BMC Genomics

Software

Literature Lab: a method of automated literature interrogation to infer biology from microarray analysis

Phillip G Febbo*[1,2,3], Mike G Mulligan[4], David A Slonina[4], Kimberly Stegmaier[5], Dolores Di Vizio[6], Paul R Martinez[4], Massimo Loda[6] and Stephen C Taylor[4]

Address: [1]Institute for Genome Science and Policy, Duke University, Durham, North Carolina, USA, [2]Department of Medicine, Duke University School of Medicine, Durham, North Carolina, USA, [3]Department of Molecular Genetics and Microbiology, Duke University School of Medicine, Durham, North Carolina, USA, [4]Acumenta Corporation, Boston, Massachusetts, USA, [5]Department of Pediatric Oncology, Dana-Farber Cancer Institute, Harvard Medical School, Boston, Massachusetts, USA and [6]Department of Pathology, Dana-Farber Cancer Institute, Harvard Medical School, Boston, Massachusetts, USA

Email: Phillip G Febbo* - phil.febbo@duke.edu; Mike G Mulligan - mmulligan@acumenta.com; David A Slonina - dslonina@acumenta.com; Kimberly Stegmaier - kimberly_stegmaier@dfci.harvard.edu; Dolores Di Vizio - dolores.divizio@childrens.harvard.edu; Paul R Martinez - pmartinez@acumenta.com; Massimo Loda - massimo_loda@dfci.harvard.edu; Stephen C Taylor - staylor@acumenta.com

* Corresponding author

Published: 18 December 2007

BMC Genomics 2007, **8**:461 doi:10.1186/1471-2164-8-461

Received: 12 July 2007
Accepted: 18 December 2007

This article is available from: http://www.biomedcentral.com/1471-2164/8/461

Abstract

Background: The biomedical literature is a rich source of associative information but too vast for complete manual review. We have developed an automated method of literature interrogation called "Literature Lab" that identifies and ranks associations existing in the literature between gene sets, such as those derived from microarray experiments, and curated sets of key terms (i.e. pathway names, medical subject heading (MeSH) terms, etc).

Results: Literature Lab was developed using differentially expressed gene sets from three previously published cancer experiments and tested on a fourth, novel gene set. When applied to the genesets from the published data including an *in vitro* experiment, an *in vivo* mouse experiment, and an experiment with human tumor samples, Literature Lab correctly identified known biological processes occurring within each experiment. When applied to a novel set of genes differentially expressed between locally invasive and metastatic prostate cancer, Literature Lab identified a strong association between the pathway term "FOSB" and genes with increased expression in metastatic prostate cancer. Immunohistochemistry subsequently confirmed increased nuclear FOSB staining in metastatic compared to locally invasive prostate cancers.

Conclusion: This work demonstrates that Literature Lab can discover key biological processes by identifying meritorious associations between experimentally derived gene sets and key terms within the biomedical literature.

Background

The accelerating expansion of biomedical research outpaces most individual attempts at comprehensive review even in relatively narrow fields. Just as the vast sequence data available for the human [1,2] and additional organisms [3-5] require sophisticated genomic browsing tools

[6-8], computational methods are required to thoroughly explore the corpus of biomedical literature. Many computational methods for interrogating the scientific literature have been developed [9]. These programs can be broadly defined as methods for information retrieval and those for information extraction [10].

Existing methods can identify significant association between individual genes and terms from the medical subject heading (MeSH) index and Gene ontology (GO) databases [11], manually curated biological lists [12], or disease-specific lists [13]. These prior methods demonstrate that genes with disease-specific differential expression can be strongly correlated with key terms within the medical literature [13,14]. In addition, gene-gene associations within the literature have been combined with multiple available databases to extend associations beyond the literature alone [15].

While these and similar approaches have underscored the potential of automated literature searching to facilitate discovery, few have provided both methods for assessing the statistical strength of identified associations and supported their methods with experimental validation. Here, we describe and apply "Literature Lab", a method of automated data retrieval confined to publicly available citations and abstracts. Literature Lab statistically assesses identified associations within the corpus of medical literature between sets of experimentally derived genes and key terms derived from curated or MeSH lists. We demonstrate that our methodology can identify previously reported relationships and can result in discovery.

Molecular Methods

Gene Expression Sets

Literature lab was applied to three gene sets derived from previously published microarray data and one gene set from as yet, unpublished dataset to determine if literature mining identified important metabolic, physiologic, or pathway activity. These gene sets include, 1) the top 100 genes with increased expression during *in vitro* exposure of human leukemia cells (HL60, AML cell line) to ATRA ("UCT"), 2) the genes differentially regulated with a transgenic model of prostate neoplasia (MPAKT) following exposure to RAD001, 3) the 70 genes used to predict outcome in localized node-negative breast cancer, and 4) genes differentially expressed between malignant epithelial cells in local v. metastatic prostate cancers. The genes lists are included in additional files and a full description of the development of each gene set is in the supplemental methods (see Additional File 1).

Western Blots

Protein from fresh ventral prostates were extracted in RIPA buffer [10 mM sodium phosphate (pH7.2) 150 mM NaCl, 1% Nonidet P-40, 0.1% SDS, 1 mM NaOVa, 1 mM DTT, 5 mM NaF, 0.1% sodium deoxycholate, 10 μg/ml leupeptin, 10 μg/ml aprotinin, and 1 mM PMSF] separated by gel electrophoresis and transferred to nitrocellulose membrane (0.45 μM) as described[16,17]. Membranes were blotted with anti-Hif-1α (kindly provided by J. Pouyssegur) and anti-tubulin (B-5-1-2) (Sigma) (1:1000). Blots were scanned and intensities were measured.

Immunohistochemistry

A prostate tissue microarray containing samples of benign prostate epithelium (n = 14), locally invasive prostate adenocarcinoma (n = 20), and metastatic prostate adenocarcinoma (n = 22) were stained for FOSB. The FOSB staining was performed as previously described [18,19]. Briefly, 5 micrometer sections were cut from the TMA block, deparaffinized, rehydrated and subjected to microwaving in 10-mM Citrate buffer (pH 6.00) in a 750W oven, for 15 minutes. The polyclonal anti-FOSB primary antibody (Santa Cruz Biotechnologies, Inc.), was incubated (1:50 dilution) at room temperature in an automated stainer (Optimax Plus 2. 0; Biogenex, San Ramon, CA). Antigen-antibody reaction was revealed with standardized development times, using the Streptavidin method with 3, 3 diaminobenzidine as substrate. Meyer Hematoxylin was used as nuclear counterstaining. FOSB nuclear positivity was scored on a total of 50 nuclei per sample. The statistical difference in FOSB staining between the unpaired populations was assessed using a two-tailed Mann Whitney test (GraphPad Prism4 Software).

Application of Established Literature Mining Tools

MeSHer data for the gene lists was obtained from the website [20] by supplying the Affymetrix U133Plus2 Ids corresponding to the list of genes using a p-value threshold of 1.0 and no p-value correction.

GOMiner data for the gene lists was obtained by following the instructions for downloading (build 148) and installing the GOMiner application and SQL database from [21]. A list of gene symbols for genes in the Affymetrix U133Plus2 probe set list was used for the 'Total' file and lists of symbols for the experimental genes used for the 'Changed' file. Other GOMiner settings used were the default values defined in the GOMiner application.

Implementation

Overview

To highlight and prioritize cellular physiology, metabolism, or pathways differentially active within a microarray experiment, Literature Lab uses an experimentally derived gene list, pre-defined sets of non-ambiguous key terms ("domains"), inclusive gene nomenclature, comprehensive literature interrogation, and a comparison of the

experimental gene set results with randomly generated gene sets. For each experimentally derived gene set, Literature Lab performs an automated literature search to determine the number of abstracts listing any of the identified genes and each term within the term list (or domain). Specific domains include those for cell metabolism (MeSH headings), cell physiology (MeSH headings), and cellular pathways (manually curated from multiple sources) (Additional File 2). Terms are ranked with a measure of association between genes and terms known as the product of frequency (PF) (see Additional File 1). The strength of the relationship between an experimental gene set and a term is determined by comparing the log (base 10) of the sum of the PF values for the experimental gene set (called the LPF) with a distribution of the same values for 1000 random gene sets of the same size as the experimental gene set. A score representing the number of standard deviations above or below the mean value for the random gene sets is assigned to each term and terms are ranked in decreasing order of this score. To highlight terms with particularly strong association with the gene list, we developed heuristic rules for labeling the relationships as "Strong" or "Moderate" (see Additional File 1 for Details) (the software used for these analyses is free for non-commercial use and available, see Availability and Results section).

