Bioprospecting the Bibleome: Adding Evidence to Support the Inflammatory Basis of Cancer

Peter L. Elkin1*, Andrew Frankel2, Ester H. Liebow-Liebling2, Jared R. Elkin2, Mark S. Tuttle1 and Steven H. Brown3,4

1Center for Biomedical Informatics, Mount Sinai School of Medicine, New York, NY, USA
2Mayo Clinic College of Medicine, Rochester, MN, USA
3Vanderbilt University, Nashville, TN, USA
4Department of Veterans Affairs, Nashville, TN, USA

Abstract

Background, cancer and question: BioProspecting is a novel approach that enabled our team to mine genetic marker related data from the New England Journal of Medicine (NEJM) utilizing Systematized Nomenclature of Medicine-Clinical Terms (SNOMED CT) and the Human Gene Ontology (HUGO). Genes associated with disorders using the Multi-threaded Clinical Vocabulary Server (MCSV) Natural Language Processing (NLP) engine, whose output was represented as an ontology-network incorporating the semantic encodings of the literature. Metabolic functions were used to identify potentially novel relationships between (genes or proteins) and (diseases or drugs). In an effort to identify genes important to transformation of normal tissue into a malignancy, we went on to identify the genes linked to multiple cancers and then mapped those genes to metabolic and signaling pathways.

Findings: Ten Genes were related to 30 or more cancers, 72 genes were related to 20 or more cancers and 191 genes were related to 10 or more cancers. The three pathways most often associated with the top 200 novel cancer markers were the Acute Phase Response Signaling, the Glucocorticoid Receptor Signaling and the Hepatic Fibrosis/Hepatic Stellate Cell Activation pathway.

Meaning and implications of the advance: This association highlights the role of inflammation in the induction and perhaps transformation of mortal cells into cancers.

BioProspecting can speed our identification and understanding of synergies between articles in the biomedical literature. By mapping these associations to known metabolic, regulatory and signaling pathways we were able to identify further evidence for the inflammatory basis of cancer.

Keywords: Bioinformatics; Biomedical literature data mining; Natural Language Processing (NLP); Ontology; Novel cancer biomarkers

Introduction

The Genomics and Bioinformatics research communities have been successful over the past several years in discovering of the genetic basis of Mendelian disorders [1]. Computational tools delivered by the field of Bioinformatics have been essential in our search for the genetic basis of Mendelian disease [2]. Bioinformatics research develops technologies in the context of clinical and genetic data standards that have been employed to create linkages between genes and disorders [3]. Unfortunately, many diseases cannot be traced to a single genetic variation where the true origin of a disease is a complex interplay of genetic variations, environmental factors, and “lifestyle” characteristics, along with some stochastic processes [4]. To advance medical science and healthcare, a broader understanding of genetic markers and their inter-relationships is required [5]. Our project to “discover” a genetic marker by mining the medical literature was an attempt to reveal, and later research, linkages between variants discerned through the mining of the medical literature [6].

New England Journal of Medicine (NEJM)

The NEJM is considered one of the highest impact peer-reviewed medical journals [7]. NEJM is the oldest continuously published medical journal. In 1812, the journal was inaugurated as the New England Journal of Medicine and Surgery [8]. The NEJM, in 1928, took on its present name after one hundred years as The Boston Medical and Surgical Journal.

The commentary associated with the presentation of the George Polk Award noted that its 1977 award to the New England Journal of Medicine “provided the first significant mainstream visibility for a publication that would achieve enormous attention and prestige in the ensuing decades” [9].

The NEJM publishes papers on original research, widely cited editorials, review articles, correspondences and case reports. The NEJM consistently has the highest impact factor of the journals of clinical medicine and in 2010 it was 53.48.

The NEJM provides on-line (electronic) access. Our study included the full text of all articles from January 1994 through December 2006.

