Meta-analysis of cancer microarray data reveals signaling pathway hotspots

Pankaj Chopra
Dept. of Electrical Engineering
Columbia University
New York, NY 10027
Email: pc2466@columbia.edu

Jaewoo Kang
Dept. of Computer Science and Engineering
Dept. of Biostatistics
Korea University
Seoul, Korea
Email: kangj@korea.ac.kr

Seung-Mo Hong
Dept. of Pathology
Johns Hopkins Medical Institutions
MD, USA
Email: shong28@jhmi.edu

***Abstract*—Recent studies have shown that the identification of deregulated bio-molecular pathways in cancer may be more important than identification of individual genes through differential expression. Since the same pathway can be deregulated by a different subset of genes, it is critical to study pathways as a whole, rather than focus on individual genes. Most papers on meta-analysis of cancer datasets focus on the identification of a set of individual genes, and not on pathways, protein families and gene ontology terms.**

We have analyzed data from 87 microarray datasets consisting of 5,126 samples and spanning 25 different types of cancer. We have identified 212 KEGG pathways, 578 protein families and 1,717 gene ontology terms that are statistically significant ($p < 0.01$), and deregulated in cancer. Many of the top pathways and proteins from our meta-analysis, e.g., Jak-Stat pathway, Annexin proteins etc. are already known to play a pivotal role in carcinogenesis. To the best of our knowledge this is the largest meta analysis of cancer pathways, protein families and gene ontology terms to date.

***Index Terms*—microarray; cancer; signaling; pathway; meta-analysis; KEGG; PFAM;GO;gene expression;**

I. Introduction

Over the last few years, there has been an explosion in the number of cancer microarray datasets available in public repositories. However, the number of research papers using multiple datasets in their analysis have been limited [1], [2]. Most of these papers have tried to determine a prognostic gene signature. While some papers have attempted to match significantly expressed genes to the Gene Ontology and KEGG pathways, there haven't been many papers that have mapped multiple cancer datasets to KEGG (Kyoto Encyclopedia of Genes and Genomes), PFAM (protein families), or GO (Gene Ontology) [3], [4]. Results from a recent paper show that analyzing pathways may be more important than individual genes, since pathways may be deregulated by a different subset of genes for different individuals [5]. An analysis of these deregulated pathways is critical to increase our understanding of cancer. Mapping and mining multiple mic! roarray datasets for pathways, protein families and gene ontology terms, may yield insights that were not possible by using just one, or at best, a few datasets. Our meta-analysis highlights key pathways that are consistently perturbed across cancer datasets.

The results presented in this paper should be viewed with the knowledge that the fold changes linked to 'over-expression' and 'under-expression' are open to biological interpretation. The meta-analysis also suffers from the drawback that the datasets have been taken from diverse platforms and they contain diverse number of genes, and that the datasets themselves have been normalized separately. Nevertheless, we believe that the global cancer map presented here presents many significant pathways, protein families and gene ontology terms that are unique to specific cancers, and also some that are common across cancer types. We have created global cancer maps for Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways, Protein Families database (PFAM), and Gene Ontology (GO).

II. Background

A. KEGG Cancer Pathways

A KEGG pathway consists of a set of bio-molecular reaction and interaction networks for metabolism and cellular processes. These are manually curated and represent our current knowledge about that pathway. These pathways incorporate products of oncogenes, tumor suppressor genes, and DNA repair genes [6]. A set of fifteen KEGG pathways have been closely identified with various types of cancer. We used this set as a 'gold' standard against which we compared the results of our meta-analysis. An example of a KEGG pathway associated with cancer is the 'mTOR pathway', which is responsible for regulation of protein translation initiation.

B. PFAM database

The PFAM database is a collection of protein families [7]. In general proteins are composed of several functional subunits, which combined and worked as a protein complex. Interaction of several of these subunits provides structural and/or functional diversity. In our analysis, we will determine if a set of genes is related to a particular protein family in a statistically significant way.

C. Gene Ontology

Gene Ontology (GO) is a collection of controlled vocabularies that describe the biology of a gene product [8]. It

978-1-4244-5121-0/09/$25.00 ©2009 IEEE

consists of approximately 20,000 terms arranged in three independent ontologies: Biological Process, Cellular Component, and Molecular function, each represented by a directed acyclic graph (DAG). In our analysis we have used the Biological Process ontology. Gene Ontology has proven to be very important for secondary analysis of microarray expression data, and a wide range of tools have been developed to aid in this analysis [9]. A comprehensive analysis of the available tools is given by Khatri [10].