Overall Architecture of the Software Implementation

The software is implemented using Sun's Java programming language [22] using the Eclipse software development environment [23]. The software consists of three major components:

• A series of programs which query PubMed for each term and each gene in the Gene Thesaurus using NCBI's Entrez Programming Utilities [24] and Sun's JAXB XML Binding classes [25]. The result for each term or gene is a file containing a list of the PubMed Ids for the term or gene. In addition, summary files containing counts of the number of abstracts for each gene and for each term are prepared.

• A series of programs which identify the PubMed Ids in common for each term/gene combination from the lists of ids for each term and each gene. The results of these programs are files containing the PubMed Ids for each term/gene combination and files containing summary counts of the number of abstracts for each term/gene combination.

• A series of programs for preparing the analysis of a specific gene list. These programs compute the described statistics for the gene list and for one thousand random gene sets of the same size as the specific gene list being processed. The final step of this process prepares a Microsoft® Excel spreadsheet containing the results of the analysis. The Apache Software Foundation's POI classes (Java API to Access Microsoft® Format Files [26] are used to format the data in a form which can be viewed in Microsoft® Excel.

Gene Annotation

Gene annotation began by obtaining gene names, symbols, and "aliases" from Stanford University's SOURCE web site [27]. Subsequent Boolean PubMed searches and manual review for precision (i.e. the ability to accurately identify abstracts related to the gene of interest) were used to develop a set of rules required for more specific automated searching (Additional File 1). In subsequent literature searches, gene terms consisting entirely of numerals, of three characters or less, or that were identified as excessively ambiguous through manual review were excluded. Algorithm-based disambiguation was not used [28]. While the database of gene terms (referred to as the gene thesaurus) is frequently updated, for all experiments herein described, the gene thesaurus updated last on 12/30/2003 was used for all experiments presented (Additional File 3). 514 of the gene terms were specifically excluded based upon the manual review (Additional File 4). All gene term curation was performed prior to testing for associations.

Topic List Curation

Topic lists are sets of terms, each of which is to be tested for significant associations with the experimentally derived gene set. Topic lists used in the experiments were defined using terms from the MeSH Thesaurus provided by the National Library of Medicine. The list of terms for each topic set consisted of all terms and all descendents (with few exceptions) as listed in MeSH. The topic sets (presented as "name [Mesh ID]") used in these experiments were Cell Metabolism [G06.535] and Cell Physiology [G04.335] (Additional File 2). A few of the descendents of these MeSH terms were ignored because very few abstracts were associated with the term. When using MeSH terms to search PubMed, the search syntax used "MeSH Subheading Explosion" so that the resulting list of abstracts included abstracts coded with descendents (if any) of the MeSH term as well as abstracts coded with the MeSH term itself.

In addition, a topic set entitled "Pathways" was derived using the pathway descriptions from BioCarta [29] with some additional curation and testing (see Additional File 1 for methods and Additional File 2 for list of pathways). All pathway term curation was performed prior to testing for associations.

Literature Search

A list of the PubMed abstract ID's for each gene and each topic were obtained by searching PubMed using the

appropriate terms and a constant date range. The PubMed Electronic Date (EDAT) was used to query a constant subset of the PubMed abstracts between 12/31/93 and 12/31/03. Whole text information was not used due to its incomplete and inconsistent availability. To determine if results changed over time with a fixed chronological time frame and pre-set search parameters, three gene lists ("Ideal", "UCT", and "Van't Veer") and three curated term lists (metabolism, physiology, pathway) were run repetitively every month for 5 months using the fixed dates above in order to assess if ongoing efforts at the NCBI to improve the search engine or annotation of the literature significantly influence our approach to literature mining.

Term Ranking

Each topic list was analyzed independently. Within each topic set, specific terms were ranked according to how many times the term was associated with any of the genes in the gene list with respect to the total number of times the term is present in PubMed. We applied two different methods to measure the degree of intersection between any term and the gene set. First, we calculated the ratio of the number of abstracts containing any gene from the gene list AND the term divided by all abstracts containing the term (referred to as "expected abstracts"). The second score (called "product of frequency") takes into account the number of abstract mentions against the entire target gene set (see Additional File 1). In order to rank terms, we compared each terms score (by either method) to a distribution of scores between the same term and randomly selected gene sets. In the case of the product of frequency, the log(PF) (LPF) more closely approximated a normal distribution thus justifying the mean and standard deviation statistics used (see Additional File 1).

Testing against random gene sets

In order to provide a metric by which to interpret the rankings and determine the likelihood of finding a match given no association, we measure each topic score given the experimental set of genes against the distribution of scores from sets of genes chosen at random. Scoring 1,000 such random sets against the topic set, we obtain estimates for the mean and standard deviation of the $F(gene\text{-}set, topic)$ score for each topic. We tested if there was a significant difference in the statistics generated using or 1000 random sets of genes containing the same number of genes as the experimental gene set.

Results

We applied Literature Lab to lists of genes generated through microarray experiments to determine if such an approach provides biological insight. We started with 3 lists of genes from previously published microarray experiments, one generated in vitro (HL60 cells treated with ATRA [30], one generated in vivo (MPAKT transgenic mice treated with RAD001 [17]), and one generated from human tumors samples (A 70 gene model of breast recurrence [31,32]) to see if Literature Lab would correctly identify known biological processes and to how altering specific variables within Literature Lab impacts the results. Finally, we tested our method to a set of genes differentially expressed between local and metastatic prostate tumors and used immunohistochemistry to confirm the lead candidate.

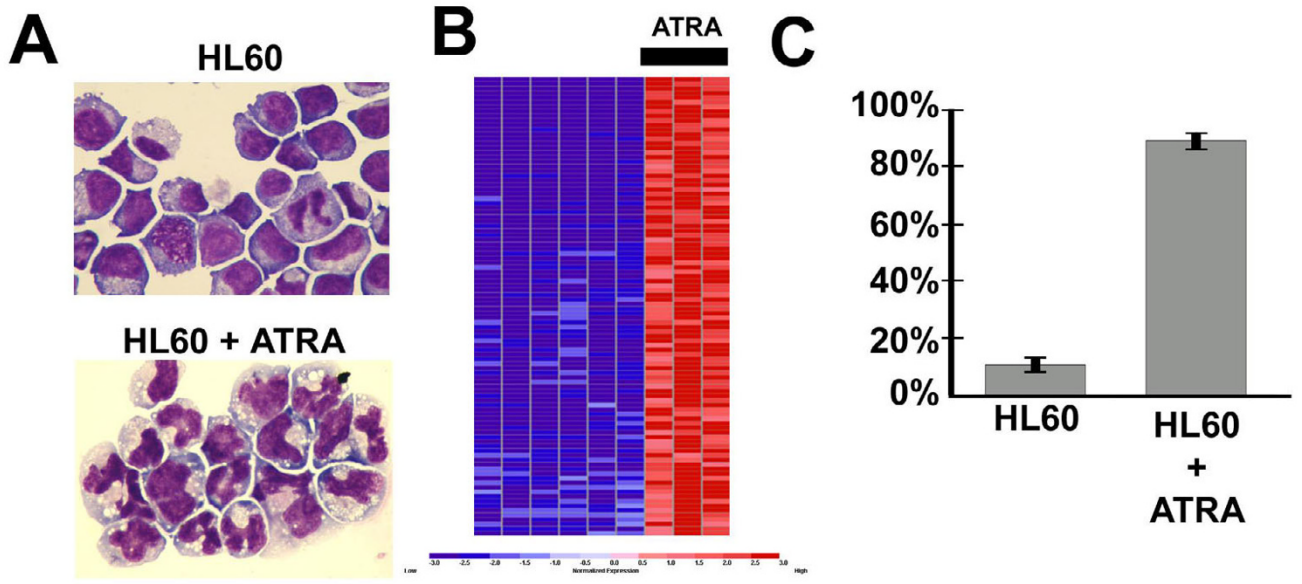
Literature Lab associates "respiratory burst" with ATRA treatment of a leukaemia cell line

Acute promyeloblastic leukemia (APML) cells differentiate when stimulated with all-trans retinoic acid (ATRA), a life-saving treatment for this disease [33]. The physiological impact of ATRA is detected by an increase in nitro blue tetrazolium (NBT) reduction (Fig. 1D), an assay established to measure the production of oxygen intermediates associated with respiratory bursts in cells [34]. We used microarray data available from a recent publication [30] to identify the top genes with increased expression in HL60 cells (a cell line used to model APML) following ATRA exposure (Fig. 1A and 1B). Genes with increased expression were analyzed with Literature Lab using key term sets for cellular metabolism, physiology, and pathways. Within the top ranked key terms for both metabolism and physiology was "respiratory burst" (Fig. 1C). This observation is confirmed by an increase in nitro blue tetrazolium (NBT) reduction, an established measure of the production of oxygen intermediates associated with respiratory bursts in cells [34].

