# Comorbidity of Bipolar Disorder with Substance Abuse: Selection of Prioritized Genes for Translational Research

**Raphael D. Isokpehi, PhD[1], Sharon A. Lewis, PhD[2], Tolulola O. Oyeleye, BS[1], Wellington K. Ayensu, MD[1], Tonya M. Gerald, PhD[3]**
**[1]Jackson State University, Jackson, MS; [2]Langston University, Langston, OK; [3]North Carolina Central University, Durham, NC**

## Abstract

*Bipolar disorder is a highly heritable mental illness. The global burden of bipolar disorder is complicated by its comorbidity with substance abuse. Several genome-wide linkage/association studies on bipolar disorder as well as substance abuse have focused on the identification and/or prioritization of candidate disease genes. A useful step for translational research of these identified/prioritized genes is to identify sets of genes that have particular kinds of publicly available data. Therefore, we have leveraged the availability of links to related resources in the Entrez Gene database to develop a web-based resource for selecting genes based on presence or absence in particular biological data resources. The utility of our approach is demonstrated using a set of 3,399 genes from multiple eukaryotes that have been studied in the context of bipolar disorder and/or substance abuse. A web resource to automate the selection of genes that contain certain database links is available at http://compbio.jsums.edu/bpd.*

## Introduction

Bipolar disorder (BPD) is a highly heritable, severe and chronic mental illness characterized by episodes of elation and high activity; alternating with periods of low mood and low energy[1,2]. This condition is less prevalent but more persistent and more impairing than major depressive disorder (MDD)[3]. Bipolar disorder poses a major challenge to the United States and the global healthcare system[4]. This burden is complicated by the comorbidity of bipolar disorder with narcotics and alcohol abuse[5].

Several studies on bipolar disorder as well as substance abuse have focused on the identification and/or prioritization of candidate genes for susceptibility[2,6,7]. Furthermore, the availability of data from genome-wide linkage/association studies and convergent functional genomics also continue to provide lists of genes associated with these diseases[8-10]. A useful step for translational research of these identified/prioritized genes is to identify sets of genes that have particular kinds of publicly available data[11].

We envisage that as genome-wide association studies of diseases continue to be published different researchers will be interested in different kinds of content and may want to intersect their own data types to see which genes have a combination of data types they are interested in. Therefore, we have leveraged the availability of links to clinical and molecular measurements as well as specialized databases in the Entrez Gene database to develop a web-based resource to automate selecting genes based on presence or absence in particular biological data resources. The utility of our approach is demonstrated using a set of genes that have been studied in the context of bipolar disorder and/or substance abuse.

The selection of prioritized gene sets for translational research on a disease can vary depending on the aspect of disease being studied[12,13]. For example, to investigate the genetic predisposition of women to predominance of depressive features in bipolar disorder, genes of interest may be those that show female-specific gene expression and contain Single Nucleotide Polymorphism (SNP) information. Knowledge that a gene has homolog in yeast or a rodent model organism may be relevant for molecular or genetic analysis of gene function. Furthermore, the availability of link to a database of images on gene expression in normal and diseased tissues could be useful to understand changes in gene expression during disease progression.

There are over 23,000 PubMed citations annotated with the Medical Subject Heading (MeSH) term: Bipolar Disorder. Furthermore, the MiSearch Adaptive PubMed Tool[14] retrieved over 170,000 citations for "substance abuse". We have compiled a list of over 3000 genes from multiple organisms that have been mapped to an integrated dataset of close to 200,000 curated PubMed citations on bipolar disorder and/or substance abuse. Furthermore, to facilitate simplified, user-defined selection of genes studied in the context of bipolar disorder and/or