III. Methods

A. *Microarray datasets*

For the meta-analysis of human cancer microarray datasets, we took a total of 87 datasets, containing 5,126 cancer tissue samples, from 25 different types of cancer, with several microarray platforms. All the microarray data, was downloaded from Gene Expression Omnibus (GEO), and Stanford Microarray Database (SMD) [11]. These datasets covered a wide variety of cancers, across many microarray platforms. To the best of our knowledge this is the largest meta-analysis of microarray datasets to date.

B. *Processing*

Similar to previous studies, we scaled the datasets after determining the log2 values of affymetrix datasets (after setting to 16,000 any value that was more than 16,000 and setting to 10 any value that was less than 10) [2], [12]. For spotted cDNA datasets, we used the log2 ratio between the measured sample and the control sample. Any gene that had more than 25% values missing was discarded from further analysis. A gene was considered to be over-expressed in a dataset if it showed greater than two-fold increase in expression levels in more than 80% samples in that dataset. Similarly, a gene was considered to be under-expressed in a dataset if it showed greater than two-fold decrease in expression levels in more than 80% tumor samples in the dataset. Then we calculated the hypergeometric probability of this set of over (under) expressed genes being associated with a particular KEGG/PFAM/GO term.

For the set of over (or under) expressed genes in a dataset, we evaluated if there are any KEGG/PFAM/GO terms that are over-represented, than would be expected by chance. We evaluate this probability by using the hypergeometric distribution of the genes. The probability of a gene set of size S containing x genes belonging to a particular KEGG/PFAM/GO term, given that the reference dataset of N genes has a total of A genes belonging to that particular KEGG/PFAM/GO term is:

$$Pr\{X = x|N, A, S\} = \frac{\binom{A}{x}\binom{N-A}{S-x}}{\binom{N}{S}}$$

where X is a random variable representing the number of over (or under) expressed genes, that are associated with a particular KEGG/PFAM/GO term. A KEGG/PFAM/GO term is considered *significant* only if it has a p-value less than 0.01.

For each microarray dataset we ordered the significant KEGG/PFAM/GO terms in order of their p-values. We then applied a rank aggregation algorithm to find a consensus ordering for the top significant KEGG/PFAM/GO terms across all the datasets [13]. Rank based aggregation algorithms are often used in comparing and combining ordered lists from different microarray platforms, which may otherwise not be directly comparable.

To further validate our results, we text mined PubMed abstracts for articles containing both the top significant pathway/protein family/ontology terms and cancer. A high number of articles would indicate that the relationship between the pathway/protein family/ontology term and cancer is well studied and documented.

IV. Results

For each dataset, we shortlisted the 'over' and 'under' expressed genes. Then we conducted two sets of analysis. In the first analysis, we determined the significant ontology and pathway terms separately for over-expressed and under-expressed genes. In the second analysis, we treated the set of 'over' and 'under' expressed genes as one set, and found significant pathway and ontology terms for this combined set.

A. *Evaluating over and under expressed genes separately for significant pathway terms*

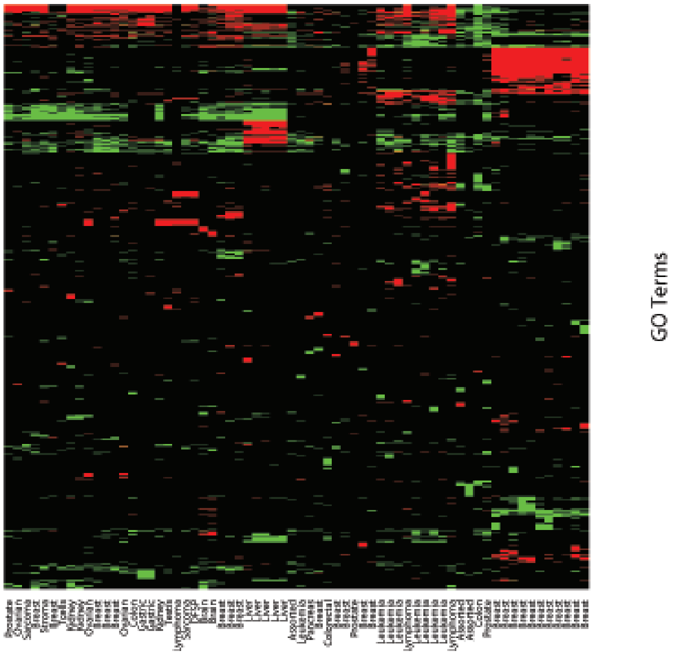

Fig. 1. Meta-analysis of human cancer datasets. The red cells represent significant GO terms associated with 'over-expressed' genes, while the green cells represent significant GO terms associated with 'under-expressed' genes.

