GENT: Gene Expression Database of Normal and Tumor Tissues

Gwangsik Shin²,¹, Tae-Wook Kang³,⁴, Sungjin Yang¹,², Su-Jin Baek³,⁵, Yong-Su Jeong⁶ and Seon-Young Kim³–⁵

¹Department of Bio and Information Technology, Graduate School, Chungbuk National University, 410 Seongbong-ro, Heungdeok-gu, Cheongju, Chungbuk, 361-763, ²NGIC inc. 381 Beonji, Mannyeon-dong, Seo-gu, Daejeon 302-834, ³Medical Genomics Research Center, ⁴Korean Bioinformation Center, ⁵Department of Functional Genomics, University of Science and Technology, KRIBB, ⁶Department of Genetic Engineering, College of Life Science and Graduate School of Biotechnology, Kyung Hee University, Yongin-si, Gyeonggi-do, 446-701, Republic of Korea. #These two authors contributed equally to this work. Corresponding author email: kimsy@kribb.re.kr

Abstract:
Background: Some oncogenes such as ERBB2 and EGFR are over-expressed in only a subset of patients. Cancer outlier profile analysis is one of computational approaches to identify outliers in gene expression data. A database with a large sample size would be a great advantage when searching for genes over-expressed in only a subset of patients.

Description: GENT (Gene Expression database of Normal and Tumor tissues) is a web-accessible database that provides gene expression patterns across diverse human cancer and normal tissues. More than 40000 samples, profiled by Affymetrix U133A or U133plus2 platforms in many different laboratories across the world, were collected from public resources and combined into two large data sets, helping the identification of cancer outliers that are over-expressed in only a subset of patients. Gene expression patterns in nearly 1000 human cancer cell lines are also provided. In each tissue, users can retrieve gene expression patterns classified by more detailed clinical information.

Conclusions: The large samples size (≥24300 for U133plus2 and ≥16400 for U133A) of GENT provides an advantage in identifying cancer outliers. A cancer cell line gene expression database is useful for target validation by in vitro experiment. We hope GENT will be a useful resource for cancer researchers in many stages from target discovery to target validation. GENT is available at http://medical.genome.kribb.re.kr/GENT/ or http://genome.kobic.re.kr/GENT/.

Keywords: gene expression, cancer, human tissues, Affymetrix
Background
Recent examples of successful cancer therapeutics such as Gleevec, Herceptin, and Iressa suggest that the concept of ‘molecular targeted therapy’ is applicable to human cancers of diverse tissue and genetic origin.1 ‘Oncogene addiction’ is a term to describe a phenomenon in which the growth and survival of tumors are impaired by the inactivation of a single oncogene.2 There are several established relationships between genetic alterations and corresponding targeted therapies, and efforts to identify further genetic alterations are underway. Mechanisms of genetic alterations include mutations (EGFR in lung cancer), translocations (BCR-ABL in chronic myeloid leukemia), and gene amplifications (ERBB2 in breast cancer).3

Interestingly, some oncogenes are altered in only a subset of cancer patients. For examples, ERBB2 is amplified and over-expressed in about 25%–30% of breast cancer patients, whereas EGFR is mutated in about 20% of lung cancer patients. Cancer Outlier Profile Analysis (COPA) is a computational method that identifies gene expression profiles that are pathogenically over-expressed in only a subset of patients.3,4 AGTR1 is an example of potential target genes identified by applying the COPA method to the Oncomine database.5

A database with a large sample size is a great advantage when searching for genes over-expressed in only a subset of patients. For example, identifying genes over-expressed in 50 out of 1000 patients is easier and more reliable than identifying genes over-expressed in 2 out of 40 patients. Although the sample size of most individual gene expression studies rarely exceeds one thousand, a data set of nearly ten thousand samples (ie, GeneSapiens database) can be created by a combined analysis of multiple data sets.6 Recent work has shown that analysis of a large microarray data set compiled from many data sets can reveal novel findings that are difficult to observe in the individual studies.7 For a combined analysis, data sets created by the Affymetrix platforms (ie, U133A and U133plus2) offer several advantages. First, most gene expression data sets have been created using the Affymetrix platforms. Second, many data sets are accompanied by raw CEL files so that users can pre-process them as they wish. We have collected human tissue gene expression data sets produced using the Affymetrix U133A and U133Plus2 platforms from public resources, and built a large-scale gene expression database of more than 40,000 samples.

