Mining PubMed for Biomarker-Disease Associations to Guide Discovery

Walter J. Jessen, Katherine T. Landschulz, Thomas G. Turi and Rachel Y. Reams
Covance Biomarker Center of Excellence, Discovery and Translational Services, Greenfield, Indiana

Introduction

Biomedical knowledge is growing exponentially; however, meta-knowledge around the data is often lacking. PubMed is a database comprising more than 21 million citations for biomedical literature from MEDLINE and additional life science journals dating back to the 1950s. To explore the use and frequency of biomarkers across human disease, we mined PubMed for biomarker-disease associations. We then ranked the top 100 linked diseases by relevance and mapped them to medical subject headings (MeSH) and, subsequently, to the Disease Ontology. To identify biomarkers for each disease, we queried Covance BioPathways, an online data resource that maps commercial biomarker assays to biological and disease pathways. We then integrated pathways-based information to describe both known and potential biomarkers as well as disease-associated genes/proteins for select diseases. This approach identifies therapeutic areas with candidate or validated biomarkers, and highlights those areas where a paucity of biomarkers exists.

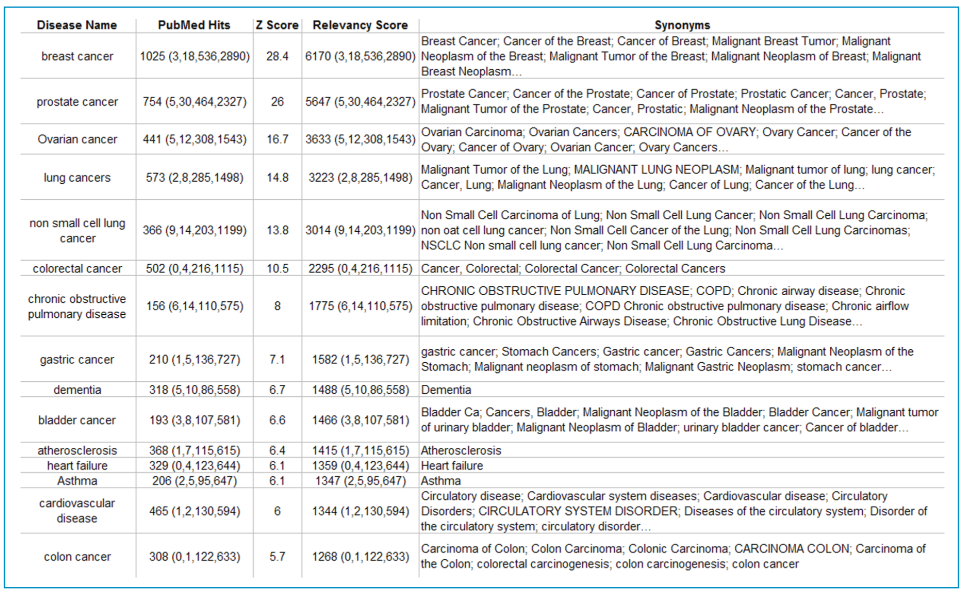
Materials and Methods

Text mining was performed using PolySearch, a web-based text mining system for extracting relationships between human diseases, genes, mutations, drugs and metabolites [Cheng et al., 2008]. The MeSH Browser (2012 MeSH) was used to map disease associations to MeSH IDs. Once MeSH IDs were assigned, the Disease Ontology was used to map DOIDs [Schriml et al., 2011]. Interaction networks were constructed in GeneGo MetaCore [Ekins et al., 2007] using the Auto expand algorithm, which gradually expands sub-networks around every object from the seed object list based on interactions identified in the literature. At every step, preference is given to objects with more connectivity to the initial object, and expansion halts when the sub-networks intersect, or when the overall network size reaches a predefined limit. Genes/proteins for which validated commercial assays exist were identified using Covance BioPathways at http://www.Covance.com/BioPathways and are indicated with a red dot ●. These genes/proteins can be considered potential biomarkers.