The significant GO terms, found by the meta-analysis for all types of human cancers, is shown in Figure 1. The global map clearly shows GO terms that are closely associated with over expressed genes, and GO terms that are closely associated

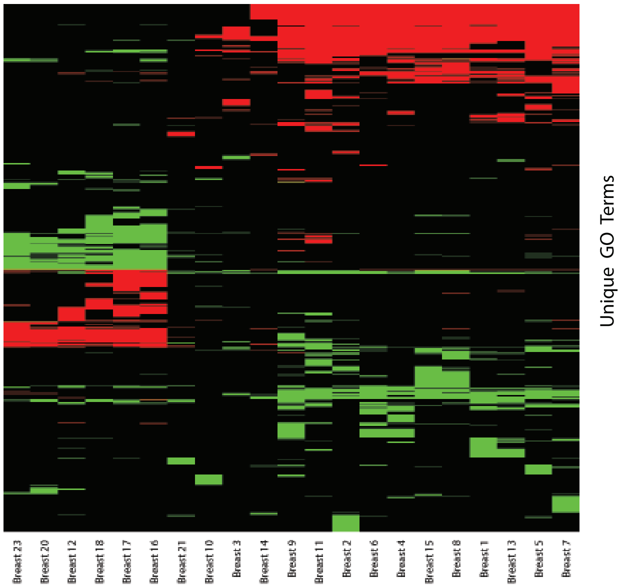

Fig. 2. Meta-analysis of human breast cancer datasets. The red cells represent significant GO terms associated with 'over-expressed' genes, while the green cells represent significant GO terms associated with 'under-expressed' genes.

with under expressed genes. This meta-analysis shows the hotspots that exist in the cancer datasets. These hotspots are most clearly discernible for breast, leukemia and liver cancers.

To determine if some types of cancer are associated with particular GO terms, we conducted a similar analysis using only human breast cancer datasets. The results are shown in Figure 2.

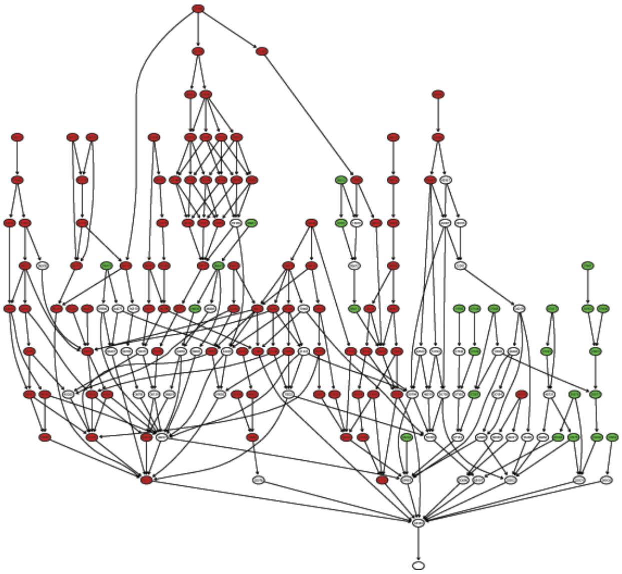

Fig. 3. Significant GO terms mapped to the Gene Ontology tree. Red nodes indicate terms associated with over-expressed genes, and green nodes indicate terms associated with under-expressed genes.

For these breast cancer datasets, we further mapped the significant GO terms to the Gene Ontology tree (Figure 3). This indicates that many over-expressed and under-expressed genes are mapped to separate parts of the Gene Ontology tree, and may be associated with different pathways.

TABLE I
KNOWN CANCER PATHWAYS FOUND BY META-ANALYSIS OF HUMAN MICROARRAY DATASETS, USING THE TOP RANKED 150 TERMS.

Cancer Types	Number of Datasets	Number of tissue samples	Number of known cancer pathways found
All types	87	5,126	11
Breast	23	1,933	12
Leukemia	8	447	14
Kidney	5	267	9
Liver	6	498	13
Prostate	6	213	13

TABLE II
TOTAL NUMBER OF SIGNIFICANT ($P < 0.01$) KEGG PATHWAYS, PROTEIN FAMILIES (PFAM), AND GENE ONTOLOGY (GO) TERMS FOUND.