Construction and Content
More than 24300 (U133plus2; 306 data sets) and 16400 (U133A, 241 data sets) samples were collected from public resources, including Gene Expression Omnibus,8 Array Express,9 and Expression Project for Oncology.10 Whenever CEL files were available (288/306 for U133plus2 and 192/241 for U133A), we pre-processed them using the MAS5 algorithm using the affy package.11 We chose the MAS5 algorithm because it is a single-array algorithm in which expression values are independent of other data. We then normalized each sample to a target density of 500. For data sets without CEL files but pre-processed by the MASS algorithm (18/306 for U133plus2 and 49/241 for U133A), we used expression measures downloaded from the web source and normalized them to a target density of 500. We identified samples described in more than one dataset, and cleaned them up from duplications. We then classified each sample according to tissue and disease types (Table 1) based on information given in each dataset. Most samples (~75%) were classified into either cancer or normal, but about 20% of samples were classified into other diseases including neurodegenerative diseases, immune-related diseases, and organ-specific diseases. In each tissue type, we also classified each sample into more detailed clinical subtypes such as estrogen receptor positive breast cancers or high grade serous carcinoma of the ovary etc, as described in the original data source. We also collected expression data for more than 3000 samples comprising nearly 1000 different cancer cell lines across tissues, and processed them using the same method (Table 2). Broad/Sanger Cell Line Project12 and GSK’s cell line project13 provided the most abundant expression data sets. For genes with multiple Affymetrix ids, we calculated the average of the multiple probes. The system implementation is based on an Apache web server, JavaScript and PHP scripts for data processing, Open flash charts and R scripts for image production, and MySQL as a backend database.

Utility and Discussion
GENT can be searched using either a gene symbol or an Affymetrix id. DB search results are presented
international multi-center study showed the effectiveness of Herceptin in patients with ERBB2-positive stomach cancer patients.\textsuperscript{14}

The pattern of ERBB2 expression in diverse cell lines provides interesting information as well. First, similar patterns are observed between cancer cell lines and tissues (Figs. 1 and 2). The ERBB2 is not only over-expressed in a subset of breast, lung, ovarian, and gastric tumor tissues, but it is also over-expressed in a subset of cell lines from those four tissues (Fig. 2). Second, information on its expression in specific cell lines is provided as raw data for convenience during in vitro experiments. Here, HCC1954 or B-BT-474 (breast), SKOV3 (ovary), and NCI-N87 or MKN7 (stomach) are good cell lines for observing the effects of siRNA knock-down of the ERBB2, a first step to
show if it is a critical oncogene (Fig. 2). The information provided by the cell line database allows users to skip laborious RT-PCR steps necessary for selecting cell lines for in vitro experiments.

As a second example, we show a pattern of MET expression in diverse normal and tumor tissues. MET is a tyrosine kinase receptor for hepatocyte growth factor and its mutations and amplifications are associated with papillary renal carcinoma. Again, the over-expression of MET in a subset of renal cancers is clearly shown (Supplementary Fig. 1). Besides, MET is over-expressed in a subset of liver, melanoma, and gastric cancer patients, suggesting that MET can be a target for a subset of liver, melanoma and gastric cancer patients.

Finally, users can search detailed phenotypes in a specific tissue of interest. For example, if a user selects a brain tissue for detailed information, patterns of expression in diverse brain diseases including Alzheimer’s disease, Parkinson’s disease, and subtypes of brain tumors are presented. We provide an example of ovarian data set (GSE12172) in Figure 3. Here, ovarian cancer is further classified into subtypes based on tumor stages. We parsed detailed clinical information given in each data set and provide them in a user-friendly manner.

One may be concerned that the GENT database may present false findings due to noise in the public data because laboratory effects are known to be present in publicly available data sets. We assessed the impact of these effects following Lukk et al’s analyses.7 We selected biological groups (with ten replicates or more) which contain at least two different laboratories. For U133A data sets, we selected 5,089 samples of 92 biological groups produced in 93 laboratories. For each of the biological groups, we computed the average correlation coefficient between the assays from different laboratories within the same group. We also calculated the average correlation coefficient between assays from the same laboratory but belonging to different biological groups. The comparison of the two similarity distributions showed that the biological effects were stronger than the laboratory effects (Fig. 4). We got similar results with the U133Plus2 data sets, too.

Also, we provide three means to help to identify and minimize false positive findings. First, raw data are provided for the searched gene with data sources (gene expression series id) so that users can check laboratory effects themselves. Second, a data set filtering option is provided so that users can include or exclude particular data sets. Finally, in our opinion, providing results separately in two platforms (U133A and U133plus2) is one way to discern between true and noisy data as congruent results between the two platforms are a sign of good data quality (Figs. 1 and 2).