Results

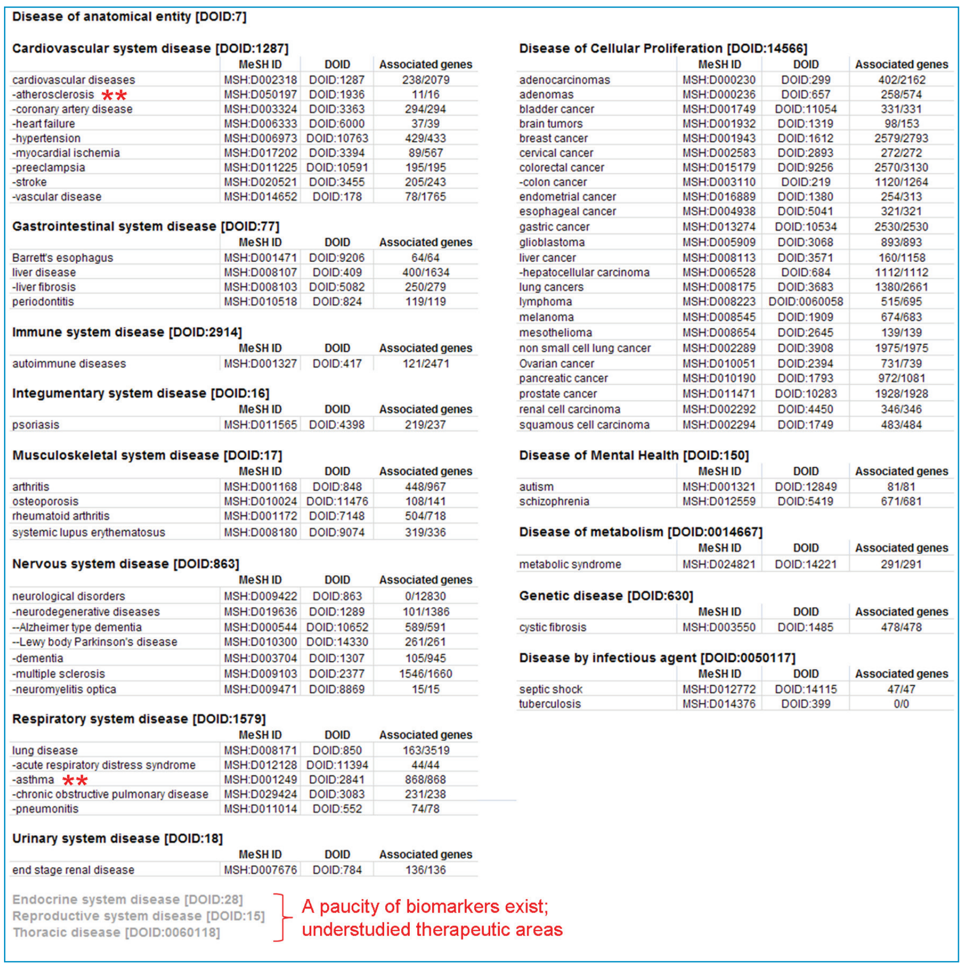
Data Extraction and Curation

In June 2011, we mined PubMed for term ("biomarker")-disease associations and identified a total of 1,181 disease associations (Table 1). We then curated the top 100 disease associations from the list, mapping each result to both medical subject (MeSH) ID and Disease Ontology ID (DOID), and then subsequently queried the GeneGo diseases ontology for associated biomarkers (Table 2). Of 100 results, 62 map to both MeSH ID and DOID and are shown below.