Cancer Types	Number of KEGG pathways	Number of PFAM	Number of GO terms
All types	212	578	1,717
Breast	210	287	877
Leukemia	205	118	515
Kidney	205	109	303
Liver	197	88	281
Prostate	205	84	319

B. Evaluating over and under expressed genes together for significant pathway terms

In our second analysis, we treated the 'over-expressed' and 'under-expressed' genes as one set. We then determined the significant pathway and ontology terms associated with this gene set, as previously described. Combining the two sets of genes is also biologically consistent because many pathways are deregulated by a combination of over-expressed and under-expressed genes.

For each human microarray dataset we ordered the significant KEGG/PFAM/GO terms in order of their p-values. We then applied a rank aggregation algorithm to find a consensus ordering for the significant KEGG/PFAM/GO terms across all the datasets [13]. From the rank ordered list of significant KEGG, PFAM, and GO terms we analyzed the top 50, 100, and 150 significant terms. The results from the top 150 ranked significant terms are presented here.

To validate our result, we compared the ranked top 150 list, obtained from datasets of all cancer types, to the list of fifteen pathways already known to be closely associated with cancer. We also determined the ranked top 150 terms from all the breast, leukemia, kidney, liver, and prostate cancer datasets, and compared these ranked lists with the 'gold' standard KEGG pathways. In all the cases, a significant proportion of the fifteen pathways were discovered by our method. The results are shown in Table I. The total number of significant KEGG/PFAM/GO terms for various cancer types is given in Table II.

To further corroborate our ranked list, we text mined PubMed for any articles that had both the ontology term/pathway and 'cancer'. Text mining is increasingly being

978-1-4244-5121-0/09/$25.00 ©2009 IEEE

used in life sciences research to keep abreast of current knowledge on a gene or protein. It is being used not only to keep the biological databases up to date but also to plan experiments and validate their results [14], [15]. The concept of using co-occurrence to indicate association has been used extensively in medical literature [16], [17].

TABLE III
TOP FIVE RANKED SIGNIFICANT KEGG TERMS USING DATASETS FROM ALL CANCER TYPES. THE SIGNIFICANT KEGG TERMS OBTAINED FROM EACH DATASET WERE ORDERED (BY INCREASING P-VALUE), AND THEN A RANK AGGREGATION ALGORITHM WAS USED TO COMBINE THESE ORDERED LISTS TO GET THE TOP RANKED KEGG TERMS FOR THE DATASETS

KEGG ID	KEGG Path	Abstracts ('cancer')
00400	Phenylalanine, tyrosine and tryptophan biosynthesis	3,201
00290	Valine, leucine and isoleucine biosynthesis	790
04630	Jak-STAT signaling pathway	328
00720	Reductive carboxylate cycle (CO2 fixation)	29
00040	Pentose and glucuronate interconversions	642

TABLE IV
TOP FIVE RANKED SIGNIFICANT PFAM TERMS USING DATASETS FROM ALL CANCER TYPES. THE SIGNIFICANT PFAM TERMS OBTAINED FROM EACH DATASET WERE ORDERED (BY INCREASING P-VALUE), AND THEN A RANK AGGREGATION ALGORITHM WAS USED TO COMBINE THESE ORDERED LISTS TO GET THE TOP RANKED PFAM TERMS FOR THE DATASETS.

PFAM ID	Protein Family	Abstracts ('cancer')
PF00047	Immunoglobulin domain	2,834
PF00191	Annexin	2,928
PF00131	Metallothionein	1,379
PF07654	Immunoglobulin C1-set domain	239
PF00061	Lipocalin / cytosolic fatty-acid binding protein family	242

TABLE V
TOP FIVE RANKED SIGNIFICANT GO TERMS USING DATASETS FROM ALL CANCER TYPES. THE SIGNIFICANT GO TERMS OBTAINED FROM EACH DATASET WERE ORDERED (BY INCREASING P-VALUE), AND THEN A RANK AGGREGATION ALGORITHM WAS USED TO COMBINE THESE ORDERED LISTS TO GET THE TOP RANKED GO TERMS FOR THE DATASETS

GOID	GO Term	Abstracts ('cancer')
GO:0006412	Translation	10,000
GO:0002009	Morphogenesis of an epithelium	1,072
GO:0009260	Ribonucleotide biosynthetic process	827
GO:0000087	M phase of mitotic cell cycle	593
GO:0055085	Transmembrane transport	865

If a credible link between a particular ontology term/pathway and cancer exists, then there would be many articles in PubMed on the subject. The ranked top five KEGG/PFAM/GO terms from our analysis, and the number of PubMed abstracts with that particular term and 'cancer', are shown in Tables V, III, and IV. For example, the third row in Table III indicates that there are 328 abstracts in PubMed that contain both 'Jak-STAT signaling pathway' (a 'gold' standard pathway) and 'cancer'.