The COPA method was originally developed to identify genomic aberrations (ie, chromosomal translocations such as TMPRSS2-ETV1) by searching for pairs of samples with mutually exclusive outliers. Currently, the Oncomine database implements the

Table 2. The number of cancer cell lines according to tissue types (U133plus2 and U133A).

| Tissue               | U133Plus2 | U133A | Total |
|----------------------|-----------|-------|-------|
| Adrenal gland        | 0         | 2     | 2     |
| Biliary tract        | 0         | 6     | 6     |
| Bladder              | 30        | 40    | 70    |
| Blood                | 229       | 142   | 371   |
| Bone                 | 12        | 32    | 44    |
| Brain                | 149       | 117   | 266   |
| Breast               | 239       | 199   | 438   |
| Cervix               | 23        | 23    | 46    |
| Colon                | 143       | 56    | 199   |
| Connective tissue    | 9         | 0     | 9     |
| Endometrium          | 0         | 11    | 11    |
| Esophagus            | 12        | 25    | 37    |
| EWT                  | 7         | 0     | 7     |
| Eye                  | 5         | 2     | 7     |
| Kidney               | 29        | 40    | 69    |
| Leukemia             | 0         | 4     | 4     |
| Liver                | 33        | 16    | 49    |
| Lung                 | 358       | 325   | 683   |
| Lymphoma             | 73        | 1     | 74    |
| Muscle               | 10        | 0     | 10    |
| Myeloma              | 7         | 24    | 31    |
| Ovary                | 24        | 43    | 67    |
| Pancreas             | 50        | 18    | 68    |
| Pharynx              | 6         | 0     | 6     |
| Placenta             | 9         | 2     | 11    |
| Prostate             | 12        | 18    | 30    |
| Rectum               | 7         | 0     | 7     |
| Sarcoma              | 8         | 0     | 8     |
| Skin                 | 129       | 73    | 202   |
| Soft tissue          | 0         | 19    | 19    |
| Stomach              | 56        | 24    | 80    |
| Testis               | 0         | 4     | 4     |
| Thyroid              | 12        | 13    | 25    |
| Upper aerodigestive  | 0         | 24    | 24    |
| Urinary tract        | 0         | 20    | 20    |
| Vulva                | 9         | 3     | 12    |
| Total                | 1714      | 1336  | 3050  |
COPA method and provides information on possible outliers. Basically, samples in a dataset are grouped by sample properties, centered by a median value and rescaled by median absolute deviation, and COPA score is calculated at multiple percentiles (ie, 1%, 5%, and 10%). Although we adopted COPA concept in the GENT database, we didn’t implement it as suggested in the original paper. Instead, we tried to increase the sample size by collating similar samples across different datasets, so outliers could be detected more reliably. We plan to add the COPA score for each gene in the future. Also, we plan to add additional genomic data such as copy number alterations and genome-wide DNA methylation data to the GENT database so that additional information can be obtained by the integrated analysis of multiple genomic data. We also plan to add gene expression data generated using other platforms (ie, Illumina or Agilent) as more

Figure 1. Pattern of ERBB2 expression across diverse normal and tumor tissues, A) U133plus2 data set, B) U133A data set.
Figure 2. Pattern of ERBB2 expression across diverse cancer cell lines in U133plus2 data. A) Data is shown as multiple boxplots. B) Data is shown as a flash chart so users can identify a cell line name by pointing a mouse on each dot.

Figure 3. Pattern of LAMB3 expression among ovarian cancer subtypes. A) A screenshot of sub-type specific search option. B) Pattern of LAMB3 expression in different stages of ovarian cancer patients from GSE12172 data set.
data accumulate. Finally, we plan to add more functions to provide more extensive clinical information which is currently provided in a limited way.

**Conclusions**

Oncomine and Gene Expression Atlas are two examples of excellent web databases for gene expression information with many useful functions. The two databases provide extensive clinical information given in each collected dataset and many useful search options. To mention a major difference between the two databases, Oncomine focuses on human cancer datasets while Gene Expression Atlas comprises more than 20 organisms including human, mouse, rat, and so on. As those two databases provide rich information, we focused on providing new features for users instead of implementing features already available in those two databases. In our opinion, two aspects of GENT are unique. The first one is the large sample size made possible by the collation of hundreds of datasets into two large datasets, and the second one is the cancer cell line gene expression database which is a convenient tool to select cell lines of interest. The large sample size (>24300 for U133plus2 and >16400 for U133A) of GENT provides an enormous advantage in identifying cancer outliers. A cancer cell line gene expression database is a useful resource for target validation by in vitro experiment. We hope GENT will be a useful resource for cancer researchers in many stages from target discovery to target validation.