We tested the impact of gene set size and literature time frame on the successful association between "respiratory burst" and the genes with increased expression following ATRA. While the specific literature time frame had relatively little impact on the association (i.e. first or last 5 years of the 10 year period) (see Additional File 5), the number of differentially expressed genes included in the analysis did impact our results (see Additional File 6) Specifically, the association between "respiratory burst" and the experimentally derived list of differentially expressed genes did not stabilize until greater than 150 genes were included. While there are no obvious rules to guide the upper limit with respect to the number of genes to be used in literature lab, the variability of results observed with shorter gene lists suggests that gene lists numbering less than 25 will not provide robust results and for experiments comparing phenotypes with marked differences, gene lists greater than 150 are encouraged.

"Hif and Hypoxia" strongly associated with mTOR inhibition

To test Literature Lab on a more challenging, *in vivo* derived dataset, we applied Literature Lab to a set of genes



D

	Score rank 1	2	3	4	5	6	7	8	9	10
CellPhysiology	Respiratory Burst	Sperm Maturation	Cell Degranulation	Chemotaxis	Cell Movement	Chemotaxis, Leukocyte	Cell Respiration	Phagocytosis	Myelopoiesis	Anoikis
Association	Moderate									
Score	2.76519077	1.87088484	1.852330255	1.83080109	1.633902	1.61500618	1.53794304	1.482401504	1.203957738	0.785818
MeshMetabolism	Pentosephosphate Pathway	Respiratory Burst	Gluconeogenesis	Tropism	Glycolysis	Biotinylation	Oxygen Consumption	Hydroxylation	Oxidation-Reduction	Energy Metabolism
Association	Strong	Moderate	Moderate							
Score	3.03757023	2.79478902	2.522628583	1.684608	1.680268	1.62021668	1.53479486	1.508730929	1.446403131	1.289381
Pathways	BTG	Tob	spermidine	pentose phosphate	CCR3	Neutrophil surface molecule	Sucrose	CCR5	CXCR4	Eosinophils Chemokine Allergy
Association	Strong		Strong	Moderate	Moderate	Moderate	Moderate	Moderate	Moderate	Moderate
Score	3.64202716	3.5687127	3.35263589	2.89526442	2.74468	2.58501325	2.49449901	2.442348294	2.329912832	2.04477

Figure 1

HL60 Leukemia cells treated with ATRA. A) May Grunwald Giemsa stain of HL60 cells prior to (upper) and following treatment with ATRA (lower). B) Heat map of the top 100 genes with increased expression in HL60 cells treated with ATRA compared to untreated HL60 cells (Red – high normalized expression, Blue – low normalized expression). C) Percentage of cells positive for nitro blue tetrazolium (NBT) reduction (Mean +/- St dev). D) Literature Lab ranking, association scores, and confidence calls for cell physiology, metabolism, and pathway terms (Association score is the log of the product of frequency (logPF)).

whose expression increased with the prostate-specific expression of myristoylated AKT (also associated with the eventual prostate phenotype of prostatic intraepithelial neoplasm) and decreased when mice were treated with RAD001, an mTOR inhibitor (Fig. 2A). This experiment was well controlled, characterized in a published manuscript [17], and has publicly available data (Gene Expression Omnibus accession number GSE1413). We performed a 10-year literature query to identify pathway terms associated with the 64 genes previously found to have the expression pattern described above [17]. Literature Lab identified "Hif" and "Hypoxia" terms in association with the list of genes thus highlighting the key biological insight discussed in the primary report (Fig. 2B). While HIF gene expression (HIF1A and HIF3A) was part of the original gene list and contributed to the strong association, Literature Lab found 12 and 17 additional genes associated with the key terms "Hif" and "Hypoxia", respectively, thus elevating these terms above others. Subsequent Western blotting for protein expression of Hif1α further supports the identified association (Fig. 2C) and demonstrates Literature Labs potential utility for data derived from animal models.

"Matrix metalloproteinase 9" and "VEGF" are associated with Breast Cancer Prognosis

Primary tumors perhaps provide the biggest challenge for analysis because of the additional associated technical and biological variation. We applied Literature Lab to the set of 70 genes associated with outcome in two seminal papers predicting breast cancer recurrence using microarray data [31,32]. Interestingly, of the cellular pathways investigated, "matrix metalloproteinase" and "VEGF" were strongly associated with the gene set (see Additional File 7). This unbiased association is strongly supported by prior studies finding MMP [35,36] and VEGF [37] activity strongly associated with the recurrent phenotype [37] and supports Literature Lab's applicability to data from human samples.

Literature drift

Given the daily growth of biomedical literature, reproducible literature search results require investigators to designate a specific chronological interval within which they queried for associations between a gene list and key terms. However, repeated queries using identical search criteria within a fixed interval demonstrated some lability hereto forward referred to as "literature drift". In sequential runs over 5 months, we measured this drift using the absolute value of the percentage change in the LPF value. Literature drift resulted in a median difference between queries of 2.51%. This degree and rate of literature drift was dependent both on the gene list and key terms as three independent sets of genes queried for three sets of key terms experienced different degrees of drift (Table 1).

By identifying the abstracts lost or gained between sequential runs, we identified a number of causal factors for literature drift:

• Changes to the various components of the NCBI search engine that result in different results for the same query. In particular, changes to the PubMed "phrase dictionary" are frequent and can yield different results for the same query at different points in time.

• The assignment of MeSH terms to abstracts subsequent to the addition of abstracts to the PubMed database.

• The editing of PubMed abstracts so as to change the title or text.

• The occasional deletion of abstracts from the database. Many of these deletions appear to be the removal of duplicates added to the PubMed database in error.

Importantly, while literature drift affected the numbers of abstracts linking a gene list with a term list, it had only minor effects on the final results of Literature Lab (Fig. 3). In addition, when the effects of literature drift were analyzed according to the strength of association between a key term and a set of genes; stronger associations were less likely to change when compared to associations with modest or no weight (11% for Strong (1 of 9), 16% for Moderate (5/32), and 27% for associations without weight (13 of 49). Thus while literature drift impacts the number of associations identified between the medial literature and a set of genes, our methodology minimizes its impact.

Discovery of increased FOSB in Metastatic Prostate Cancer

While the Literature Lab results from the HL60 cell line, MPAKT mouse, and breast cancer tumors are compelling, these sets had known biological associations and were used to help evaluate the methods herein described; they cannot be viewed as independent tests of our method. In order to test Literature Lab's ability to identify valid gene-key term associations through automated literature interrogation, we applied literature lab to microarray data obtained after RNA amplification of local and metastatic prostate cancer specimens. We first identified the top 100 genes up-regulated in the metastatic samples in order to identify the top pathway(s) associated with metastasis. FOSB was the top pathway and the only pathway term meeting heuristic criteria for a "strong" association (Fig. 4A). Expression of the FOSB gene was associated with metastasis but the additional identification of associations between FOSB and 5 other genes having increased expression in the metastatic tumors resulted in the top ranking of FOSB and a "strong" heuristic label. Based on

