Nucleic Acids Research, 2006, Vol. 34, Database issue ***D613–D616***
doi:10.1093/nar/gkj022

ITTACA: a new database for integrated tumor transcriptome array and clinical data analysis

Adil Elfilali[1,2], Séverine Lair[1,2], Catia Verbeke[1,2], Philippe La Rosa[1], François Radvanyi[2] and Emmanuel Barillot[1,*]

[1]Institut Curie, Service Bioinformatique, 26 rue d'Ulm, Paris, 75248 cedex 05, France and
[2]Institut Curie, CNRS UMR 144, 26 rue d'Ulm, Paris, 75248 cedex 05, France

Received August 9, 2005; Revised and Accepted September 19, 2005

ABSTRACT

Transcriptome microarrays have become one of the tools of choice for investigating the genes involved in tumorigenesis and tumor progression, as well as finding new biomarkers and gene expression signatures for the diagnosis and prognosis of cancer. Here, we describe a new database for Integrated Tumor Transcriptome Array and Clinical data Analysis (ITTACA). ITTACA centralizes public datasets containing both gene expression and clinical data. ITTACA currently focuses on the types of cancer that are of particular interest to research teams at Institut Curie: breast carcinoma, bladder carcinoma and uveal melanoma. A web interface allows users to carry out different class comparison analyses, including the comparison of expression distribution profiles, tests for differential expression and patient survival analyses. ITTACA is complementary to other databases, such as GEO and SMD, because it offers a better integration of clinical data and different functionalities. It also offers more options for class comparison analyses when compared with similar projects such as Oncomine. For example, users can define their own patient groups according to clinical data or gene expression levels. This added flexibility and the user-friendly web interface makes ITTACA especially useful for comparing personal results with the results in the existing literature. ITTACA is accessible online at http://bioinfo.curie.fr/ittaca.

INTRODUCTION

Recently, transcriptomic data from DNA microarrays have allowed researchers to analyze gene expression and explore the molecular basis of disease. Analysis of the transcriptome allows researchers in oncology to study the regulatory mechanisms and biochemical pathways involved in cell transformation and tumor progression. The different types of cancer can also be classified according to their gene expression signatures. This can uncover gene expression patterns that correlate with various anatomo-pathological or clinical characteristics of tumors, and allow novel diagnostic biomarkers or therapeutic targets to be identified.

Increasing amounts of DNA microarray data are being generated as more cancer research laboratories start to use this technology. These data are being made available to the research community by the efforts of the Microarray Gene Expression Data society (MGED). The MGED encourages researchers to deposit their data in public repositories following the Minimum Information About a Microarray Experiment (MIAME) guidelines (1,2). Complete microarray datasets are starting to become available, either as Supplementary Data in publications or in public databases, such as Gene Expression Omnibus (GEO) (3) or ArrayExpress (4). However, the data are often not used to their full potential, as they are generally analyzed with one particular clinical or biological question in mind. It is also difficult to take advantage of these public datasets for new studies because the data are not always easy to access and analyze. In particular, clinical data are often missing in public databases, although it may be described in the publication or supplied as Supplementary Data.

Integrated Tumor Transcriptome Array and Clinical data Analysis (ITTACA) was developed by the Bioinformatics group and the Molecular Oncology group of the Institute Curie, a leading center in cancer research and treatment. It aims to centralize public clinical and transcriptomic datasets and make them available for analysis on the web. It allows users access to transcriptome and clinical data for tumors from experiments in published articles. ITTACA allows various class comparison analyses to be carried out between two different user-defined groups and to visualize gene expression profiles. ITTACA allows users to reanalyze transcriptomic data from the scientific literature in addition to presenting the results of predefined analyses. Researchers can therefore

*To whom correspondence should be addressed. Email: emmanuel.barillot@curie.fr

Table 1. Dataset content of ITTACA

Dataset references	Tissue	Platform	No. of samples	Sources
(15) (Dataset from supervised class analysis)	Bladder carcinoma	Affymetrix HugeneFL Genechip	40	Gene Expression Omnibus
(15) (Dataset from unsupervised class analysis)	Bladder carcinoma	Affymetrix HugeneFL Genechip	31	Gene Expression Omnibus
(16)	Breast carcinoma	Non-commercial	58	Lund University
(17)	Breast carcinoma	Non-commercial	17	PNAS
(18)	Breast carcinoma	Non-commercial	22	NHGRI
(19)	Breast carcinoma	Affymetrix HG-U95Av2	89	Duke Institute for Genome Sciences and Policy
(20)	Bladder carcinoma	Clontech	50	Düsseldorf University
(21)	Uveal melanoma	Affymetrix HG-U133A and HG-U133B	25	Cancer Research
(22)	Breast carcinoma	Non-commercial	122	PNAS
(23)	Breast carcinoma	Non-commercial	99	PNAS
(24)	Uveal melanoma	Affymetrix HG-U95Av2	20	Duisburg-Essen University
(10)	Breast carcinoma	Hu25K, Agilent	117	Nature
(9)	Breast carcinoma	Affymetrix HugeneFL Genechip	49	PNAS