SNOMED CT (Systematized Nomenclature of Medicine-Clinical Terms)

SNOMED CT (SCT) is a large-scale general medical ontology used to describe current medical and health knowledge [10]. We employed SCT to represent clinical disorders, proteins, chemicals and metabolic functions identified within the biomedical literature. SCT is the most

*Corresponding author: Peter L. Elkin, M.D., MACP, FACMI, Mount Sinai School of Medicine, Center for Biomedical Informatics, New York, USA, Tel: 212-860-3837; Fax: 212-824-2329; E-mail: ontolimatics@gmail.com

Received July 19, 2011; Accepted May 03, 2012; Published May 05, 2012

Citation: Elkin PL, Frankel A, Liebow-Liebling EH, Elkin JR, Tuttle MS, et al. (2012) Bioprospecting the Bibleome: Adding Evidence to Support the Inflammatory Basis of Cancer. Metabolomics 2:112. doi:10.4172/2153-0769.1000112

Copyright: © 2012 Elkin PL et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.
comprehensive of the general medical and biological Ontologies. SCT is maintained by the International Heath Terminology Standards Development Organization (IHTSDO) and is designed to represent the health and medical domains. Although very broad with over 291,000 concepts and over 1.2 million relations, it does not provide complete coverage of all medical content. SCT was formed through a merger of SNOMED RT (Reference Terminology) in the late 1990s with another large ontology, Clinical Terms v3, developed in the United Kingdom.

Ontologies, to be practically useful for computation, need to be accurate in two specific ways; firstly, they need to match as closely as possible our understanding of the natural world. Secondly, ontologies need to be at least 95% complete in their domain coverage. This means the ontology needs to be constructed with very close attention to the meanings of its concepts. In other words the concepts used in the ontology need to represent closely the understanding we have of the real world [11]. The variety of linguistic “usage” of those terms needs to be harnessed to ensure that the ontology has adequate content coverage. One should adequately explore the diversity of semantic roles of the concepts to ensure that only those roles that are useful are included in the ontological modeling, and to remove ambiguity in the use of those roles. Precision of each definition is particularly important in establishing the relationships between concepts and identifying the fundamental atomic concepts and how they combine or “fit together” systematically in compositional expressions. Ontologies also consist of levels of abstraction. In ontological computation there are two basic forms of abstraction, aggregation and generalization. Aggregation hierarchies are created from the ontological indexing by linking like objects. Generalization hierarchies are used throughout SCT as the basic mechanisms for relating content (e.g. All Oncological disorders).

SCT provides concepts to represent diagnoses, findings, procedures and testing. In a previous article we Elkin et al. published a cohort study reporting that SCT provided 92.3% coverage of common medical problems from cases from the Mayo Clinic [12]. This NLP system has been used in subsequent studies for electronic quality monitoring [13].

HUGO

The Human Genome Organization (HUGO) is an organization involved in the Human Genome Project. HUGO was established to foster collaboration between genome scientists around the world, in 1989. The HUGO Gene Nomenclature Committee (HGNC) is one of HUGO’s committees that assign a unique gene name and symbol for each human gene.

Materials and Methods

We parsed the full-text content of the New England Journal of Medicine (1994-2006) using the Multi-threaded Clinical Vocabulary Server. The MCVS creates an ontology-network of the semantic encodings of genes, proteins, disorders and drugs and metabolic functions identified in the literature that are organized by the section of the article including the tables.

The indexing was done utilizing SNOMED-CT and the HUGO Ontologies. We utilized SNOMED CT as it provides robust clinical indexing. SNOMED-CT has >370,000 concepts and >1,000,000 terms (in our lab we added another 790,000 terms to improve its clinical relevance) and HUGO has >26,000 human gene names. This concept based indexing represents a broad and consistent data infrastructure across articles from the literature [14].