TABLE VI
TOP FIVE SIGNIFICANT KEGG TERMS USING DATASETS FROM BREAST CANCER

KEGG ID	KEGG Path	Abstracts ('cancer')	Abstracts ('breast cancer')
00230	Purine metabolism	10,000	1,157
00010	Glycolysis / Gluconeogenesis	2,464	149
00360	Phenylalanine metabolism	4,090	194
00220	Urea cycle and metabolism of amino groups	900	51
00510	N-Glycan biosynthesis	145	9

TABLE VII
TOP FIVE RANKED SIGNIFICANT PFAM TERMS USING DATASETS FROM BREAST CANCER

PFAM ID	Protein Family	Abstracts ('cancer')	Abstracts ('breast) cancer')
PF00227	Proteasome A-type and B-type	131	8
PF01454	MAGE family	175	22
PF07654	Immunoglobulin C1-set domain	239	42
PF00055	Laminin N-terminal (Domain VI)	53	4
PF00458	WHEP-TRS domain	26	1

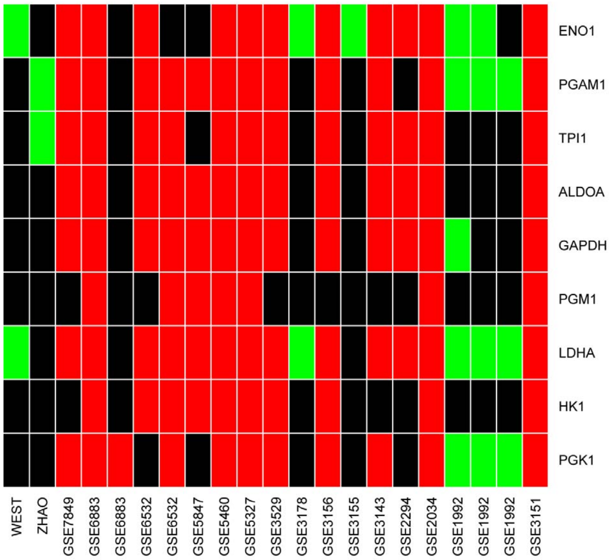

Fig. 4. Over-expressed genes, associated with the glycolysis pathway (ref. Table VI, are indicated by the color red. Each column represents a breast cancer dataset in which the glycolysis pathway was deregulated

We conducted a similar analysis for human breast cancer [1] (Tables VIII, VI, and VII), and leukemia (Tables XI, IX, and X). For this analysis, we also mined PubMed for number of abstracts that had the ontology/pathway term and 'breast cancer' (or 'leukemia','kidney'...). This is indicated in column 4 in the tables. The large number of articles in PubMed indicate that the ranked ontology/pathway terms are also closely associated with the corresponding cancer type.

[1] The number of PubMed abstracts were capped at 10,000

TABLE VIII
TOP FIVE RANKED SIGNIFICANT GO TERMS USING DATASETS FROM BREAST CANCER

GO ID	GO Term	Abstracts ('cancer')	Abstracts ('breast cancer')
GO:0007599	Hemostasis	6,583	315
GO:0009611	Response to wounding	5,632	515
GO:0065008	Regulation of biological quality	3,369	525
GO:0051325	Interphase	10,000	1,395
GO:0000398	Nuclear mRNA splicing via spliceosome	680	75

TABLE IX
TOP FIVE RANKED SIGNIFICANT KEGG TERMS USING DATASETS FROM LEUKEMIA

KEGG ID	KEGG Path	Abstracts ('cancer')	Abstracts ('leukemia')
00020	Citrate cycle (TCA cycle)	441	30
00340	Histidine metabolism	1,606	201
00100	Biosynthesis of steroids	10,000	2,629
00590	Arachidonic acid metabolism	2,652	464
00230	Purine metabolism	10,000	5,144

TABLE X
TOP FIVE RANKED SIGNIFICANT PFAM TERMS USING DATASETS FROM LEUKEMIA

PFAM ID	Protein Family	Abstracts ('cancer')	Abstracts ('leukemia')
PF00969	Class II histocompatibility antigen, beta domain	4,185	821
PF00074	Pancreatic ribonuclease	539	75
PF00219	Insulin-like growth factor binding protein	3,341	149
PF00079	Serpin (serine protease inhibitor)	5,367	456
PF00047	Immunoglobulin domain	2,834	366