ITTACA currently contains 13 public gene expression datasets with both transcriptome and clinical data for human cancer tissue samples from bladder (3 datasets) carcinomas, breast (8 datasets) carcinomas and uveal melanoma (2 datasets).

verify their results and build on previous work. For example, they can assess the differential expression of the gene they are particularly interested in or compare their own transcriptome array results with results in the existing literature.

DATA CONTENT

ITTACA currently contains publicly available gene expression datasets with both transcriptomic and clinical data for human cancer tissue samples. Complete datasets were obtained through online resources or provided on request by the authors of publications. Currently, ITTACA focuses on three types of cancer, for which the data integration has been as exhaustive as possible: breast carcinoma, bladder carcinoma and uveal melanoma. Table 1 lists all of the public datasets currently available in ITTACA (12 article references: 2 for bladder carcinoma, 8 for breast carcinoma and 2 for uveal melanoma).

Datasets are selected based on the availability of relevant anatomo-clinical data. For each studied sample, the following data must also be available: normalized gene expression data (from different commercial or non-commercial platforms), proper annotations for probes or probe sets (at least one identifier or accession number to a public databank) and anatomo-pathological and/or clinical characteristics.

Probe annotations, such as GeneId, Gene name, Gene symbol, Unigene Identifier and Gene Ontology identifier, may help users perform a query in ITTACA. Therefore, publicly available annotation tools, such as MatchMiner (5), Onto-Miner (6) from Onto-Tools for non-commercial microarrays, and microarray-provider annotation tools, such as NetAffx (7) for Affymetrix GeneChip arrays, are used to enhance the original gene annotation datasets in ITTACA.

For the clinical data, different anatomo-pathological and clinical characteristics are allowed in ITTACA. The following characteristics are stored: histological information about samples, such as tissue type (normal, tumoral), tumor type (primary or not), TNM classification, staging (G), specific markers [like Estrogen Receptor (ER) or Progesterone Receptor for breast cancers], clinical outcome of patients, such as patient status (alive or dead and causes of death), relapse events and delays and, if available, therapy information. These clinical data were generally obtained from the Supplementary Data or the tables in the publications.

ACCESS AND DATA ANALYSIS

The following describes the steps to follow when using ITTACA and the different analyses available (Figure 1). In ITTACA, the user starts by choosing the publication of interest (Figure 1, part I). Groups of patients can then be defined to be compared or a list of genes of interest can be selected (or both). Groups can be defined either by choosing patients manually or by selecting patients based on clinical parameters, survival time or gene expression thresholds (Figure 1, part II). Once the patient groups and/or list of genes are selected, several analyses can be carried out (Figure 1, part III).

ITTACA can generate Kaplan-Meier survival curves for a set of patient groups and can compare them using the log-rank test, which assesses the significance of the difference in survival. Genes that are differentially expressed between these groups can also be identified using the SAM (Significance Analysis of Microarrays) (8) algorithm. The user supplies an acceptable false discovery rate (FDR) and an ordered list of the most significantly differentially expressed genes is returned.

Alternatively, different statistical tests can be used to assess the differential expression of the user-selected genes (Student's *t*-test or Wilcoxon test). Expression profiles can also be viewed in ITTACA using barplots and expression distribution graphs.

We will illustrate a possible use of ITTACA by taking a real example. We wanted to confirm the results found in the study by West *et al*. (9) using another breast cancer dataset from the study by van't Veer *et al*. (10).