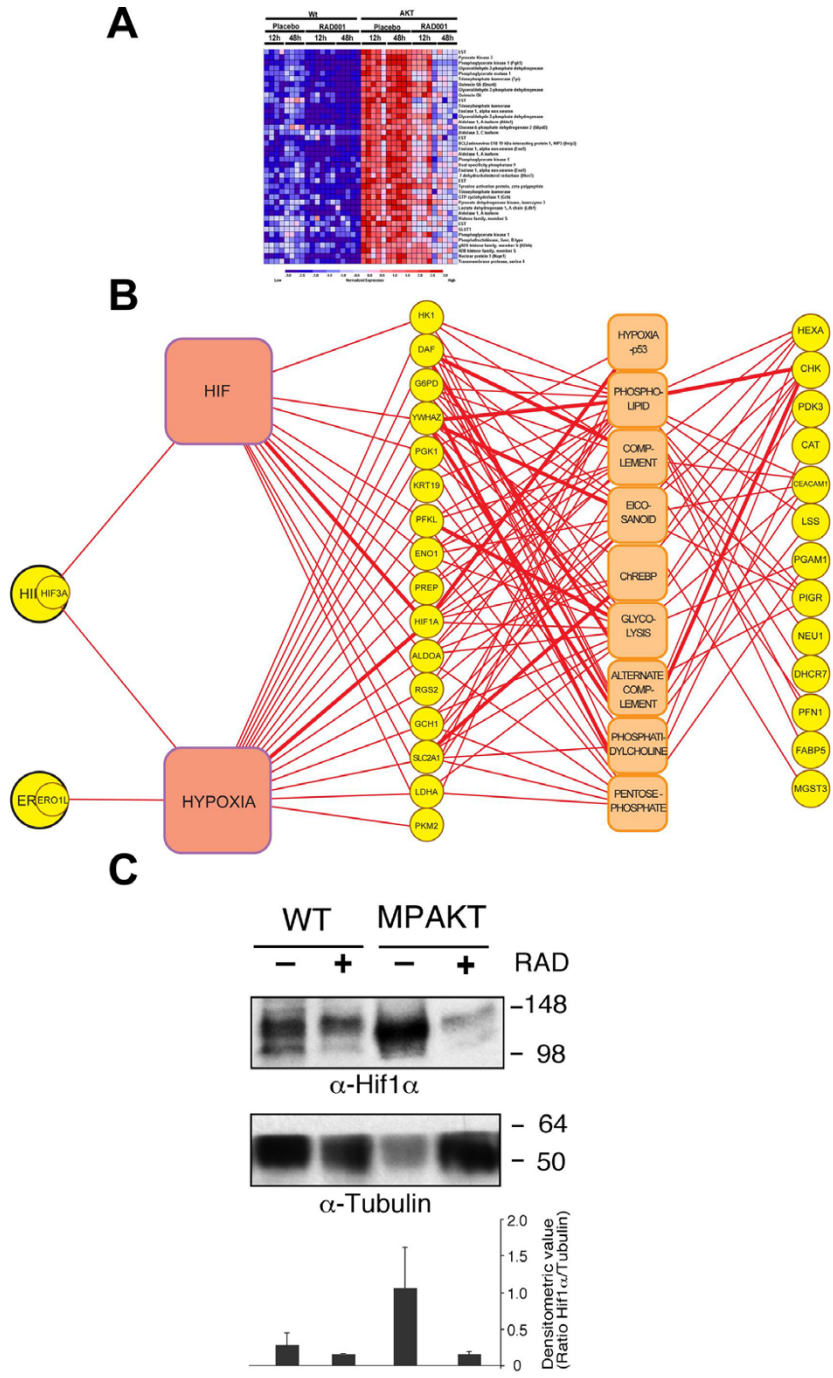


Figure 2

MPAKT mouse treated with RAD001. A) Heat map of normalized gene expression for the top 64 genes correlating with AKT expression and RAD001 treatment in the prostates of transgenic MPAKT mice. B) Association of pathway terms (Squares) with the 64 genes (circles) associated with RAD001 treatment of MPAKT mice. Size of square indicates confidence of associations, thickness of connecting line correlates with the number of abstracts linking any pair of term and gene. C) Protein lysates prepared from the VP of individual MPAKT and WT mice either treated with RAD001 (+) or with placebo (-) for 48 hours were immunoblotted with anti-Hif1α and anti-tubulin as indicated. Lower panel shows the densitometric ratio of Hif1α and tubulin.

Table 1: Median of Absolute Value of Percentage Change in LPF over 5 Months

	Metabolism Terms	Physiology Terms	*Pathways Terms*
MPAKT Gene Set	1.90%	1.98%	3.09%
HL60 Gene Set	1.32%	3.08%	3.41%
vant Veer Gene Set	1.05%	2.22%	1.91%

this result, we performed FOSB immunohistochemistry on a tissue microarray containing benign, local malignant, and metastatic malignant prostate tissue. Nuclear FOSB was significantly increased in the metastatic tumors compared to the locally invasive tumors ($p = 0.0013$, two-tailed Mann Whitney test) thus confirming the association highlighted by Literature Lab (Fig. 4B and Fig 4C).

Comparison with GOMiner, MESHER and GeneCite

To determine how Literature Lab compares with existing, publicly available sample annotation and literature mining technologies, we imported the AKT mouse and metastatic prostate cancer gene lists into GOMiner [21], MeSHer [20], and GeneCite [38]. When the AKT mouse gene set was imported into GOMiner, 222 GO terms were found to be significantly associated ($p \leq 0.05$) with the genes differentially expressed (see Additional File 8). Among the significant GO terms were "response to hypoxia" and "glycolysis" thus demonstrating that GOMiner is able to identify similar underlying biology but these terms were among a large number of significant terms and not prominent. Here we note that GOMiner and Literature Lab represent different approaches (enrichment analysis v. Literature mining, respectively) and, as such, the results are not directly comparable. However, the purposes of Literature Lab and GOMiner are similar (i.e. the identification of biological processes implicated by differential gene expression) and our results suggest

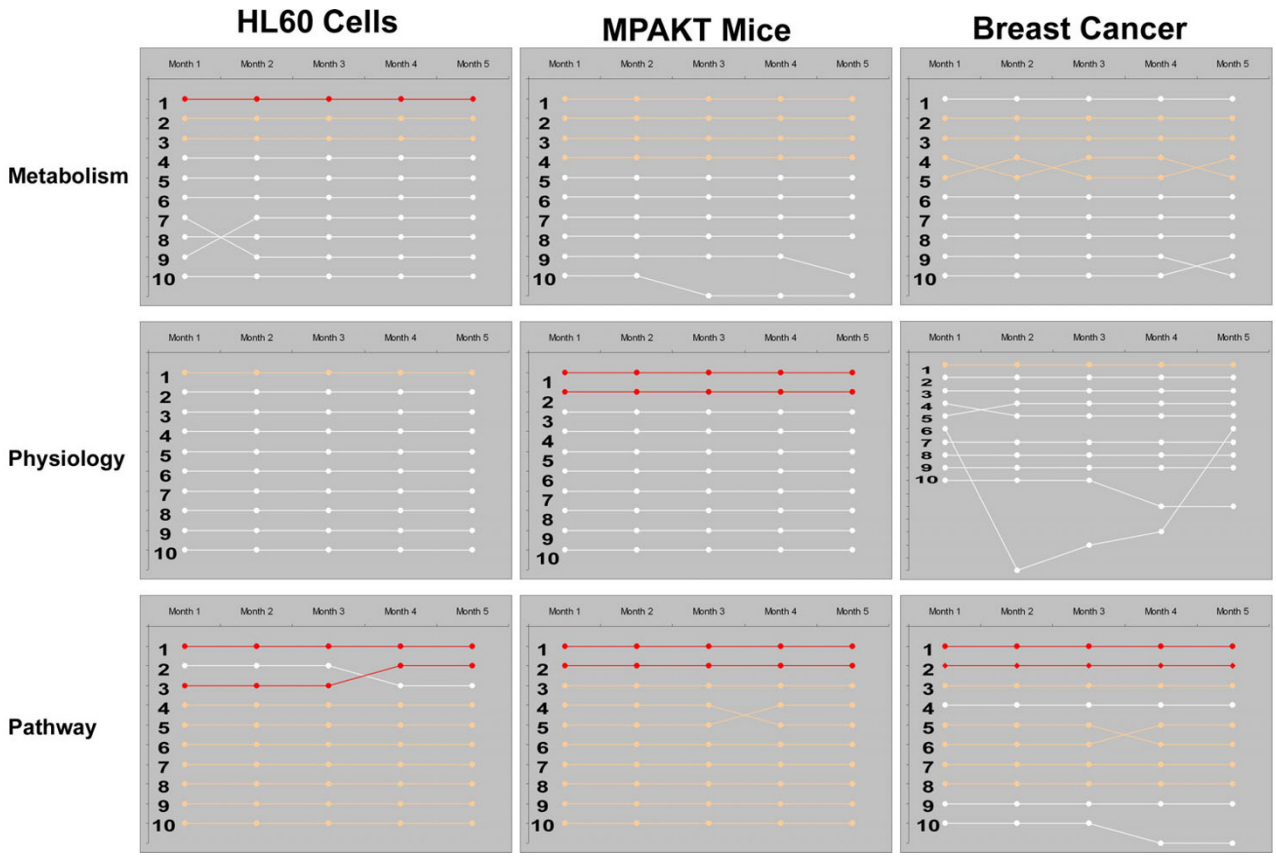
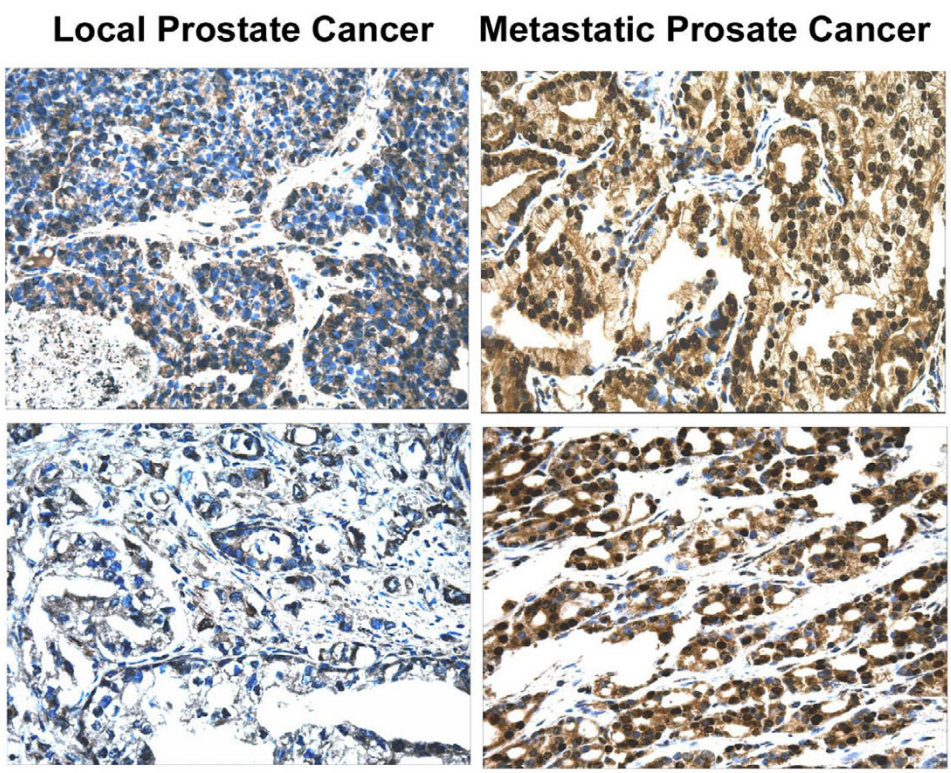