Disease Name	PubMed Hits	Z Score	Relevancy Score	Synonyms
breast cancer	1025 (3,18,536,2890)	28.4	6170 (3,18,536,2890)	Breast Cancer; Cancer of the Breast; Cancer of Breast; Malignant Breast Tumor; Malignant Neoplasm of the Breast; Malignant Tumor of the Breast; Malignant Neoplasm of Breast; Malignant Breast Neoplasm...
prostate cancer	754 (5,30,464,2327)	26	5647 (5,30,464,2327)	Prostate Cancer; Cancer of the Prostate; Cancer of Prostate; Prostatic Cancer; Cancer, Prostate; Malignant Tumor of the Prostate; Cancer, Prostatic; Malignant Neoplasm of the Prostate...
Ovarian cancer	441 (5,12,308,1543)	16.7	3633 (5,12,308,1543)	Ovarian Carcinoma; Ovarian Cancers; CARCINOMA OF OVARY; Ovary Cancer; Cancer of the Ovary; Cancer of Ovary; Ovarian Cancer; Ovary Cancers...
lung cancers	573 (2,8,285,1498)	14.8	3223 (2,8,285,1498)	Malignant Tumor of the Lung; MALIGNANT LUNG NEOPLASM; Malignant tumor of lung; lung cancer; Cancer, Lung; Malignant Neoplasm of the Lung; Cancer of Lung; Cancer of the Lung...
non small cell lung cancer	366 (9,14,203,1199)	13.8	3014 (9,14,203,1199)	Non Small Cell Carcinoma of Lung; Non Small Cell Lung Cancer; Non Small Cell Lung Carcinoma; non oat cell lung cancer; Non Small Cell Cancer of the Lung; Non Small Cell Lung Carcinomas; NSCLC Non small cell lung cancer; Non Small Cell Lung Carcinoma...
colorectal cancer	502 (0,4,216,1115)	10.5	2295 (0,4,216,1115)	Cancer, Colorectal; Colorectal Cancer; Colorectal Cancers
chronic obstructive pulmonary disease	156 (6,14,110,575)	8	1775 (6,14,110,575)	CHRONIC OBSTRUCTIVE PULMONARY DISEASE; COPD; Chronic airway disease; Chronic obstructive pulmonary disease; COPD Chronic obstructive pulmonary disease; Chronic airflow limitation; Chronic Obstructive Airways Disease; Chronic Obstructive Lung Disease...
gastric cancer	210 (1,5,136,727)	7.1	1582 (1,5,136,727)	gastric cancer; Stomach Cancers; Gastric cancer; Gastric Cancers; Malignant Neoplasm of the Stomach; Malignant neoplasm of stomach; Malignant Gastric Neoplasm; stomach cancer...
dementia	318 (5,10,86,558)	6.7	1488 (5,10,86,558)	Dementia
bladder cancer	193 (3,8,107,581)	6.6	1466 (3,8,107,581)	Bladder Ca; Cancers, Bladder; Malignant Neoplasm of the Bladder; Bladder Cancer; Malignant tumor of urinary bladder; Malignant Neoplasm of Bladder; urinary bladder cancer; Cancer of bladder...
atherosclerosis	368 (1,7,115,615)	6.4	1415 (1,7,115,615)	Atherosclerosis
heart failure	329 (0,4,123,644)	6.1	1359 (0,4,123,644)	Heart failure
Asthma	206 (2,5,95,647)	6.1	1347 (2,5,95,647)	Asthma
cardiovascular disease	465 (1,2,130,594)	6	1344 (1,2,130,594)	Circulatory disease; Cardiovascular system diseases; Cardiovascular disease; Circulatory Disorders; CIRCULATORY SYSTEM DISORDER; Diseases of the circulatory system; Disorder of the circulatory system; circulatory disorder...
colon cancer	308 (0,1,122,633)	5.7	1268 (0,1,122,633)	Carcinoma of Colon; Colon Carcinoma; Colonic Carcinoma; CARCINOMA COLON; Carcinoma of the Colon; colorectal carcinogenesis; colon carcinogenesis; colon cancer

Table 1. A representative list of term ("biomarker")-disease associations mined from PubMed in June 2011. The top 100 disease associations were ranked by Z Score. The Z-score indicates the number of standard deviations that the relevancy score is above the mean; larger Z-scores denote stronger associations. The top 100 data set is available under the Open Data Commons Attribution License at http://BiomarkerCommons.org.