West *et al*. (9) looked for genes differentially expressed in 49 breast tumors between two groups having different ER status (25 ER+ versus 24 ER−) using Affymetrix oligonucleotide arrays containing 7129 genes. The study used standard binary regression models combined with singular value

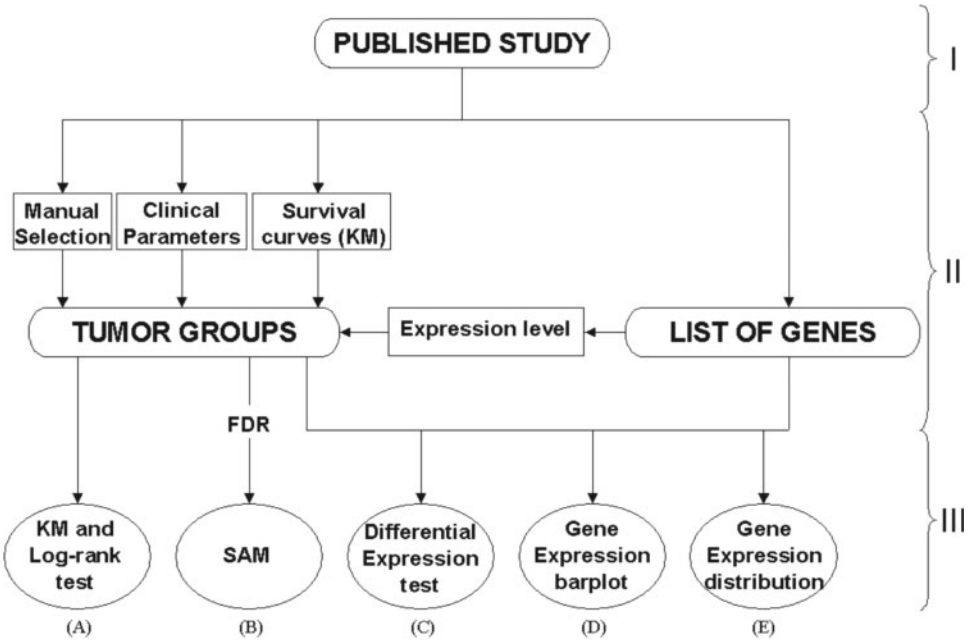

Figure 1. The workflow/outline of ITTACA. The first step (I, upper part of the figure) in ITTACA is the choice of the publication the user wishes to study. Then (step II, middle part of the figure) the user can define groups of tumors or select genes (or both). There are four methods for building tumor groups (represented by a right-angled box): manual choice from the list of tumors, selection based on clinical parameters or on survival time (Kaplan-Meier curve), or on a gene expression threshold. The ovals represent the different analyses that can be carried out with ITTACA (step III, bottom part of the figure): (**A**) KM is the Kaplan-Meier survival curve; it also includes a log-rank test to assess significance of different survival. (**B**) SAM (8) is a statistical method for finding significantly differentially expressed genes in a set of microarray experiments for a given FDR. (**C**) The differential expression tests used by ITTACA are the Wilcoxon (non-parametric) test and the Student's (parametric) *t*-test. ITTACA allows a descriptive analysis of the data, with a gene expression barplot (**D**) and/or the sample's frequency according to the expression level for a gene (**E**) (frequency distribution).

decompositions, and with stochastic regularization using Bayesian analysis. This resulted in a list of 100 differentially expressed genes.

The data from the study by van't Veer *et al*. (10) contain gene expression profiles of primary breast tumors from 115 sporadic young patients (under 55 years old) and 2 familial cases. The study by van't Veer *et al*. (10) aimed to identify a gene signature that was strongly predictive of a short interval to distant metastases in lymph node-negative patients.

The full dataset from van't Veer *et al*. (10) contains the expression of 24 481 genes in 117 samples. From these data, we defined two groups of tumors having different ER status (75 samples ER+ versus 42 samples ER−). Next, we searched for the 100 differentially expressed genes identified by West *et al*. (9); 93 of these were present in the 24 481-gene microarray used in the study by van't Veer *et al*. (10). Among these 93 genes, 88 were found to be significantly differentially expressed between ER+ and ER− groups for a *P*-value threshold of 0.05 using the Wilcoxon test, and 38 genes were also confirmed using the Student's *t*-test (the other genes did not pass the normality test) with the same *P*-value threshold (without correcting for multiple testing). This result, available as Supplementary Data on the ITTACA web site, shows that ITTACA allows a rapid 'interpublication' analysis using different groups of samples than those initially chosen in the original studies.

ITTACA is publicly accessible at http://bioinfo.curie.fr/ittaca. Online documentation and a video tutorial are available on the ITTACA website demonstrating its functionality and guiding users in their analyses.

SYSTEM IMPLEMENTATION

ITTACA is a web-based tool combining a MySQL relational database with a dynamic web interface written in PHP and JavaScript. The website is powered by an Apache server. The website requires a HTML 4.0-compliant browser with JavaScript enabled. It does not require any particular visual plug-in tool.

The data pre-processing uses scripts written in Perl using Perl DBI and Perl xSV packages.