**Availability and Requirements**

The GENT database is freely available at http://medical-genome.kribb.re.kr/GENT/ or http://genome.kobic.re.kr/GENT/.
Lists of Abbreviations Used
GENT, Gene Expression database of Normal and Tumor tissues; COPA, Cancer Outlier Profile Analysis.

Acknowledgements
We thank Dr. Sang-Chul Kim for his helpful comments and R codes. This work was supported by grants from Basic Science Research Program through the National Research Foundation of Korea (NRF2010-0008143) funded by the Ministry of Education, Science and Technology (MOEST), National R & D Program for Cancer Control, Ministry for Health and Welfare, Republic of Korea (1020360), and KRIIBB Research Initiative Grant.

Disclosures
This manuscript has been read and approved by all authors. This paper is unique and not under consideration by any other publication and has not been published elsewhere. The authors and peer reviewers report no conflicts of interest. The authors confirm that they have permission to reproduce any copyrighted material.

References
1. Stuart D, Sellers WR. Linking somatic genetic alterations in cancer to therapeutics. *Curr Opin Cell Biol*. 2009;21(2):304–10.
2. Weinstein IB, Joe A. Oncogene addiction. *Cancer Res*. 2008;68(9):3077–80; discussion 3080.
3. MacDonald JW, Ghosh D. COPA—cancer outlier profile analysis. *Bioinformatics*. 2006;22(23):2950–1.
4. Tomlins SA, Rhodes DR, Perner S, et al. Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. *Science*. 2005;310(5748):644–8.
5. Rhodes DR, Ateeq B, Cao Q, et al. AGTR1 overexpression defines a subset of breast cancer and confers sensitivity to losartan, an AGTR1 antagonist. *Proc Natl Acad Sci U S A*. 2009;106(25):10284–9.
6. Kilpinen S, Autio R, Ojala K, et al. Systematic bioinformatic analysis of expression levels of 17,330 human genes across 9,783 samples from 175 types of healthy and pathological tissues. *Genome Biol*. 2008;9(9):R139.
7. Lukk M, Kapushesky M, Nikkila J, et al. A global map of human gene expression. *Nat Biotechnol*. 2010;28(4):322–4.
8. Barrett T, Troup DB, Wilhite SE, et al. NCBi GEO: archive for high-throughput functional genomic data. *Nucleic Acids Res*. 2009;37(Database issue):D885–90.
9. Parkinson H, Kapushesky M, Kolesnikov N, et al. ArrayExpress update—from an archive of functional genomics experiments to the atlas of gene expression. *Nucleic Acids Res*. 2009;37(Database issue):D868–72.
10. Expression Project for Oncology [http://www.integen.org/expo/].
11. Zilliox MJ, Irizarry RA. A gene expression bar code for microarray data. *Nat Methods*. 2007;4(11):911–3.
12. Broad/Sanger Cancer Cell Line Project [http://www.broadinstitute.org/cgi-bin/cancer/publications/pub_paper.cgi?mode=view&paper_id=189].
13. GSK’s Cancer Cell Line Project [https://array.nci.nih.gov/caarray/project/woost-00041/].
14. Petrelli NJ, Winer EP, Brahmer J, et al. Clinical Cancer Advances 2009: major research advances in cancer treatment, prevention, and screening—a report from the American Society of Clinical Oncology. *J Clin Oncol*. 2009;27(35):6052–69.
15. Kapushesky M, Emam I, Holloway E, et al. Gene expression atlas at the European bioinformatics institute. *Nucleic Acids Res*. 2010;38 (Database issue):D690–8.
Supplementary Data

Figure S1. Pattern of MET expression across diverse normal and tumor tissues.

Publish with Libertas Academica and every scientist working in your field can read your article

“I would like to say that this is the most author-friendly editing process I have experienced in over 150 publications. Thank you most sincerely.”

“The communication between your staff and me has been terrific. Whenever progress is made with the manuscript, I receive notice. Quite honestly, I’ve never had such complete communication with a journal.”

“LA is different, and hopefully represents a kind of scientific publication machinery that removes the hurdles from free flow of scientific thought.”

Your paper will be:
- Available to your entire community free of charge
- Fairly and quickly peer reviewed
- Yours! You retain copyright

http://www.la-press.com