We identified co-occurrences of genes and metabolic functions, proteins and metabolic functions, diseases and metabolic functions and drugs and metabolic functions. Next, we matched these data sets linking pairs (e.g. Disorders and Genes) with a common metabolic function, identifying functional relationships between proteins and diseases, for example. Candidate functional relationships were identified between genes and diseases, proteins and drugs, genes and drugs, and drugs and diseases. Next, we identified the disjoint sets where, for example, a gene and a disease match across metabolic function but have not been mentioned together in any previous NEJM journal article.

We compiled the genes from the disjoint set (potentially novel markers) that were related to more than three tissue types of cancer. We sorted the genes by the number of cancers associated with the gene. Then we took the 200 genes from the top of this list and mapped them against the metabolic pathways and signaling pathways available within Ingenuity™ [15]. Ingenuity is commercially available pathway analysis software. The pathways associated with the highest number of these possibly novel genes are reported.

We compared the chance that a gene would be related to 20 or more cancers and by random chance with the rate of identification in our study and analyzed the results using the McNemar test.

The conceptual process can be visualized as shown in Figure 1.

Results

Novel relationships

SCT contained 574 metabolic functions that were used to link to Genes, Proteins, Disorders, and Drugs.

We identified 1756 genes that were related to 10,303 different types of cancer. Ten genes were related to 30 or more cancers. The gene associated with the most cancers was MYELIN OLIGODENDROCYTE GLYCOPROTEIN and was related to 35 types of cancer. There were 72 genes related to 20 or more cancers and there were 191 genes related to 10 or more cancers. We have annotated the top 50 genes and present them in table 1. In table 2 we show the clinical pathways containing the top 200 genes from this dataset and which we then mapped against the metabolic and signaling pathways (See Table 2) to identify the most common pathways that involved this set of genes. The pathways with the greatest overlap with this dataset were the Acute Phase Response