Figure 3
Literature Drift. Literature Lab results for repeat analysis performed on data from HL60 cells, MPAKT mice, and Breast Cancer tumors. The ranking of the top 10 terms from the Metabolism, Physiology, and Pathway lists are presented when the same analysis was performed within fixed dates (12/31/93 and 12/31/03). The ranking of each term is followed across the 5 monthly repeats by connecting lines. Color of line indicates strength of association (red = "strong"; tan = "moderate", white = no call).

A

Pathways	FOSB	Cytomegalovirus	Omega Oxidation	Phospholipase C	SLRP	acute inflammatory	Cytokine Inflammatory	myosin	Tubby	chemokine leukocyte
Association	Strong	Moderate			Moderate	Moderate				
Score	3.811579	2.732648	2.425799	2.323982	2.263882	2.187303	1.850983	1.825681	1.763941	1.756581

B



C

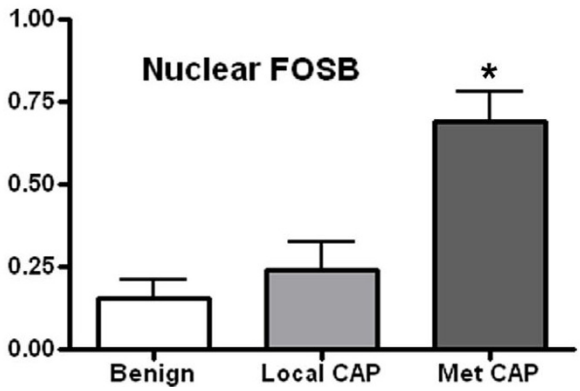


Figure 4

FOSB Identification and Immunohistochemistry. A) Ranking, association scores, and confidence calls for pathway terms associated with the 100 genes with increased expression in metastatic prostate tumors compared to local prostate tumors (Association score is the log of the product of frequency (logPF)). B) Two examples of local (left) and metastatic (right) prostate cancer FOSB expression measured by immunohistochemistry. C) Percentage of epithelial cell nuclei staining positive for benign prostate epithelium (n = 14), local prostate cancer (n = 20), and metastatic prostate cancer (n = 22) (* Met v. Local, p = 0.0013).

that direct literature mining and our statistical approach provide insight that cannot be fully reproduced using GOMiner.

MESHER did not find a significant association between "Hypoxia", "HIF", and "Hypoxia p53" and the AKT mouse gene list and there was very poor comparison between MESHER and Literature Lab despite each method using MeSH terms (see Additional File 9). GeneCite identifies the abstracts associating the terms and genes but has no direct measure of significance other than the number of abstracts (which favours more general terms over more specific terms). In addition, GeneCite has lower precision and recall when compared to Literature Lab due to the lack of a thesaurus for gene nomenclature (see Additional File 10). For example, many of the abstracts for the CAT gene refer to felines and CAT scans and not to the CAT gene.

When the genes differentially expressed between local and metastatic prostate cancers were imported into GOMiner and MESHER, there was poor overlap with both. FOSB, the term significantly associated with metastatic prostate cancer and subsequently validated by immunohistochemistry, was either not present in the library (GOMiner) or not associated with the gene list (MESHER, "Biogenetics-MeSH – "Genes, fos"). Results for GeneCite (see Additional File 11) exhibit the same limitations previously described for the AKT mouse gene set.

Discussion

Full utilization of publicly available, data-rich resources remains a universal challenge in contemporary scientific investigation. As technologies have diminished the cost and time associated with data collection, content within diverse repositories of data have increased exponentially. The medical literature is one such data repository and a repository that continues to grow rapidly. While investigators frequently use computational tools to interrogate genomic or gene expression data repositories, few use similar tools when reviewing the literature.

Literature Lab represents a method to comprehensively interrogate the literature for associations between a list of genes and a list of key terms in an unbiased manner in order to highlight potentially important biological processes implicated by the gene list. While there are many methods by which to develop a gene list, we have designed Literature Lab to aid in the analysis of microarray experiments which frequently associate the expression of hundreds to thousands of seemingly unrelated genes with cellular behaviors, in vivo phenotypes, or disease outcomes. We developed and refined our methodology using gene sets from previously published work and successfully tested Literature Lab on a novel dataset. The pathway term FOSB was ranked highest by Literature Lab and highlighted as having a "strong" association; an increase in nuclear FOSB staining was subsequently confirmed with immunohistochemistry.

Literature Lab is complementary to the increasingly prevalent pathway oriented approaches to the analysis of microarray data (Reviewed in [39]). As a general approach, these methods look for significant differential expression within a microarray experiment using predetermined aggregations of genes (alternatively called gene sets, metagenes, or gene modules) rather than individual genes [40]. Successful gene sets can identify underlying genetic abnormalities or signal transduction networks driving disease pathologies and help effectively bridge microarray data with biological significance [41,42].

Some pathway approaches methods use the literature and publicly available annotations (Gene Ontology) to develop gene sets and use these gene sets to interrogate expression data [43,44]. Literature Lab offers the opportunity to use a gene set derived from microarray data to interrogate the biomedical literature without a priori classification or annotation. As such, Literature Lab can appropriately interrogate the literature as it grows and evolves. When compared to two publicly available methods of analysis (GOMiner and MeSHer), the results of Literature Lab were more comparable with GOMiner. However, the statistical evaluation of associations identified by Literature Lab help improve the specificity of findings (highlighting strong associations) while maintaining sensitivity (neither GOMiner nor MeSHer identified the association between FOSB and genes with differential expression between local and metastatic prostate cancer). It should be noted, however, that given the difference in the approaches, our results cannot be interpreted as demonstrating the superiority of Literature Lab over GOMiner

Literature Lab remains dependent on the strength of the term lists and while we have demonstrated the use of lists for metabolism, physiology, and pathways, further development is focused on creating lists to include disease, pharmacological agents, drug toxicities, and many additional classes.

We initially anticipated that fixing the chronological interval for a query would ensure exact reproduction of the results. However, we identified literature drift within fixed retrospective intervals. While the degree of literature drift seems to range from minimal to moderate depending on the specific gene list, Literature Lab successfully limits the effects of literature drift especially for associations identified as "strong" with the current heuristics. Thus, while literature drift is unlikely to have significant impact on the

associations identified by Literature Lab, some variation in the specific weights and rankings of associations will change even when investigators define a fixed chronological interval within which they perform their query.

For this initial description, we focused on developing a robust measure of association, a relatively useful measure of significance, and heuristic rules to highlight the most important associations. Clearly, the specifics of our methods will be the subject of further investigation and refinement. While we have identified some critical elements of success (avoiding measures of association that are driven by single gene-term associations and having a gene set size of 25 or more), work is ongoing to explore the effects of refining the genes based upon the statistical association between their expression and the phenotype, limiting Literature Lab to specific journals of high quality content, and increasing the number of sets of key terms with which to test the association between gene expression.

Conclusion

The methodology herein described for Literature Lab highlights the biomedical literature as a content-rich resource amenable to automated, comprehensive interrogation. As with most complex data, successful comprehensive interrogation requires filtering out the noise and finding valuable information. Our current methods of gene annotation, key term curation, and literature interrogation, can find strong associations and are likely to benefit a diverse scientific community.