Statistical analyses are performed with the R statistical package (11) (http://www.R-project.org) and Bioconductor (12) for SAM (8).

CONCLUSION

ITTACA is a public database that gathers published clinical and transcriptomic datasets for bladder and breast carcinomas, and uveal melanoma. These data are made available on the Internet for users to reanalyze and study. Several analysis tools allow users to define new patient groups, search for differentially expressed genes, assess differential gene expression, study patient survival time and perform some statistical analyses of gene expression.

Other tools such as Oncomine (13) allow access to public cancer transcriptome data and allow meta-analyses to be carried out on several published datasets. However, the use of Oncomine is restricted to the analysis of groups pre-defined in the publication. GEO (3), Array Express (4) and SMD (14) provide public non-cancer-specific transcriptome data

repositories and are starting to offer some analysis tools online, but they often lack clinical data and are far from exhaustive. Therefore, ITTACA is a complementary tool in research.

The collection of additional relevant microarray datasets is ongoing. ITTACA will be regularly updated with new datasets for breast and bladder carcinomas and uveal melanoma. We intend to be as exhaustive as possible for these three cancers. The ITTACA database is flexible and will accept submissions of datasets for other cancers. New data from pediatric tumors are currently being integrated. We are also planning to make ITTACA compliant with the MIAME standard (1), as well as developing data import and export functionality in MAGE-ML (MicroArray Gene Expression Markup Language) to facilitate exchanges with other repositories.

SUPPLEMENTARY DATA

Supplementary Data are available at http://bioinfo.curie.fr/ittaca/NAR2006SupplementaryData.

ACKNOWLEDGEMENTS

The authors thank the Bioinformatics group of Institut Curie for their support in statistics and computational developments, the Molecular Oncology group of Institut Curie, CNRS UMR 144 for their comments. Funding to pay the Open Access publication charges for this article was provided by Institut Curie.

Conflict of interest statement. None declared.