![Figure 1: Illustrates the synergy between two journal articles, linking a gene to a disease through a common metabolic function.](image-url)
| Gene                                                                 | Number of Cancers | Annotation                                                                 |
|----------------------------------------------------------------------|-------------------|---------------------------------------------------------------------------|
| MYELIN OLIGODENDROCYTE GLYCOPROTEIN                                 | 35                | Neuro Tumors                                                              |
| LIPOPOLYSACCHARIDE BINDING PROTEIN                                   | 34                | inflammatory; antimicrobial peptide LL-37 or Hcap 18 is a precursor.     |
| CATHELICIDIN ANTIMICROBIAL PEPTIDE                                   | 34                | Vascular Proliferation                                                    |
| POLYMERIC IMMUNOGLOBULIN RECEPTOR                                    | 33                | Transports Immunoglobulin across cell membranes                           |
| TELOMERASE REVERSE TRANSCRIPTASE                                     | 32                | Essential for transformation but unable to accomplish conversion on its own |
| DIHYDROPYRIDINIDE DEHYDROGENASE                                      | 31                | Degrades S-FU, deficiency leads to medulloblastoma, (Diasso - Auto Recessive Inheritance) |
| ORNITHINE CARBAOYLTRANSFERASE                                        | 30                | Trans-species; proliferation associated; Chemoprevention which can be an inhibitor of OTC (DMFO) |
| GASTRIC INHIBATORY POLYPEPTIDE                                       | 30                | glucose-dependent insulinoitropic polypeptide                             |
| HYPERGONADOTROPIC HYPOGONADISM                                       | 30                | hypogonadism with elevated gonadotropins                                  |
| UROPORPHYRINOGEN DECARBOXYLASE                                       | 30                | PORPHYRIA CUTANEA TARDA                                                   |
| ASILOGLYCOPROTEIN RECEPTOR 1                                         | 29                | Hepatic Cell signaling - to Endoplasmic Reticulum                         |
| VASOACTIVE INTESTINAL PEPTIDE                                        | 29                | VIP promotes TH2 differentiation and inhibits TH1 responses by regulating macrophage costimulatory signals and probably IL12/IFN-gamma production; Decreases with Age. |
| PHOSPHOLIPID TRANSFER PROTEIN                                        | 29                | polymorphisms decreased HDL; transcription factor-binding motifs, SP1 and AP-2 |
| DIHYDOORotate DEHYDROGENASE                                          | 28                | catalyzes the fourth enzymatic step in de novo pyrimidine biosynthesis; For cell replication. |
| CILIARY NEUROTROPHIC FACTOR                                          | 27                | Errors cause Weight Gain; some earlier presentation of ALS                |
| ARYL HYDROCARBON RECEPTOR                                            | 25                | Halogenated aromatic hydrocarbons cause mediated by the enzyme produced by this gene; basic helix-loop-helix/PAS family transcription factors, regulates the effects of Estrogen Receptors |
| BREAKPOINT CLUSTER REGION                                             | 25                | CML                                                                       |
| PHENYLALANINE HYDOXYLASE                                             | 25                | PKU                                                                       |
| PHOSPHOGLYCERATE KINASE 1                                            | 25                | functions in glycolysis but is secreted by tumor cells and participates in the angiogenic process as a disulfide reductase; anti-angiogenic and slows tumor growth; deficiency causes hemolytic anemia |
| ISLET AMYLIOID POLYPEPTIDE                                           | 25                | diabetes Mellitus type 1 and 2, Insulorna,                               |
| CHOLINE ACETYLTTRANSFERASE                                           | 25                | deficiency causes myasthenic symptoms                                     |
| NUCLEOSIDE PHOSPHORYLASE                                             | 24                | deficiency led to lymphoma, lymphopenia,                                 |
| LEUKOTRIENE A4 HYDROLASE                                             | 24                | inflammatory mediator, rarely in African Americans                       |
| GLUTATHIONE PEROXIDASE 1                                             | 24                | Hemolytic Anemia, 6 copy repeats associated with myeloid leukemias        |
| PROSTAGLANDIN E SYNTHASE                                             | 24                | A p53 induced gene; PIG12 gene (syonym) encodes a microsomal glutathione S-transferase; is anti-inflammatory and can lead to apoptosis. Can improve Hepatocellular carcinoma by blocking PG1 and PG3 receptors. |
| ERYTHROPOIETIN RECEPTOR                                              | 23                | proerythroblast cell lines that expressed Epor and had rearranged and inactivated expression of the p53 suppressor oncogene |
| GROWTH HORMONE RECEPTOR                                              | 23                | leading to synthesis and secretion of insulin-like growth factor 1 ; GHR belongs to the cytokine superfamily of receptors that depend on JAK tyrosine kinases (see 147795) for activation of STATs |
| MELANOCORTIN 4 RECEPTOR                                              | 23                | decreases body weight.                                                    |
| LEUKOTRIENE C4 SYNTHASE                                              | 23                | potent lipid mediators of tissue inflammation                             |
| ARGININOSUCCINATE LYASE                                              | 23                | deficiency results in defective cleavage of Argininosuccinic acid (ASA), a precursor to fumarate in the citric acid cycle, which causes accumulation of ASA in cells and an excessive excretion of ASA in urine (argininosuccinic aciduria). Deficiency characterized by hyperammonemia in affected individuals. |
| HISTIDINE DECARBOXYLASE                                              | 23                | the only histamine-synthesizing enzyme; mouse models w/ gene removed are characterized by undetectable tissue histamine levels. |
| DIHYDROFOLATE REDUCTASE                                              | 23                | converts Dihydrofollic acid (vitamin B9), which interacts with bacteria during cell division and can be targeted with drug analgols to prevent nucleic acid synthesis, to tetrahydrofolic acid. |
| ADENOSINE A2A RECEPTOR                                               | 22                | a potent biologic mediator that modulates the activity of numerous cell types, including various neuronal populations, platelets, neutrophils and mast cells, and smooth muscle cells in bronchi and vasculature, helping to protect cells and tissues during stress situations such as ischemia. Abundant in basal ganglia, vasculature and platelets, and stimulates adenylyl cyclase. It is a major target of caffeine. Knockout mouse models show reduced exploratory activity, and caffeine, which normally stimulates exploratory behavior, became a depressant of exploratory activity. They scored higher in anxiety tests, and male mice were more aggressive toward intruders. Their response to acute pain stimuli was slower. Blood pressure and heart rate were increased, as well as platelet aggregation. The specific A2a agonist CGS 21680 lost its biologic activity in all systems tested. |
| Gene Name                          | Annotation |
|-----------------------------------|------------|
| INTERLEUKIN 4 RECEPTOR            | 22         |
| INTERLEUKIN 6 RECEPTOR            | 22         |
| PROMYELOCYTIC LEUKEMA             | 22         |
| PANCREATIC POLYPEPTIDE             | 22         |
| THYMIDYLYATE SYNTHETASE           | 22         |
| DEOXYHYPUSINE SYNTHASE            | 22         |
| XANTHINE DEHYDROGENASE            | 22         |
| SORBITOL DEHYDROGENASE            | 22         |
| SPLEEN TYROSINE KINASE            | 22         |
| ANKYLOSING SPONDYLITIS            | 22         |
| EOSINOPHIL PEROXIDASE             | 21         |
| HYALURONAN SYNTHASE 3             | 21         |
| ADENOSINE A3 RECEPTOR             | 21         |
| HISTONE DEACETYLASE 2             | 21         |
| HISTONE DEACETYLASE 4             | 21         |