TABLE XI
TOP FIVE RANKED SIGNIFICANT GO TERMS USING DATASETS FROM LEUKEMIA

GO ID	GO Term	Abstracts ('cancer')	Abstracts ('leukemia')
GO:0048513	Organ development	4,795	404
GO:0048869	Cellular developmental process	805	120
GO:0007155	Cell adhesion	10,000	2777
GO:0007596	Blood coagulation	6,718	1,465
GO:0007154	Cell communication	8,552	944

V. DISCUSSION

In our cancer map, we have shown that there are many pathway 'hotspots' associated with over-expressed and under-expressed genes. Even for a particular type of cancer (e.g., breast cancer) there are distinct pathways/ontology terms that are closely associated with over and under expressed genes. Many of the top ranked pathways found by our analysis are already closely associated with cancer.

In a recent genome-wide study of glioblasoma, the deregulation of critical cancer related pathways (TP53, RB1, PI3K/PTEN, ...) was compared to the deregulation of their associated genes [5]. It was found that a different subset of genes, associated with the same pathway, resulted in the pathway being deregulated. In fact, each subset consisted of mutually exclusive genes, i.e., for the same deregulated pathway, different patients had a different subset of genes deregulated. This result supports our view that approaching the pathway, rather than each single gene in a pathway, may provide for a more comprehensive understanding of the biology of cancer.

One of the top 5 ranked significant PFAM terms from all cancer data sets is the annexin family (Table IV). Annexin super-family is calcium binding membranous proteins which consist of 13 members from annexin 1 to annexin13. They involve in diverse cellular functions in normal status, which include apoptosis cellular division, vesicular trafficking, calcium signaling, and growth regulations. In carcinogenesis, each annexin provides different roles [18]. For example, annexin 2 promotes angiogenesis, tumor progression, and metastasis by working as receptor proteins for plasminogen and tPA. Annexin expression is changed based on specific types of cancers [19]. Annexin 1 expression is increased in adenocarcinomas from esophagus, stomach, colon, lung and pancreas and hepatocellular carcinomas, on the other had Annexin 1 expression is decreased in squamous carcinomas from esophagus and head and neck, breast invasive ductal carcinomas, ! and prostatic adenocarcinomas [20]. The link between Jak/Stat signaling pathway (Table III) and cancer has already been extensively studied [21]–[24].

Similarly, one of the top 5 ranked significant KEGG terms from breast cancer is 'glycolysis/gluconeogenesis' (Table VI). It is well known that up regulation of aerobic glycolysis ('Warburg effect') is one of the characteristics of cancers [25]. Although, aerobic glycolysis is an inefficient way of generating ATP (adenosine 5'-triphosphate) compared to oxidative phosphorylation, it provides not only energy but also several other compounds, such as nucleotides, amino acids, and lipids, to cancer cells. These provide nutrients for the production of new cells [26]. In the present study, we observed several genes in glycolysis pathway, including ENO1, PGAM1, TPI1, ALDOA, GAPDH, PGM1, LDHA, HK1, and PGK1 that were up-regulated (Figure 4) in a significant number of breast cancer datasets.

This meta-analysis can serve as an exploratory tool for narrowing down the search space for oncological terms and pathways related to cancer. From the thousands of ontological terms, this meta-analysis reveals hotspots that need to be examined further in a wet lab environment. This higher level perspective will help in a better understanding of oncogenesis and metastasis of cancer.

VI. CONCLUSION

We conducted a meta-analysis of cancer datasets and analyzed the regulatory pathways that are perturbed by over and under expressed genes. To the best of our knowledge, this meta-analysis with eighty seven human cancer microarray datasets and across twenty five cancer types is the largest such study. We have shown that the meta-analysis picked out pathways already closely associated with cancer, but also

978-1-4244-5121-0/09/$25.00 ©2009 IEEE

implicated other pathways that may be closely associated with cancer. We corroborated our results by text mining PubMed. Significant ontology and pathways terms found by our analysis have been studied by researchers, and considerable literature exists linking cancer to the significant ontology and pathway terms that were found by our analysis. This meta-analysis has revealed 'hotspots' in pathways and gene ontology terms that may be causally linked to cancer. Future works could focus on these hotspots to further enhance our understanding of the deregulated pathways that are crucial to the cause and progression of cancer.