Availability and Requirement

The instructions, software, and data required to perform an analysis of a gene list using the techniques described herein can be obtained from http://www.acumenta.com/freeware/instructions.html. Sun's Java Runtime environment Version 1.4 or higher is required in order to run the software (and may be downloaded from [22]). The software runs on both Windows platforms (Windows 2000 and later) and Linux platforms. Memory of 1 GB and 5 GB of available disk space are required (much of the disk space requirement is for temporary storage during the analysis). In addition, the results of the analysis are presented in a Microsoft Excel spreadsheet for viewing on any system having a spreadsheet viewer capable of rendering the Microsoft Excel format.

• **Project name**: Literature Lab

• **Project home page**: http://www.acumenta.com/freeware/instructions.html

• **Operating system(s)**: Windows (2000 +) or Linux

• **Programming language**: e.g. Java

• **Other requirements**: Java Runtime environment 1.4 or higher

• **License**: Not required for Academic users

• **Any restrictions to use by non-academics**: License needed from Acumenta

Authors' contributions

PGF oversaw all aspects of the research including development of the concept of Literature Lab, provision of microarray results for analysis by Literature Lab, evaluation of all results from Literature Lab, and drafting of the manuscript. MGM helped develop and implement the software for Literature Lab. DAS helped develop and implement Literature Lab. KS oversaw all experiments with the leukemia cells, provided microarray data, provided the NBT data, and helped in the interpretation of Literature Lab results. DDV performed, analyzed, and interpreted the FOSB immunohistochemistry on prostate cancer specimens. PRM participated in the development, conceptualization, and implementation of Literature Lab. ML oversaw the FOSB immunohistochemistry and helped with the interpretation of Literature Lab results. SCT lead the development of Literature Lab from conceptualization to implementation, performed all Literature Lab analyses, and helped draft the manuscript.

Additional material

Additional file 1

Supplemental Methods.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-8-461-S1.doc]

Additional file 2

***Term Lists**. Excel spreadsheet of term lists used for physiology, metabolism, and pathway domains.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-8-461-S2.xls]

Additional file 3

***Gene Thesaurus**. Excel spreadsheet of all genes with all aliases used for all experiments in this manuscript.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-8-461-S3.xls]

Additional file 4

***Excluded Terms**. Excel spreadsheet of terms excluded from search due to ambiguity or limited associated information.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-8-461-S4.xls]

Additional file 5

Effects of Time Frame. *Standard Literature Lab interrogations performed on the HL60 gene set for key term lists for metabolism, pathways, and physiology. The only difference between the two interrogations involved time frame: One search included citations from the first 5 years of the standard 10 year interval (12/31/93 and 12/30/98, "1st5 yr") and the other the latter 5 years (12/31/98–12/31/03, "2nd5 yr"). Log_{10}(PF) ("Score") was used to order key terms and standard heuristics were applied to identify associations as "strong" or "moderate".*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-8-461-S5.JPEG]

Additional file 6

Effects of Gene Set Size. *Standard Literature Lab interrogations performed on the HL60 gene set for key term lists for metabolism with increasing size of the gene set. "Metabolism-X" where X equals 10, 25, 50, 100, 150, 200, 250, and 300 and represents the number of genes in the gene set.. Log_{10}(PF) ("Score") was used to order key terms and standard heuristics were applied to identify associations as "strong" or "moderate".*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-8-461-S6.JPEG]

Additional file 7

Results of Breast Cancer Gene Set. *The 70-genes associated with outcome for localized breast cancer were used as a gene set and interrogated by Literature Lab for associations between members of the gene set and key terms included on the lists of Physiology, Metabolizm, and Pathways. Log_{10}(PF) ("Score") was used to order key terms and standard heuristics were applied to identify associations as "strong" or "moderate".*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-8-461-S7.JPEG]

Additional file 8

GOMiner. *Excel spreadsheet of results from GOMiner using gene list from MPAKT experiment annotated for overlap with Literature Lab.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-8-461-S8.xls]

Additional file 9

Mesher. *Excel spreadsheet of results from MESHer using gene list from MPAKT experiment annotated for overlap with Literature Lab.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-8-461-S9.xls]

Additional file 10

GeneCite MPAKT. *Excel spreadsheet of results from GeneCite using gene list from MPAKT experiment annotated for overlap with Literature Lab.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-8-461-S10.xls]

Additional file 11

GeneCite Metastatic Prostate Cancer. *Excel spreadsheet of results from GeneCite using gene list from local v metastatic prostate cancer experiment annotated for overlap with Literature Lab.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-8-461-S11.xls]

Acknowledgements

P.G.F is a Damon Runyon-Lily Clinical Investigator (Grant #CA-29-05) and, in addition, was supported by NCI grants CA89031 and CA123175 during the design, analysis, and preparation of this manuscript. MGM, DAS, PRM, and SCT were supported by Acumenta, Inc during the design, analysis, and preparation of this manuscript. All analysis of data by investigators supported by Acumenta was performed in a blinded fashion; Acumenta supported investigators were unaware of the experiments from which the gene lists uploaded into Literature Lab were derived. KS was supported by the Department of Pediatric Oncology during the design and analysis of this work. DDV and ML were supported by the Department of Pathology at Dana Farber Cancer Institute during the design and analysis of this work.

References

1. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, Funke R, Gage D, Harris K, Heaford A, Howland J, Kann L, Lehoczky J, LeVine R, McEwan P, McKernan K, Meldrim J, Mesirov JP, Miranda C, Morris W, Naylor J, Raymond C, Rosetti M, Santos R, Sheridan A, Sougnez C, Stange-Thomann N, Stojanovic N, Subramanian A, Wyman D, Rogers J, Sulston J, Ainscough R, Beck S, Bentley D, Burton J, Clee C, Carter N, Coulson A, Deadman R, Deloukas P, Dunham A, Dunham I, Durbin R, French L, Grafham D, Gregory S, Hubbard T, Humphray S, Hunt A, Jones M, Lloyd C, McMurray A, Matthews L, Mercer S, Milne S, Mullikin JC, Mungall A, Plumb R, Ross M, Shownkeen R, Sims S, Waterston RH, Wilson RK, Hillier LW, McPherson JD, Marra MA, Mardis ER, Fulton LA, Chinwalla AT, Pepin KH, Gish WR, Chissoe SL, Wendl MC, Delehaunty KD, Miner TL, Delehaunty A, Kramer JB, Cook LL, Fulton RS, Johnson DL, Minx PJ, Clifton SW, Hawkins T, Branscomb E, Predki P, Richardson P, Wenning S, Slezak T, Doggett N, Cheng JF, Olsen A, Lucas S, Elkin C, Uberbacher E, Frazier M, Gibbs RA, Muzny DM, Scherer SE, Bouck JB, Sodergren EJ, Worley KC, Rives CM, Gorrell JH, Metzker ML, Naylor SL, Kucherlapati RS, Nelson DL, Weinstock GM, Sakaki Y, Fujiyama A, Hattori M, Yada T, Toyoda A, Itoh T, Kawagoe C, Watanabe H, Totoki Y, Taylor T, Weissenbach J, Heilig R, Saurin W, Artiguenave F, Brottier P, Bruls T, Pelletier E, Robert C, Wincker P, Smith DR, Doucette-Stamm L, Rubenfield M, Weinstock K, Lee HM, Dubois J, Rosenthal A, Platzer M, Nyakatura G, Taudien S, Rump A, Yang H, Yu J, Wang J, Huang G, Gu J, Hood L, Rowen L, Madan A, Qin S, Davis RW, Federspiel NA, Abola AP, Proctor MJ, Myers RM, Schmutz J, Dickson M, Grimwood J, Cox DR, Olson MV, Kaul R, Shimizu N, Kawasaki K, Minoshima S, Evans GA, Athanasiou M, Schultz R, Roe BA, Chen F, Pan H, Ramser J, Lehrach H, Reinhardt R, McCombie WR, de la Bastide M, Dedhia N, Blocker H, Hornischer K, Nordsiek G, Agarwala R, Aravind L, Bailey JA, Bateman A, Batzoglou S, Birney E, Bork P, Brown DG, Burge CB, Cerutti L, Chen HC, Church D, Clamp M, Copley RR, Doerks T, Eddy SR, Eichler EE, Furey TS, Galagan J, Gilbert JG, Harmon C, Hayashizaki Y, Haussler D, Hermjakob H, Hokamp K, Jang W, Johnson LS, Jones TA, Kasif S, Kaspryzk A, Kennedy S, Kent WJ, Kitts P, Koonin EV, Korf I, Kulp D, Lancet D, Lowe TM, McLysaght A, Mikkelsen T, Moran JV, Mulder N, Pollara VJ, Ponting CP, Schuler G, Schultz J, Slater G, Smit AF, Stupka E, Szustakowski J, Thierry-Mieg D, Thierry-Mieg J, Wagner L, Wallis J, Wheeler R, Williams A, Wolf YI, Wolfe KH, Yang SP, Yeh RF, Collins F, Guyer MS, Peterson J, Felsenfeld A, Wetterstrand KA, Patrinos A, Morgan MJ, Szustakowki J, de Jong P, Catanese JJ, Osoegawa K, Shizuya H, Choi S, Chen YJ: **Initial sequencing and analysis of the human genome.** *Nature* 2001, **409(6822):**860-921.

2. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, Gocayne JD, Amanatides P, Ballew RM, Huson DH, Wortman JR, Zhang Q, Kodira CD, Zheng XH, Chen L, Skupski M, Subramanian G, Thomas PD, Zhang J, Gabor Miklos GL, Nelson C, Broder S, Clark AG, Nadeau J, McKusick VA, Zinder N, Levine AJ, Roberts RJ, Simon M, Slayman C, Hunkapiller M, Bolanos R, Delcher A, Dew I, Fasulo D, Flanigan M, Florea L, Halpern A, Hannenhalli S, Kravitz S, Levy S, Mobarry C, Reinert K, Remington K, Abu-Threideh J, Beasley E, Biddick K, Bonazzi V, Brandon R, Cargill M, Chandramouliswaran I, Charlab R, Chaturvedi K, Deng Z, Di Francesco V, Dunn P, Eilbeck K, Evangelista C, Gabrielian AE, Gan W, Ge W, Gong F, Gu Z, Guan P, Heiman TJ, Higgins ME, Ji RR, Ke Z, Ketchum KA, Lai Z, Lei Y, Li Z, Li J, Liang Y, Lin X, Lu F, Merkulov GV, Milshina N, Moore HM, Naik AK, Narayan VA, Neelam B, Nusskern D, Rusch DB, Salzberg S, Shao W, Shue B, Sun J, Wang Z, Wang A, Wang X, Wang J, Wei M, Wides R, Xiao C, Yan C, Yao A, Ye J, Zhan M, Zhang W, Zhang H, Zhao Q, Zheng L, Zhong F, Zhong W, Zhu S, Zhao S, Gilbert D, Baumhueter S, Spier G, Carter C, Cravchik A, Woodage T, Ali F, An H, Awe A, Baldwin D, Baden H, Barnstead M, Barrow I, Beeson K, Busam D, Carver A, Center A, Cheng ML, Curry L, Danaher S, Davenport L, Desilets R, Dietz S, Dodson K, Doup L, Ferriera S, Garg N, Gluecksmann A, Hart B, Haynes J, Haynes C, Heiner C, Hladun S, Hostin D, Houck J, Howland T, Ibegwam C, Johnson J, Kalush F, Kline L, Koduru S, Love A, Mann F, May D, McCawley S, McIntosh T, McMullen I, Moy M, Moy L, Murphy B, Nelson K, Pfannkoch C, Pratts E, Puri V, Qureshi H, Reardon M, Rodriguez R, Rogers YH, Romblad D, Ruhfel B, Scott R, Sitter C, Smallwood M, Stewart E, Strong R, Suh E, Thomas R, Tint NN, Tse S, Vech C, Wang G, Wetter J, Williams S, Williams M, Windsor S, Winn-Deen E, Wolfe K, Zaveri J, Zaveri K, Abril JF, Guigo R, Campbell MJ, Sjolander KV, Karlak B, Kejariwal A, Mi H, Lazareva B, Hatton T, Narechania A, Diemer K, Muruganujan A, Guo N, Sato S, Bafna V, Istrail S, Lippert R, Schwartz R, Walenz B, Yooseph S, Allen D, Basu A, Baxendale J, Blick L, Caminha M, Carnes-Stine J, Caulk P, Chiang YH, Coyne M, Dahlke C, Mays A, Dombroski M, Donnelly M, Ely D, Esparham S, Fosler C, Gire H, Glanowski S, Glasser K, Glodek A, Gorokhov M, Graham K, Gropman B, Harris M, Heil J, Henderson S, Hoover J, Jennings D, Jordan C, Jordan J, Kasha J, Kagan L, Kraft C, Levitsky A, Lewis M, Liu X, Lopez J, Ma D, Majoros W, McDaniel J, Murphy S, Newman M, Nguyen T, Nguyen N, Nodell M, Pan S, Peck J, Peterson M, Rowe W, Sanders R, Scott J, Simpson M, Smith T, Sprague A, Stockwell T, Turner R, Venter E, Wang M, Wen M, Wu D, Wu M, Xia A, Zandieh A, Zhu X: **The sequence of the human genome.** *Science* 2001, **291(5507):**1304-1351.
3. Myers EW, Sutton GG, Delcher AL, Dew IM, Fasulo DP, Flanigan MJ, Kravitz SA, Mobarry CM, Reinert KH, Remington KA, Anson EL, Bolanos RA, Chou HH, Jordan CM, Halpern AL, Lonardi S, Beasley EM, Brandon RC, Chen L, Dunn PJ, Lai Z, Liang Y, Nusskern DR, Zhan M, Zhang Q, Zheng X, Rubin GM, Adams MD, Venter JC: **A whole-genome assembly of Drosophila.** *Science* 2000, **287(5461):**2196-2204.
4. Gibbs RA, Weinstock GM, Metzker ML, Muzny DM, Sodergren EJ, Scherer S, Scott G, Steffen D, Worley KC, Burch PE, Okwuonu G, Hines S, Lewis L, DeRamo C, Delgado O, Dugan-Rocha S, Miner G, Morgan M, Hawes A, Gill R, Celera, Holt RA, Adams MD, Amanatides PG, Baden-Tillson H, Barnstead M, Chin S, Evans CA, Ferriera S, Fosler C, Glodek A, Gu Z, Jennings D, Kraft CL, Nguyen T, Pfannkoch CM, Sitter C, Sutton GG, Venter JC, Woodage T, Smith D, Lee HM, Gustafson E, Cahill P, Kana A, Doucette-Stamm L, Weinstock K, Fechtel K, Weiss RB, Dunn DM, Green ED, Blakesley RW, Bouffard GG, De Jong PJ, Osoegawa K, Zhu B, Marra M, Schein J, Bosdet I, Fjell C, Jones S, Krzywinski M, Mathewson C, Siddiqui A, Wye N, McPherson J, Zhao S, Fraser CM, Shetty J, Shatsman S, Geer K, Chen Y, Abramzon S, Nierman WC, Havlak PH, Chen R, Durbin KJ, Egan A, Ren Y, Song XZ, Li B, Liu Y, Qin X, Cawley S, Worley KC, Cooney AJ, D'Souza LM, Martin K, Wu JQ, Gonzalez-Garay ML, Jackson AR, Kalafus KJ, McLeod MP, Milosavljevic A, Virk D, Volkov A, Wheeler DA, Zhang Z, Bailey JA, Eichler EE, Tuzun E, Birney E, Mongin E, Ureta-Vidal A, Woodwark C, Zdobnov E, Bork P, Suyama M, Torrents D, Alexandersson M, Trask BJ, Young JM, Huang H, Wang H, Xing H, Daniels S, Gietzen D, Schmidt J, Stevens K, Vitt U, Wingrove J, Camara F, Mar Alba M, Abril JF, Guigo R, Smit A, Dubchak I, Rubin EM, Couronne O, Poliakov A, Hubner N, Ganten D, Goesele C, Hummel O, Kreitler T, Lee YA, Monti J, Schulz H, Zimdahl H, Himmelbauer H, Lehrach H, Jacob HJ, Bromberg S, Gullings-Handley J, Jensen-Seaman MI, Kwitek AE, Lazar J, Pasko D, Tonellato PJ, Twigger S, Ponting CP, Duarte JM, Rice S, Goodstadt L, Beatson SA, Emes RD, Winter EE, Webber C, Brandt P, Nyakatura G, Adetobi M, Chiaromonte F, Elnitski L, Eswara P, Hardison RC, Hou M, Kolbe D, Makova K, Miller W, Nekrutenko A, Riemer C, Schwartz S, Taylor J, Yang S, Zhang Y, Lindpaintner K, Andrews TD, Caccamo M, Clamp M, Clarke L, Curwen V, Durbin R, Eyras E, Searle SM, Cooper GM, Batzoglou S, Brudno M, Sidow A, Stone EA, Venter JC, Payseur BA, Bourque G, Lopez-Otin C, Puente XS, Chakrabarti K, Chatterji S, Dewey C, Pachter L, Bray N, Yap VB, Caspi A, Tesler G, Pevzner PA, Haussler D, Roskin KM, Baertsch R, Clawson H, Furey TS, Hinrichs AS, Karolchik D, Kent WJ, Rosenbloom KR, Trumbower H, Weirauch M, Cooper DN, Stenson PD, Ma B, Brent M, Arumugam M, Shteynberg D, Copley RR, Taylor MS, Riethman H, Mudunuri U, Peterson J, Guyer M, Felsenfeld A, Old S, Mockrin S, Collins F: **Genome sequence of the Brown Norway rat yields insights into mammalian evolution.** *Nature* 2004, **428(6982):**493-521.
5. **Initial sequence of the chimpanzee genome and comparison with the human genome.** *Nature* 2005, **437(7055):**69-87.
6. Ensembl: **Ensembl.** [http://www.ensembl.org].
7. UCSCBrowser: **UCSC Browser.** [http://genome.ucsc.edu/].
8. NCBI_Browser: **NCBI Browser.** [http://www.ncbi.nlm.nih.gov/mapview/].
9. de Bruijn B, Martin J: **Getting to the (c)ore of knowledge: mining biomedical literature.** *Int J Med Inform* 2002, **67(1-3):**7-18.
10. Muller HM, Kenny EE, Sternberg PW: **Textpresso: an ontology-based information retrieval and extraction system for biological literature.** *PLoS Biol* 2004, **2(11):**e309.
11. Jenssen TK, Laegreid A, Komorowski J, Hovig E: **A literature network of human genes for high-throughput analysis of gene expression.** *Nat Genet* 2001, **28(1):**21-28.
12. Alako BT, Veldhoven A, van Baal S, Jelier R, Verhoeven S, Rullmann T, Polman J, Jenster G: **CoPub Mapper: mining MEDLINE based on search term co-publication.** *BMC Bioinformatics* 2005, **6(1):**51.
13. LaBaer J: **Mining the literature and large datasets.** *Nat Biotechnol* 2003, **21(9):**976-977.
14. Rubinstein R, Simon I: **MILANO--custom annotation of microarray results using automatic literature searches.** *BMC Bioinformatics* 2005, **6:**12.
15. Delong M, Yao G, Wang Q, Dobra A, Black EP, Chang JT, Bild A, West M, Nevins JR, Dressman H: **DIG--a system for gene annotation and functional discovery.** *Bioinformatics* 2005, **21(13):**2957-2959.
16. Majumder PK, Yeh JJ, George DJ, Febbo PG, Kum J, Xue Q, Bikoff R, Ma H, Kantoff PW, Golub TR, Loda M, Sellers WR: **Prostate intraepithelial neoplasia induced by prostate restricted Akt activation: the MPAKT model.** *Proc Natl Acad Sci U S A* 2003, **100(13):**7841-7846.
17. Majumder PK, Febbo PG, Bikoff R, Berger R, Xue Q, McMahon LM, Manola J, Brugarolas J, McDonnell TJ, Golub TR, Loda M, Lane HA, Sellers WR: **mTOR inhibition reverses Akt-dependent prostate intraepithelial neoplasia through regulation of apoptotic and HIF-1-dependent pathways.** *Nat Med* 2004, **10(6):**594-601.
18. Loda M, Capodieci P, Mishra R, Yao H, Corless C, Grigioni W, Wang Y, Magi-Galluzzi C, Stork PJ: **Expression of mitogen-activated protein kinase phosphatase-1 in the early phases of human epithelial carcinogenesis.** *Am J Pathol* 1996, **149(5):**1553-1564.
19. Rossi S, Graner E, Febbo P, Weinstein L, Bhattacharya N, Onody T, Bubley G, Balk S, Loda M: **Fatty acid synthase expression defines distinct molecular signatures in prostate cancer.** *Mol Cancer Res* 2003, **1(10):**707-715.
20. MeSHer: **MeSHer.** [http://compbio.dfci.harvard.edu/mesher.html].
21. GOMiner: **GOMiner.** [http://discover.nci.nih.gov/gominer/].
22. Microsystems S: **Java Software.** [http://java.sun.com].
23. Eclipse: **Eclipse Software.** [http://www.eclipse.org].
24. Utilities NCBIP: **NCBI Programming Utilities.** [http://eutils.ncbi.nlm.nih.gov/entrez/query/static/eutils_help.html].
25. Classes XMLB: **XML Binding Classes.** [http://java.sun.com/developer/technicalArticles/WebServices/jaxb].
26. Classes APOI: **Apachi POI Classes.** [http://poi.apache.org].
27. database SSOURCE: **Stanford SOURCE.** [http://genome-www5.stanford.edu/cgi-bin/source/sourceSearch].
28. Schijvenaars BJ, Mons B, Weeber M, Schuemie MJ, van Mulligen EM, Wain HM, Kors JA: **Thesaurus-based disambiguation of gene symbols.** *BMC Bioinformatics* 2005, **6:**149.