Table 1: Top 50 Genes associated with Multiple Cancer Tissue Types and their annotations.

Signaling pathway (See Figure 2) respectively, the Glucocorticoid Receptor Signaling and the Hepatic Fibrosis/Hepatic Stellate Cell Activation pathway that includes targets such as TNFa-NFkB.

The chance that a gene is related to twenty or more cancers as identified by our bioprospecting methods is not truly related cancer is very small (that all findings would be false positives; $3.9 \times 10^{-24}$). If we compare the chance that a gene would be related to twenty or more cancers based on chance alone with the current findings the results are highly statistically significant ($p<0.001$; McNemar Test).

**Discussion**

The biomedical literature is a repository of our accumulated biomedical knowledge. Much of the value contained in this resource is locked in free-text and is therefore not in a form that is easily amenable
The table below lists the signaling pathways and the number of genes in common:

| Signaling Pathways                              | Number of genes in common |
|-------------------------------------------------|---------------------------|
| Actin Cytoskeleton Signaling                    | 2                         |
| Acute Phase Response Signaling                  | 7                         |
| Amyotrophic Lateral Sclerosis Signaling         | 1                         |
| Antigen Presentation Pathway                    | 1                         |
| Aryl Hydrocarbon Receptor Signaling             | 5                         |
| B Cell Receptor Signaling                       | 1                         |
| BMP signaling pathway                           | 1                         |
| Calcium Signaling                               | 1                         |
| Cell Cycle: G1/S Checkpoint Regulation          | 1                         |
| Geramide Signaling                              | 1                         |
| Circadian Rhythm Signaling                      | 1                         |
| Coagulation System                              | 3                         |
| Complement System                               | 2                         |
| Dopamine Receptor Signaling                     | 2                         |
| Eicosanoid Signaling                            | 2                         |
| ERK/MAPK Signaling                              | 1                         |
| Erythropoietin Signaling                        | 2                         |
| Estrogen Receptor Signaling                     | 2                         |
| Fc Epsilon Signaling                            | 3                         |
| FXR/RXR Activation                              | 4                         |
| Glucocorticoid Receptor Signaling               | 9                         |
| Glutamate Receptor Signaling                    | 1                         |
| Hepatic Cholestasis                             | 1                         |
| Hepatic Fibrosis/Hepatic Stellate Cell Activation| 9                         |
| Huntington's Disease Signaling                  | 3                         |
| Hypoxia Signaling in the Cardiovascular System  | 2                         |
| It-10 Signaling                                 | 4                         |
| It-2 Signaling                                  | 2                         |
| It-4 Signaling                                  | 3                         |
| It-6 Signaling                                  | 3                         |
| Insulin Receptor Signaling                      | 1                         |
| Interferon Signaling                            | 1                         |
| LPS/L-1 Mediated Inhibition of RXR Function      | 3                         |
| LXR/RXR Activation                              | 5                         |
| Mitochondrial Dysfunction                       | 4                         |
| NRF2-mediated Oxidative Stress Response         | 3                         |
| PS3                                             | 3                         |
| PPAR Signaling                                  | 1                         |
| PPARa/RXRa Activation                           | 4                         |
| PXR/RXR Activation                              | 1                         |
| RAR Activation                                  | 1                         |
| Role of BRCA1 in DNA Damage Response            | 1                         |
| T Cell Receptor Signaling                       | 1                         |
| Toll-Like Receptor Signaling                    | 1                         |
| TR/RXR Activation                               | 2                         |
| VDR/RXR Activation                              | 5                         |
| Xenobiotic Metabolism Signaling                 | 1                         |
| T Cell Receptor Signaling                       | 1                         |
| Toll-Like Receptor Signaling                    | 1                         |
| TR/RXR Activation                               | 2                         |
| VDR/RXR Activation                              | 5                         |
| Xenobiotic Metabolism Signaling                 | 1                         |