Disease of anatomical entity [DOID:7]

Cardiovascular system disease [DOID:1287]

	MeSH ID	DOID	Associated genes
cardiovascular diseases	MSH:D002318	DOID:1287	238/2079
-atherosclerosis **	MSH:D050197	DOID:1936	11/16
-coronary artery disease	MSH:D003324	DOID:3363	294/294
-heart failure	MSH:D006333	DOID:6000	37/39
-hypertension	MSH:D006973	DOID:10763	429/433
-myocardial ischemia	MSH:D017202	DOID:3394	89/567
-preeclampsia	MSH:D011225	DOID:10591	195/195
-stroke	MSH:D020521	DOID:3455	205/243
-vascular disease	MSH:D014652	DOID:178	78/1765

Gastrointestinal system disease [DOID:77]

	MeSH ID	DOID	Associated genes
Barrett's esophagus	MSH:D001471	DOID:9206	64/64
liver disease	MSH:D008107	DOID:409	400/1634
-liver fibrosis	MSH:D008103	DOID:5082	250/279
periodontitis	MSH:D010518	DOID:824	119/119

Immune system disease [DOID:2914]

	MeSH ID	DOID	Associated genes
autoimmune diseases	MSH:D001327	DOID:417	121/2471

Integumentary system disease [DOID:16]

	MeSH ID	DOID	Associated genes
psoriasis	MSH:D011565	DOID:4398	219/237

Musculoskeletal system disease [DOID:17]

	MeSH ID	DOID	Associated genes
arthritis	MSH:D001168	DOID:848	448/967
osteoporosis	MSH:D010024	DOID:11476	108/141
rheumatoid arthritis	MSH:D001172	DOID:7148	504/718
systemic lupus erythematosus	MSH:D008180	DOID:9074	319/336

Nervous system disease [DOID:863]

	MeSH ID	DOID	Associated genes
neurological disorders	MSH:D009422	DOID:863	0/12830
-neurodegenerative diseases	MSH:D019636	DOID:1289	101/1386
--Alzheimer type dementia	MSH:D000544	DOID:10652	589/591
--Lewy body Parkinson's disease	MSH:D010300	DOID:14330	261/261
-dementia	MSH:D003704	DOID:1307	105/945
-multiple sclerosis	MSH:D009103	DOID:2377	1546/1660
-neuromyelitis optica	MSH:D009471	DOID:8869	15/15

Respiratory system disease [DOID:1579]

	MeSH ID	DOID	Associated genes
lung disease	MSH:D008171	DOID:850	163/3519
-acute respiratory distress syndrome	MSH:D012128	DOID:11394	44/44
-asthma **	MSH:D001249	DOID:2841	868/868
-chronic obstructive pulmonary disease	MSH:D029424	DOID:3083	231/238
-pneumonitis	MSH:D011014	DOID:552	74/78

Urinary system disease [DOID:18]

	MeSH ID	DOID	Associated genes
end stage renal disease	MSH:D007676	DOID:784	136/136

Endocrine system disease [DOID:28]
Reproductive system disease [DOID:15]
Thoracic disease [DOID:0060118]

A paucity of biomarkers exist; understudied therapeutic areas

Disease of Cellular Proliferation [DOID:14566]