Acknowledgements

Correspondence should be addressed to Kang. This work was partially supported by the Second Brain Korea 21 Project Grant, the Korea Research Foundation Grant funded by the Korean Government(MOEHRD, Basic Research Promotion Fund)(KRF-2008-331-D00481), the Korea Science and Engineering Foundation(KOSEF) grant funded by the Korea government(MEST)(R01-2008-000-20564-0), and the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (2009-0077688).

References

[1] J. P. de Magalhaes, J. Curado, and G. M. Church, "Meta-analysis of age-related gene expression profiles identifies common signatures of aging," *Bioinformatics*, vol. 25, no. 7, pp. 875–881, 2009.

[2] E. Segal, N. Friedman, D. Koller, and A. Regev, "A module map showing conditional activity of expression modules in cancer." *Nat Genet*, vol. 36, no. 10, pp. 1090–8, Oct 2004. [Online]. Available: 10.1038/ng1434

[3] T. Werner, "Bioinformatics applications for pathway analysis of microarray data," *Current Opinion in Biotechnology*, vol. 19, no. 1, pp. 50 – 54, 2008, analytical biotechnology. [Online]. Available: http://www.sciencedirect.com/science/article/B6VRV-4RN489C-1/2/628da2c300e0dd20dba7e043b06fa3d4

[4] M. Pas, S. V. Hemert, and B. Hulsegge, "A pathway analysis tool for analyzing microarray data of species with low physiological information," *Advances in Bioinformatics*, vol. 2008, no. 1, 2008. [Online]. Available: http://www.sciencedirect.com/science/article/B6VRV-4RN489C-1/2/628da2c300e0dd20dba7e043b06fa3d4

[5] D. W. Parsons, S. Jones, X. Zhang, J. C.-H. Lin, R. J. Leary, P. Angenendt, P. Mankoo, H. Carter, I.-M. Siu, G. L. Gallia, A. Olivi, R. McLendon, B. A. Rasheed, S. Keir, T. Nikolskaya, Y. Nikolsky, D. A. Busam, H. Tekleab, J. Diaz, Luis A., J. Hartigan, D. R. Smith, R. L. Strausberg, S. K. N. Marie, S. M. O. Shinjo, H. Yan, G. J. Riggins, D. D. Bigner, R. Karchin, N. Papadopoulos, G. Parmigiani, B. Vogelstein, V. E. Velculescu, and K. W. Kinzler, "An Integrated Genomic Analysis of Human Glioblastoma Multiforme," *Science*, vol. 321, no. 5897, pp. 1807–1812, 2008.

[6] M. Kanehisa, M. Araki, S. Goto, M. Hattori, M. Hirakawa, M. Itoh, T. Katayama, S. Kawashima, S. Okuda, T. Tokimatsu, and Y. Yamanishi, "Kegg for linking genomes to life and the environment," *Nucl. Acids Res.*, pp. gkm882+, December 2007. [Online]. Available: http://dx.doi.org/10.1093/nar/gkm882

[7] R. D. Finn, J. Tate, J. Mistry, P. C. Coggill, S. J. Sammut, H.-R. Hotz, G. Ceric, K. Forslund, S. R. Eddy, E. L. L. Sonnhammer, and A. Bateman, "The Pfam protein families database," *Nucl. Acids Res.*, vol. 36, no. 1, pp. D281–288, 2008.

[8] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock, "Gene ontology: tool for the unification of biology. the gene ontology consortium." *Nat Genet*, vol. 25, no. 1, pp. 25–29, May 2000.

[9] J. Liu, W. Wang, and J. Yang, "A framework for ontology-driven subspace clustering," in *KDD '04: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, 2004, pp. 623–628.

[10] P. Khatri and S. Draghici, "Ontological analysis of gene expression data: current tools, limitations, and open problems," *Bioinformatics*, vol. 21, no. 18, pp. 3587–3595, 2005.

[11] T. Barrett, D. B. Troup, S. E. Wilhite, P. Ledoux, D. Rudnev, C. Evangelista, I. F. Kim, A. Soboleva, M. Tomashevsky, and R. Edgar, "Ncbi geo: mining tens of millions of expression profiles–database and tools update," *Nucl. Acids Res.*, vol. 35, pp. D760–765, 2007.

