the algorithms. As large-scale expression data accumulate fast, more expression data will be integrated to improve gene and cancer type coverage. A classification system will be established to address the expression privilege of pTAAs in NTs, as in the case of tumor antigens.

### Citing HPtaa

Users are requested to cite this article and quote the HPtaa home page URL (http://www.hptaa.org).

## SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

## ACKNOWLEDGEMENTS

We thank Beijing Cofly Bioinformatics Company (http://www.co-fly.net) for help with the analysis of microarray cancer series. This work has been supported by grants from National 863 program in China (no. 2001AA215411) and National Natural Science Foundation of China (no. 30531160045 and no. 30570393) and Ludwig Institute for Cancer Research, New York (KSP #003). Funding to pay the Open Access publication charges for this article was provided by National Natural Science Foundation of China (no. 30531160045).

*Conflict of interest statement*. None declared.

## REFERENCES

1. Gretzer,M.B. and Partin,A.W. (2003) PSA markers in prostate cancer detection. *Urol. Clin. North Am.*, **30**, 677–686.
2. Sell,S. (1978) AFP as a marker for liver cell injury: differentiation of tumor growth, hepatotoxicity, and carcinogenesis. *UCLA Forum Med. Sci.*, **20**, 51–58.
3. Davis,I.D., Chen,W., Jackson,H., Parente,P., Shackleton,M., Hopkins,W., Chen,Q., Dimopoulos,N., Luke,T., Murphy,R. *et al.* (2004) Recombinant NY-ESO-1 protein with ISCOMATRIX adjuvant induces broad integrated antibody and $CD4^+$ and $CD8^+$ T cell responses in humans. *Proc. Natl Acad. Sci. USA*, **101**, 10697–10702.
4. Gilboa,E. (2004) The promise of cancer vaccines. *Nature Rev. Cancer*, **4**, 401–411.
5. Chen,Y.T., Venditti,C.A., Theiler,G., Stevenson,B.J., Iseli,C., Gure,A.O., Jongeneel,C.V., Old,L.J. and Simpson,A.J. (2005) Identification of CT46/HORMAD1, an immunogenic cancer/testis antigen encoding a putative meiosis-related protein. *Cancer Immun.*, **5**, 9.
6. Scanlan,M.J., Gordon,C.M., Williamson,B., Lee,S.Y., Chen,Y.T., Stockert,E., Jungbluth,A., Ritter,G., Jager,D., Jager,E. *et al.* (2002) Identification of cancer/testis genes by database mining and mRNA expression analysis. *Int. J. Cancer*, **98**, 485–492.
7. Dong,X.Y., Su,Y.R., Qian,X.P., Yang,X.A., Pang,X.W., Wu,H.Y. and Chen,W.F. (2003) Identification of two novel CT antigens and their capacity to elicit antibody response in hepatocellular carcinoma patients. *Br. J. Cancer*, **89**, 291–297.
8. Segal,N.H., Blachere,N.E., Guevara-Patino,J.A., Gallardo,H.F., Shiu,H.Y., Viale,A., Antonescu,C.R., Wolchok,J.D. and Houghton,A.N. (2005) Identification of cancer-testis genes expressed by melanoma and soft tissue sarcoma using bioinformatics. *Cancer Immun.*, **5**, 2.
9. Strausberg,R.L., Greenhut,S.F., Grouse,L.H., Schaefer,C.F. and Buetow,K.H. (2001) *In silico* analysis of cancer through the Cancer Genome Anatomy Project. *Trends Cell Biol.*, **11**, S66–S71.
10. Hamosh,A., Scott,A.F., Amberger,J.S., Bocchini,C.A. and McKusick,V.A. (2005) Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.*, **33**, D514–D517.
11. Bairoch,A., Apweiler,R., Wu,C.H., Barker,W.C., Boeckmann,B., Ferro,S., Gasteiger,E., Huang,H., Lopez,R., Magrane,M. *et al.* (2005) The Universal Protein Resource (UniProt). *Nucleic Acids Res.*, **33**, D154–D159.
12. Harris,M.A., Clark,J., Ireland,A., Lomax,J., Ashburner,M., Foulger,R., Eilbeck,K., Lewis,S., Marshall,B., Mungall,C. *et al.* (2004) The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.*, **32**, D258–D261.
13. Su,A.I., Wiltshire,T., Batalov,S., Lapp,H., Ching,K.A., Block,D., Zhang,J., Soden,R., Hayakawa,M., Kreiman,G. *et al.* (2004) A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc. Natl Acad. Sci. USA*, **101**, 6062–6067.
14. Su,A.I., Cooke,M.P., Ching,K.A., Hakak,Y., Walker,J.R., Wiltshire,T., Orth,A.P., Vega,R.G., Sapinoso,L.M., Moqrich,A. *et al.* (2002) Large-scale analysis of the human and mouse transcriptomes. *Proc. Natl Acad. Sci. USA*, **99**, 4465–4470.
15. Shmueli,O., Horn-Saban,S., Chalifa-Caspi,V., Shmoish,M., Ophir,R., Benjamin-Rodrig,H., Safran,M., Domany,E. and Lancet,D. (2003) GeneNote: whole genome expression profiles in normal human tissues. *C. R. Biol.*, **326**, 1067–1072.