	MeSH ID	DOID	Associated genes
adenocarcinomas	MSH:D000230	DOID:299	402/2162
adenomas	MSH:D000236	DOID:657	258/574
bladder cancer	MSH:D001749	DOID:11054	331/331
brain tumors	MSH:D001932	DOID:1319	98/153
breast cancer	MSH:D001943	DOID:1612	2579/2793
cervical cancer	MSH:D002583	DOID:2893	272/272
colorectal cancer	MSH:D015179	DOID:9256	2570/3130
-colon cancer	MSH:D003110	DOID:219	1120/1264
endometrial cancer	MSH:D016889	DOID:1380	254/313
esophageal cancer	MSH:D004938	DOID:5041	321/321
gastric cancer	MSH:D013274	DOID:10534	2530/2530
glioblastoma	MSH:D005909	DOID:3068	893/893
liver cancer	MSH:D008113	DOID:3571	160/1158
-hepatocellular carcinoma	MSH:D006528	DOID:684	1112/1112
lung cancers	MSH:D008175	DOID:3683	1380/2661
lymphoma	MSH:D008223	DOID:0060058	515/695
melanoma	MSH:D008545	DOID:1909	674/683
mesothelioma	MSH:D008654	DOID:2645	139/139
non small cell lung cancer	MSH:D002289	DOID:3908	1975/1975
Ovarian cancer	MSH:D010051	DOID:2394	731/739
pancreatic cancer	MSH:D010190	DOID:1793	972/1081
prostate cancer	MSH:D011471	DOID:10283	1928/1928
renal cell carcinoma	MSH:D002292	DOID:4450	346/346
squamous cell carcinoma	MSH:D002294	DOID:1749	483/484

Disease of Mental Health [DOID:150]

	MeSH ID	DOID	Associated genes
autism	MSH:D001321	DOID:12849	81/81
schizophrenia	MSH:D012559	DOID:5419	671/681

Disease of metabolism [DOID:0014667]

	MeSH ID	DOID	Associated genes
metabolic syndrome	MSH:D024821	DOID:14221	291/291

Genetic disease [DOID:630]

	MeSH ID	DOID	Associated genes
cystic fibrosis	MSH:D003550	DOID:1485	478/478

Disease by infectious agent [DOID:0050117]

	MeSH ID	DOID	Associated genes
septic shock	MSH:D012772	DOID:14115	47/47
tuberculosis	MSH:D014376	DOID:399	0/0

Table 2. The curated list of disease associations minded from PubMed and organized by high-level Disease Ontology. Each specific disease association has a unique MeSH ID, DOID and number of associated genes as defined in the GeneGo MetaCore knowledgebase.

Disease Interaction Network and Biomarker Assay Identification

For illustrative purposes, we constructed an interaction network around disease-associated genes for two diseases–one with few associated genes (atherosclerosis) and one with many associated genes (asthma)–using a network building algorithm in GeneGo MetaCore. For each interaction network gene set, we then queried Covance BioPathways, a publicly accessible, web-based data source that integrates biological and disease pathway maps with validated Covance assays and antibody products, to identify commercially available biomarker assays.

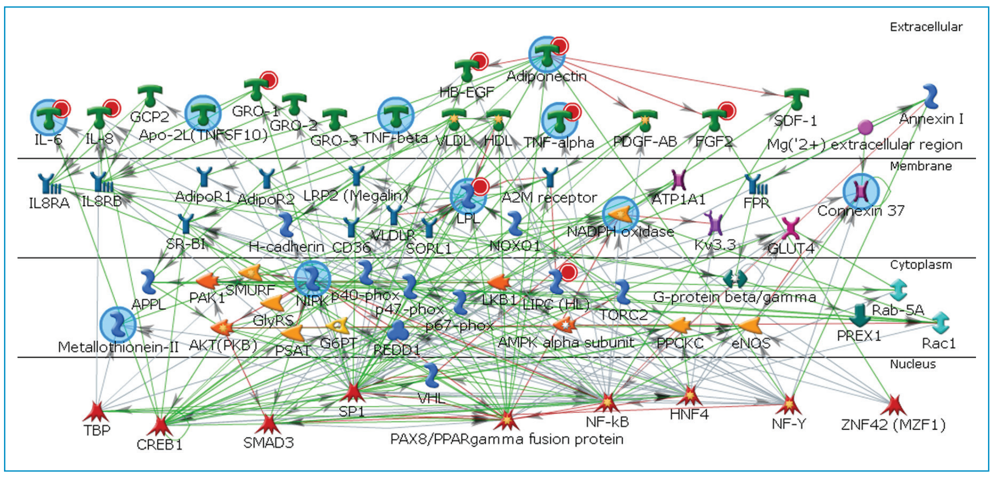
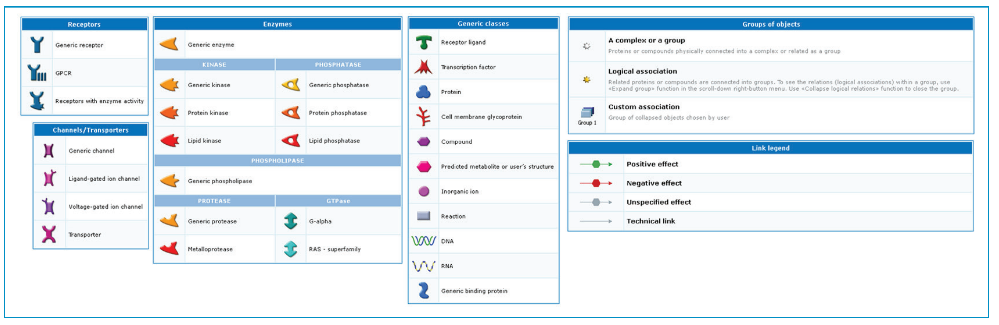