[12] P. Wirapati, C. Sotiriou, S. Kunkel, P. Farmer, S. Pradervand, B. Haibe-Kains, C. Desmedt, M. Ignatiadis, T. Sengstag, F. Schutz, D. Goldstein, M. Piccart, and M. Delorenzi, "Meta-analysis of gene expression profiles in breast cancer: toward a unified understanding of breast cancer subtyping and prognosis signatures," *Breast Cancer Research*, vol. 10, no. 4, p. R65, 2008. [Online]. Available: http://breast-cancer-research.com/content/10/4/R65

[13] V. Pihur, S. Datta, and S. Datta, "Rankaggreg, an r package for weighted rank aggregation," *BMC Bioinformatics*, vol. 10, no. 1, p. 62, 2009. [Online]. Available: http://www.biomedcentral.com/1471-2105/10/62

[14] J. Baumgartner, William A., K. B. Cohen, L. M. Fox, G. Acquaah-Mensah, and L. Hunter, "Manual curation is not sufficient for annotation of genomic databases," *Bioinformatics*, vol. 23, no. 13, pp. i41–48, 2007. [Online]. Available: http://bioinformatics.oxfordjournals.org/cgi/content/abstract/23/13/i41

[15] M. Krallinger, A. Valencia, and L. Hirschman, "Linking genes to literature: text mining, information extraction, and retrieval applications for biology," *Genome Biology*, vol. 9, no. Suppl 2, p. S8, 2008. [Online]. Available: http://genomebiology.com/2008/9/S2/S8

[16] H. Yu, V. Hatzivassiloglou, C. Friedman, A. Rzhetsky, and J. W. Wilbur, "Automatic extraction of gene and protein synonyms from medline and journal articles," in *Proc. AMIA Symp*, vol. 23, 2002, pp. 919–923. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.60.6455

[17] A. Cohen, W. Hersh, C. Dubay, and K. Spackman, "Using co-occurrence network structure to extract synonymous gene and protein names from medline abstracts," *BMC Bioinformatics*, vol. 6, no. 1, p. 103, 2005. [Online]. Available: http://www.biomedcentral.com/1471-2105/6/103

[18] M. J. Hayes and S. E. Moss, "Annexins and disease," *Biochemical and Biophysical Research Communications*, vol. 322, no. 4, pp. 1166 – 1170, 2004. [Online]. Available: http://www.sciencedirect.com/science/article/B6WBK-4D2FHTM-8/2/c08a3bacfe60ab661e1c865c2608b84b

[19] S. Mussunoor and G. I. Murray, "The role of annexins in tumour development and progression," *The Journal of Pathology*, vol. 216, no. 2, pp. 131–140, 2008.

[20] M. C. Sharma and M. Sharma, "The role of annexin ii in angiogenesis and tumor progression: a potential therapeutic target," *Curr Pharm Des.*, vol. 13, no. 35, pp. 3568–75, 2007.

[21] V. Boudny and J. Kovarik, "Jak/stat signaling pathways and cancer. janus kinases/signal transducers and activators of transcription." *Neoplasma*, vol. 49, no. 6, pp. 349 – 355, 2002.

[22] J.-P. Spano, G. Milano, C. Rixe, and R. Fagard, "Jak/stat signalling pathway in colorectal cancer: A new biological target with therapeutic implications," *European Journal of Cancer*, vol. 42, no. 16, pp. 2668 – 2670, 2006. [Online]. Available: http://www.sciencedirect.com/science/article/B6T68-4KVXHHH-4/2/56cce47d1e203d1fd368f646273484bf

[23] S. N. Constantinescu, M. Girardot, and C. Pecquet, "Mining for jak-stat mutations in cancer." *Trends in biochemical sciences*, vol. 33, no. 3, pp. 122–131, March 2008. [Online]. Available: http://dx.doi.org/10.1016/j.tibs.2007.12.002

[24] J. R. Caceres-Cortes, "A potent anti-carcinoma and anti-acute myeloblastic leukemia agent, ag490." *Anti-cancer agents in medicinal chemistry*, vol. 8, no. 7, pp. 717–722, October 2008. [Online]. Available: http://view.ncbi.nlm.nih.gov/pubmed/18855573

[25] R. A. Gatenby and R. J. Gillies, "Why do cancers have high aerobic glycolysis?" *Nat Rev Cancer*, vol. 4, no. 11, pp. 891–899, November 2004. [Online]. Available: http://dx.doi.org/10.1038/nrc1478

[26] Vander, L. C. Cantley, and C. B. Thompson, "Understanding the warburg effect: The metabolic requirements of cell proliferation," *Science*, vol. 324, no. 5930, pp. 1029–1033, May 2009. [Online]. Available: http://dx.doi.org/10.1126/science.1160809

978-1-4244-5121-0/09/$25.00 ©2009 IEEE