Table 2: The signaling pathways reviewed and the number of genes associated with each pathway from this potentially novel cancer-related dataset.

We examined all of the genes related to more than three cancers. In this analysis, we identified 10 genes related to thirty or more cancers, 72 genes related to twenty or more cancers, and 191 genes related to ten or more cancers. It is possible that the genes in each of these two will serve to help researchers identify the cure for cancer. We then mapped the top 200 genes to known metabolic and signaling pathways. The three pathways that were most highly correlated with the sets of potentially novel genes – cancer diagnosis relationships were all inflammatory pathways (i.e., the Acute Phase Response Signaling Pathway, the Glucocorticoid Signaling Pathway and the Hepatic Fibrosis Pathway). This highlights the importance of inflammation as a stimulus for the transformation of normal tissue to malignancy. Perhaps the common basis for transformation of cells to malignancies has already been published but remains “hidden” within synergies between articles within the vast amounts of biomedical literature. This NLP-based data mining experiment shows the utility of data mining the literature for scientific discovery. The authors believe that the entire medical literature should be exposed in this way to provide improved computable access to the knowledge contained in the text of the biomedical literature. Future research should include the addition of other journals content in the compendium used to find synergies across articles from the biomedical literature.

Our goal was to identify synergy between articles within the literature indicating potential relationships between genes or proteins and drugs or diseases that have heretofore not been previously recognized. This generated a marker discovery database that is searchable from multiple perspectives such that a disease-oriented researcher could search the disease they are interested in and find all of the genes, proteins, and drugs associated with that disease organized by function. Alternatively, researchers could ask for all genes related to a medication. A clinical trialist might ask what new diseases might be treated by an existing medication. Researchers could either access this information with regard to known synergies (where the gene and disease have been mentioned in the same article) that verifies the utility of the algorithm or discontinuities (where the gene and disease have a functional relationship; however, the two entities have never previously been mentioned in the same journal article). This may indicate the possibility of a novel relationship that can then be taken back to the bench for further definition, identification, and research. Proteins and drugs, for example, that have a functional relationship but that have never been recognized to affect one another, may be candidates for further basic science analysis thereby leading to more rapid marker and treatment discovery. We see this as a potential and very promising method for improving the research productivity.