Figure 1. Atherosclerosis interaction network. Disease-associated genes are indicated with blue halos; genes without a halo were included by the network building algorithm. Biomarkers that have commercially validated assays are indicated with a red dot ● ; they either are known or can be considered potential atherosclerosis biomarkers.



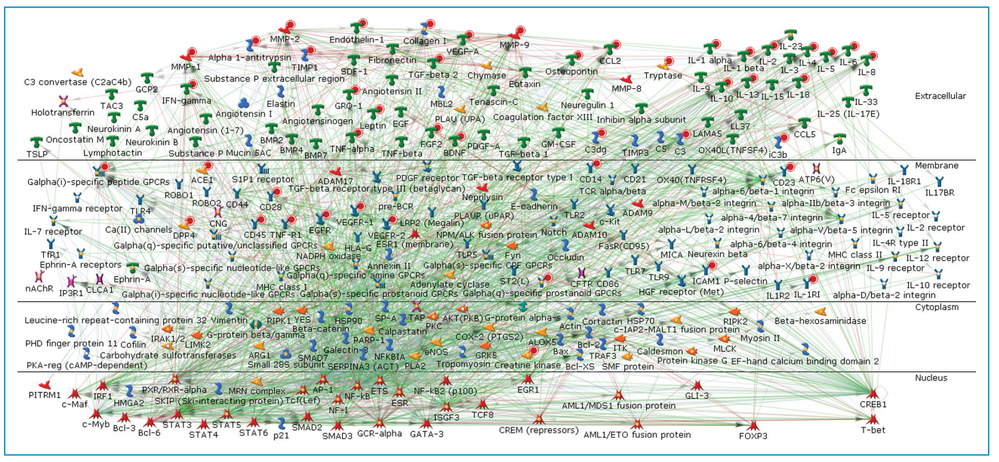


Figure 2. Asthma interaction network. All nodes shown are disease-associated genes. Biomarkers that have commercially validated assays are indicated with a red dot ● ; they are either known or can be considered potential asthma biomarkers.

Discussion

Given the molecular interdependencies within a cell, a disease is rarely a consequence of a single gene abnormality but instead reflects the perturbation of a complex network of biological and signaling pathways. The approach described here describes the detection and ranking of human disease based on research/clinical activity surrounding biomarkers. It also enables the identification of therapeutic areas with candidate or validated biomarkers. The strategy takes an integrative approach to identify candidate disease biomarkers by combining disease-associated genes/proteins with commercially validated assays for known biomarkers. We first constructed a system-level model of disease that incorporates molecular interactions across biological and signaling pathways. We then identified each gene/protein in the model that has an existing commercially validated assay. This research offers an alternative, comprehensive view of key relationships and pathway perturbations that may identify biomarkers of disease emergence or progression.