REFERENCES

1. Brazma,A., Hingamp,P., Quackenbush,J., Sherlock,G., Spellman,P., Stoecker,C., Aach,J., Ansorge,W., Ball,C.A., Causton,H.C. *et al*. (2001) Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nature Genet*., **29**, 365–371.
2. Ball,C.A., Brazma,A., Causton,H., Chervitz,S., Edgar,R., Hingamp,P., Matese,J.C., Parkinson,H., Quackenbush,J., Ringwald,M. *et al*. (2004) Submission of microarray data to public repositories. *PLoS Biol*., **2**, e317.
3. 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**, D562–566.
4. Parkinson,H., Sarkans,U., Shojatalab,M., Abeygunawardena,N., Contrino,S., Coulson,R., Farne,A., Lara,G.G., Holloway,E., Kapushesky,M. *et al*. (2005) ArrayExpress: a public repository for microarray gene expression data at the EBI. *Nucleic Acids Res*., **33**, D553–D555.
5. Bussey,K.J., Kane,D., Sunshine,M., Narasimhan,S., Nishizuka,S., Reinhold,W.C., Zeeberg,B., Ajay,W. and Weinstein,J.N. (2003) MatchMiner: a tool for batch navigation among gene and gene product identifiers. *Genome Biol*., **4**, R27.
6. Khatri,P., Sellamuthu,S., Malhotra,P., Amin,K., Done,A. and Draghici,S. (2005) Recent additions and improvements to the Onto-Tools. *Nucleic Acids Res*., **33**, W762–W765.
7. Liu,G., Loraine,A.E., Shigeta,R., Cline,M., Cheng,J., Valmeekam,V., Sun,S., Kulp,D. and Siani-Rose,M.A. (2003) NetAffx: Affymetrix probesets and annotations. *Nucleic Acids Res*., **31**, 82–86.
8. Tusher,V.G., Tibshirani,R. and Chu,G. (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl Acad. Sci. USA*, **98**, 5116–5121.
9. West,M., Blanchette,C., Dressman,H., Huang,E., Ishida,S., Spang,R., Zuzan,H., Olson,J.A.,Jr, Marks,J.R. and Nevins,J.R. (2001) Predicting the clinical status of human breast cancer by using gene expression profiles. *Proc. Natl Acad. Sci. USA*, **98**, 11462–11467.
10. van't Veer,L.J., Dai,H., van de Vijver,M.J., He,Y.D., Hart,A.A., Mao,M., Peterse,H.L., van der Kooy,K., Marton,M.J., Witteveen,A.T. *et al*. (2002) Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, **415**, 530–536.
11. R Development Core Team (2004) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-00-3.
12. Gentleman,R.C., Carey,V.J., Bates,D.M., Bolstad,B., Dettling,M., Dudoit,S., Ellis,B., Gautier,L., Ge,Y., Gentry,J. *et al*. (2004) Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol*., **5**, R80.
13. Rhodes,D.R., Yu,J., Shanker,K., Deshpande,N., Varambally,R., Ghosh,D., Barrette,T., Pandey,A. and Chinnaiyan,A.M. (2004) ONCOMINE: a cancer microarray database and integrated data-mining platform. *Neoplasia*, **6**, 1–6.
14. Ball,C.A., Awad,I.A., Demeter,J., Gollub,J., Hebert,J.M., Hernandez-Boussard,T., Jin,H., Matese,J.C., Nitzberg,M., Wymore,F. *et al*. (2005) The Stanford Microarray Database accommodates additional microarray platforms and data formats. *Nucleic Acids Res*., **33**, D580–D582.
15. Dyrskjot,L., Thykjaer,T., Kruhoffer,M., Jensen,J.L., Marcussen,N., Hamilton-Dutoit,S., Wolf,H. and Orntoft,T.F. (2003) Identifying distinct classes of bladder carcinoma using microarrays. *Nature Genet*., **33**, 90–96.
16. Gruvberger,S., Ringner,M., Chen,Y., Panavally,S., Saal,L.H., Borg,A., Ferno,M., Peterson,C. and Meltzer,P.S. (2001) Estrogen receptor status in breast cancer is associated with remarkably distinct gene expression patterns. *Cancer Res*., **61**, 5979–5984.
17. Hedenfalk,I., Ringner,M., Ben-Dor,A., Yakhini,Z., Chen,Y., Chebil,G., Ach,R., Loman,N., Olsson,H., Meltzer,P., Borg,A. and Trent,J. (2003) Molecular classification of familial non-BRCA1/BRCA2 breast cancer. *Proc. Natl Acad. Sci. USA*, **100**, 2532–2537.
18. Hedenfalk,I., Duggan,D., Chen,Y., Radmacher,M., Bittner,M., Simon,R., Meltzer,P., Gusterson,B., Esteller,M., Kallioniemi,O.P., Wilfond,B., Borg,A. and Trent,J. (2003) Gene-expression profiles in hereditary breast cancer. *N. Engl. J. Med*., **344**, 539–548.
19. Huang,E., Cheng,S.H., Dressman,H., Pittman,J., Tsou,M.H., Horng,C.F., Bild,A., Iversen,E.S., Liao,M., Chen,C.M., West,M., Nevins,J.R. and Huang,A.T. (2003) Gene expression predictors of breast cancer outcomes. *Lancet*, **361**, 1590–1596.
20. Modlich,O., Prisack,H.B., Pitschke,G., Ramp,U., Ackermann,R., Bojar,H., Vogeli,T.A. and Grimm,M.O. (2004) Identifying superficial, muscle-invasive, and metastasizing transitional cell carcinoma of the bladder: use of cDNA array analysis of gene expression profiles. *Clin. Cancer Res*., **10**, 3410–3421.
21. Onken,M.D., Worley,L.A., Ehlers,J.P. and Harbour,J.W. (2004) Gene expression profiling in uveal melanoma reveals two molecular classes and predicts metastatic death. *Cancer Res*., **64**, 7205–7209.
22. Sorlie,T., Tibshirani,R., Parker,J., Hastie,T., Marron,J.S., Nobel,A., Deng,S., Johnsen,H., Pesich,R., Geisler,S., Demeter,J., Perou,C.M., Lonning,P.E., Brown,P.O., Borresen-Dale,A.L. and Botstein,D. (2003) Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proc. Natl Acad. Sci. USA*, **100**, 8418–8423.
23. Sotiriou,C., Neo,S.Y., McShane,L.M., Korn,E.L., Long,P.M., Jazaeri,A., Martiat,P., Fox,S.B., Harris,A.L. and Liu,E.T. (2003) Breast cancer classification and prognosis based on gene expression profiles from a population-based study. *Proc. Natl Acad. Sci. USA*, **100**, 10393–10398.
24. Tschentscher,F., Husing,J., Holter,T., Kruse,E., Dresen,I.G., Jockel,K.H., Anastassiou,G., Schilling,H., Bornfeld,N., Horsthemke,B., Lohmann,D.R. and Zeschnigk,M. (2003) Tumor classification based on gene expression profiling shows that uveal melanomas with and without monosomy 3 represent two distinct entities. *Cancer Res*., **63**, 2578–2584.