16. Bhattacharjee,A., Richards,W.G., Staunton,J., Li,C., Monti,S., Vasa,P., Ladd,C., Beheshti,J., Bueno,R., Gillette,M. *et al.* (2001) Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. *Proc. Natl Acad. Sci. USA*, **98**, 13790–13795.
17. Chen,X., Cheung,S.T., So,S., Fan,S.T., Barry,C., Higgins,J., Lai,K.M., Ji,J., Dudoit,S., Ng,I.O. *et al.* (2002) Gene expression patterns in human liver cancers. *Mol. Biol. Cell*, **13**, 1929–1939.
18. Chen,X., Leung,S.Y., Yuen,S.T., Chu,K.M., Ji,J., Li,R., Chan,A.S., Law,S., Troyanskaya,O.G., Wong,J. *et al.* (2003) Variation in gene expression patterns in human gastric cancers. *Mol. Biol. Cell*, **14**, 3208–3215.
19. Garber,M.E., Troyanskaya,O.G., Schluens,K., Petersen,S., Thaesler,Z., Pacyna-Gengelbach,M., van de Rijn,M., Rosen,G.D., Perou,C.M., Whyte,R.I. *et al.* (2001) Diversity of gene expression in adenocarcinoma of the lung [Erratum (2002) *Proc. Natl Acad. Sci. USA,* 99, 1098.]. *Proc. Natl Acad. Sci. USA*, **98**, 13784–13789.
20. Iacobuzio-Donahue,C.A., Maitra,A., Olsen,M., Lowe,A.W., van Heek,N.T., Rosty,C., Walter,K., Sato,N., Parker,A., Ashfaq,R. *et al.* (2003) Exploration of global gene expression patterns in pancreatic adenocarcinoma using cDNA microarrays. *Am. J. Pathol.*, **162**, 1151–1162.
21. Lapointe,J., Li,C., Higgins,J.P., van de Rijn,M., Bair,E., Montgomery,K., Ferrari,M., Egevad,L., Rayford,W., Bergerheim,U. *et al.* (2004) Gene expression profiling identifies clinically relevant subtypes of prostate cancer. *Proc. Natl Acad. Sci. USA*, **101**, 811–816.
22. Lenburg,M.E., Liou,L.S., Gerry,N.P., Frampton,G.M., Cohen,H.T. and Christman,M.F. (2003) Previously unidentified changes in renal cell carcinoma gene expression identified by parametric analysis of microarray data. *BMC Cancer*, **3**, 31.
23. Mecham,B.H., Klus,G.T., Strovel,J., Augustus,M., Byrne,D., Bozso,P., Wetmore,D.Z., Mariani,T.J., Kohane,I.S. and Szallasi,Z. (2004) Sequence-matched probes produce increased cross-platform consistency and more reproducible biological results in microarray-based gene expression measurements. *Nucleic Acids Res.*, **32**, e74.
24. Ramaswamy,S., Tamayo,P., Rifkin,R., Mukherjee,S., Yeang,C.H., Angelo,M., Ladd,C., Reich,M., Latulippe,E., Mesirov,J.P. *et al.* (2001) Multiclass cancer diagnosis using tumor gene expression signatures. *Proc. Natl Acad. Sci. USA*, **98**, 15149–15154.
25. Schaner,M.E., Ross,D.T., Ciaravino,G., Sorlie,T., Troyanskaya,O., Diehn,M., Wang,Y.C., Duran,G.E., Sikic,T.L., Caldeira,S. *et al.* (2003) Gene expression patterns in ovarian carcinomas. *Mol. Biol. Cell*, **14**, 4376–4386.
26. Ye,Q.H., Qin,L.X., Forgues,M., He,P., Kim,J.W., Peng,A.C., Simon,R., Li,Y., Robles,A.I., Chen,Y. *et al.* (2003) Predicting hepatitis B virus-positive metastatic hepatocellular carcinomas using gene expression profiling and supervised machine learning. *Nature Med.*, **9**, 416–423.
27. Zhao,H., Langerod,A., Ji,Y., Nowels,K.W., Nesland,J.M., Tibshirani,R., Bukholm,I.K., Karesen,R., Botstein,D., Borresen-Dale,A.L. *et al.* (2004) Different gene expression patterns in invasive lobular and ductal carcinomas of the breast. *Mol. Biol. Cell*, **15**, 2523–2536.
28. Pontius,J.U. and Schuler,G.D. (2003) UniGene: a unified view of the transcriptome. NCBI, Bethesda, MD. *The NCBI Handbook*. National Library of Medicine (US), NCBI.
29. Barrett,T., Suzek,T.O., Troup,D.B., Wilhite,S.E., Ngau,W.C., Ledoux,P., Rudnev,D., Lash,A.E., Fujibuchi,W. and Edgar,R. (2005) NCBI GEO: mining millions of expression profiles—database and tools. *Nucleic Acids Res.*, **33**, D562–D566.
30. Scanlan,M.J., Simpson,A.J. and Old,L.J. (2004) The cancer/testis genes: review, standardization, and commentary. *Cancer Immun.*, **4**, 1.
31. Novellino,L., Castelli,C. and Parmiani,G. (2005) A listing of human tumor antigens recognized by T cells: March 2004 update. *Cancer Immunol. Immunother.*, **54**, 187–207.
32. Qiu,X., Zhu,X., Zhang,L., Mao,Y., Zhang,J., Hao,P., Li,G., Lv,P., Li,Z., Sun,X. *et al.* (2003) Human epithelial cancers secrete immunoglobulin G with unidentified specificity to promote growth and survival of tumor cells. *Cancer Res.*, **63**, 6488–6495.