This longitudinal research program aims to support the development of novel clinical and translational methods that can encompass a wide range of techniques including new methods of phenotyping where one could use SNOMED CT to phenotype the patients described in this paper. New biomarkers for research are a potential output of this project. In addition, this project may benefit clinical informatics for longitudinal studies that aim to rapidly look at specific associations that may lead to either additional retrospective analysis or prospective...
clinical trials. The special deliverable of this project is an interface where one can search the already recognized (concordant) relationships, and also disjoint or previously undocumented relationships between genes or proteins and drugs or diseases. As organizing concepts are included, searching for classes (e.g. concepts like beta blockers for drugs or cardiovascular diseases) will be possible without having to articulate the individual sub-classes of information.

Although there are many other Ontologies available [16] and other literature which could be included in such an analysis, this study provides a proof of concept that useful synergies can be identified from the knowledge available across articles in the literature. A computable method for identifying these synergies has the potential to speed scientific discovery.

Biomedical Informatics has the potential to help us to uncover novel genetic linkage to human disease. This can lead to significant improvements in patient care with more swift knowledge discovery and translational research, bringing that knowledge more rapidly to the bedside and thereby empowering clinical implementation of personalized /individualized medicine.

**Acknowledgement**

The authors wish to thank Miss Priyadarshini K for her editorial support of our manuscript. This work has been supported in part by grants from Centers for Disease Control and Prevention (PH00022 and HK00014) and from the NIH NCRR, CTSA UL1RR029887.

**References**

1. Butte AJ, Chen R (2006) Finding disease-related genomic experiments within
an international repository: first steps in translational bioinformatics. AMIA Annu Symp Proc 2006: 106-10.

2. Elkin PL (2003) Primer on medical genomics part V: bioinformatics. Mayo Clin Proc 78: 57-64.

3. Husser CS, Buchhalter JR, Raffo OS, Shabo A, Brown SH et al. (2006) Standardization of microarray and pharmacogenomics data. Methods Mol Biol 316: 111-57.

4. Pritchard JK, Cox NJ (2002) The allelic architecture of human disease genes: common disease-common variant...or not? Hum Mol Genet 11: 2417-2423.

5. Butte AJ, Kohane IS (1999) Unsupervised knowledge discovery in medical databases using relevance networks. Proc AMIA Symp 1999: 711-5.

6. Grivell L (2002) Mining the biblome: searching for a needle in a haystack? New computing tools are needed to effectively scan the growing amount of scientific literature for useful information. EMBO Rep 3: 200-203.

7. The New England Journal of Medicine. Accessed April 13, 2010

8. Garland J (1962) A voice in the wilderness. The “New England Journal of Medicine” since 1812. Br Med J 1: 105-108.

9. Long Island University. Previous George Polk Award Winners. Accessed August 19, 2010

10. Cimino JJ, Zhu X (2006) The practical impact of ontologies on biomedical informatics. Yearb Med Inform 2006: 124-35.

11. Ceusters W, Elkin P, Smith B (2007) Negative findings in electronic health records and biomedical ontologies: A realist approach. Int J Med Inform 3: S326-33

12. Elkin PL, Brown SH, Husser CS, Bauer BA, Wahner-Roedler D et al.(2006) Evaluation of the Content Coverage of SNOMED CT: Ability of SNOMED Clinical Terms to Represent Clinical Problem Lists. Mayo Clin Proc 81: 741-748.

13. Brown SH, Speroff T, Fielstein EM, Bauer BA, Wahner-Roedler DL, et al. (2006) eQuality: electronic quality assessment from narrative clinical reports. Mayo Clin Proc 81:1472-1481.

14. Elkin PL, Tuttle MS, Trusko BE, Brown SH (2009) Bioprospecting: novel marker discovery obtained by mining the bibleome. BMC Bioinformatics 10 Suppl 2: S9.

15. Ingenuity Systems. Redwood City, CA. Accessed August 20, 2010.

16. Whetzel PL, Noy NF, Shah NH, Alexander PR, Nyulas C, et al. (2011) BioPortal: enhanced functionality via new Web services from the National Center for Biomedical Ontology to access and use ontologies in software applications. Nucleic Acids Res 39: W541-545.