29. Biocarta: **Biocarta.** [http://www.biocarta.com/genes/allPathways.asp].
30. Stegmaier K, Ross KN, Colavito SA, O'Malley S, Stockwell BR, Golub TR: **Gene expression-based high-throughput screening(GE-HTS) and application to leukemia differentiation.** *Nat Genet* 2004, **36(3):**257-263.
31. van de Vijver MJ, He YD, van't Veer LJ, Dai H, Hart AA, Voskuil DW, Schreiber GJ, Peterse JL, Roberts C, Marton MJ, Parrish M, Atsma D, Witteveen A, Glas A, Delahaye L, van der Velde T, Bartelink H, Rodenhuis S, Rutgers ET, Friend SH, Bernards R: **A gene-expression signature as a predictor of survival in breast cancer.** *N Engl J Med* 2002, **347(25):**1999-2009.
32. van 't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AA, Mao M, Peterse HL, van der Kooy K, Marton MJ, Witteveen AT, Schreiber GJ, Kerkhoven RM, Roberts C, Linsley PS, Bernards R, Friend SH: **Gene expression profiling predicts clinical outcome of breast cancer.** *Nature* 2002, **415(6871):**530-536.
33. Tallman MS, Andersen JW, Schiffer CA, Appelbaum FR, Feusner JH, Ogden A, Shepherd L, Willman C, Bloomfield CD, Rowe JM, Wiernik PH: **All-trans-retinoic acid in acute promyelocytic leukemia.** *N Engl J Med* 1997, **337(15):**1021-1028.
34. Schopf RE, Mattar J, Meyenburg W, Scheiner O, Hammann KP, Lemmel EM: **Measurement of the respiratory burst in human monocytes and polymorphonuclear leukocytes by nitro blue tetrazolium reduction and chemiluminescence.** *J Immunol Methods* 1984, **67(1):**109-117.
35. Duffy MJ, Maguire TM, Hill A, McDermott E, O'Higgins N: **Metalloproteinases: role in breast carcinogenesis, invasion and metastasis.** *Breast Cancer Res* 2000, **2(4):**252-257.
36. Gupta GP, Nguyen DX, Chiang AC, Bos PD, Kim JY, Nadal C, Gomis RR, Manova-Todorova K, Massague J: **Mediators of vascular remodelling co-opted for sequential steps in lung metastasis.** *Nature* 2007, **446(7137):**765-770.
37. Sledge GW Jr.: **Vascular endothelial growth factor in breast cancer: biologic and therapeutic aspects.** *Semin Oncol* 2002, **29(3 Suppl 11):**104-110.
38. GeneCite: **GeneCite.** .
39. Stuart JM, Segal E, Koller D, Kim SK: **A gene-coexpression network for global discovery of conserved genetic modules.** *Science* 2003, **302(5643):**249-255.
40. Bild A, Febbo PG: **Application of a priori established gene sets to discover biologically important differential expression in microarray data.** *Proc Natl Acad Sci U S A* 2005, **102(43):**15278-15279.
41. Bild AH, Yao G, Chang JT, Wang Q, Potti A, Chasse D, Joshi MB, Harpole D, Lancaster JM, Berchuck A, Olson JA, Marks JR, Dressman HK, West M, Nevins JR: **Oncogenic pathway signatures in human cancers as a guide to targeted therapies.** *Nature* 2005.
42. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP: **Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles.** *Proc Natl Acad Sci U S A* 2005, **102(43):**15545-15550.
43. Mootha VK, Lindgren CM, Eriksson KF, Subramanian A, Sihag S, Lehar J, Puigserver P, Carlsson E, Ridderstrale M, Laurila E, Houstis N, Daly MJ, Patterson N, Mesirov JP, Golub TR, Tamayo P, Spiegelman B, Lander ES, Hirschhorn JN, Altshuler D, Groop LC: **PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes.** *Nat Genet* 2003, **34(3):**267-273.
44. Zeeberg BR, Feng W, Wang G, Wang MD, Fojo AT, Sunshine M, Narasimhan S, Kane DW, Reinhold WC, Lababidi S, Bussey KJ, Riss J, Barrett JC, Weinstein JN: **GoMiner: a resource for biological interpretation of genomic and proteomic data.** *Genome Biol* 2003, **4(4):**R28.