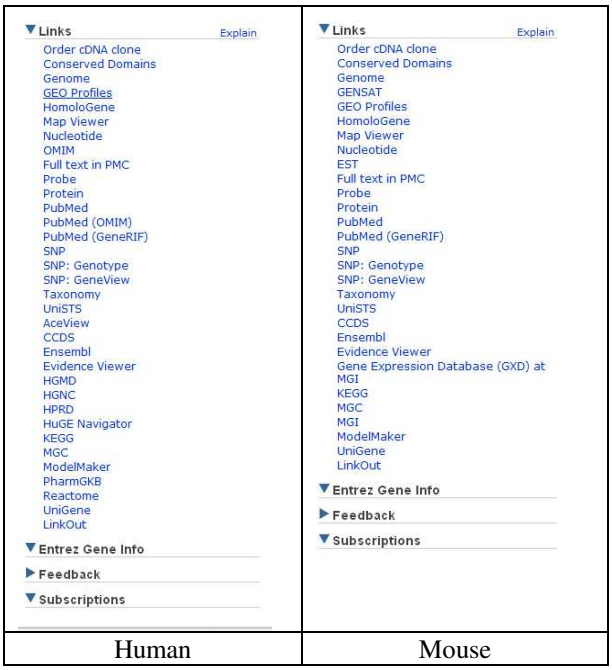
substance abuse, each gene was tagged with a 60-digit binary signature. The signature encodes the presence or absence of selected links from the Entrez Gene gene information record to complex molecular and clinical measurements as well as specialized databases. A web-resource at http://compbio.jsums.edu/bpd was developed to enable the pattern mining of the gene-link binary matrix of the signature collection.

## Methods

**Compilation of Multi-Organism Gene Set on Bipolar Disorder and Substance Abuse.** A non-redundant list of Medical Subject Heading (MeSH) curated PubMed citations were obtained by integrating the search results on the PubMed literature database[15] obtained with the following texts separately: "Bipolar Disorder" and "Substance Abuse". Genes mapped to each citation were extracted from the 'gene2pubmed.gz' file available from the Entrez Gene download website on September 9, 2008. We realized that the genes retrieved from a mapping of PubMed citation to Entrez Gene could be from genomes other than the human genome. Thus, in order to obtain an enriched set of genes, the putative homologous genes reported in the HomoloGene record for each gene was extracted.

**Selection of Links to Molecular and Clinical Measurements; and Specialized Databases from Entrez Gene Records.** The name of databases under the "Links" section of the each Entrez Gene record (Figure 1) was programmatically extracted. Links with more that 3 gene records were selected. In addition, links that provide similar information and have identical number of records were removed from the Links Set. For example, in our dataset SNP and SNP: GeneView had identical record count.

**Binary-encoding the Availability of Database Links for Genes.** A binary-encoding strategy was used to obtain a comprehensive integrative view of how the links are distributed across the gene set. Therefore, for each gene, the presence (encoded as 1) or absence (encoded 0) of a selected link was determined and then used to construct a signature whose digit length is equal to the number of selected links. The signatures were then collated into a binary matrix which was then mined for patterns.




Human | Mouse

**Figure 1**. Differences in Links count and types on Entrez Gene record page to molecular and clinical measurements as well as specialized databases for catechol-O-methyltransferase (COMT) gene of human (GeneID: 1312) and mouse (GeneID: 12846).

**Use Case Scenario of Pattern Mining of Binary Matrix of Genes and Links.** A web resource to allow for selection of genes based on the availability of links to selected resources in the Entrez Gene record. We used the interface to search for genes that have PubChem BioAssay link.

## Results

**Compilation of Multi-Organism Gene Set on Bipolar Disorder and Substance Abuse.** The search for MeSH curated PubMed citations on "Bipolar Disorder" and "Substance Abuse" yielded 23,253 and 172,988 citations respectively. An integration of the two sets of PubMed Identifiers (PMID) resulted in a total of 194,675 unique PubMed citations which mapped to 519 genes in the Entrez Gene database. Enrichment of this gene set with putative homologs available in the HomoloGene database resulted in 3,399 genes from 21 eukaryotic organisms (Table 1).

**Selection of Links to Molecular and Clinical Measurements; and Specialized Databases from Entrez Gene Records.** A total of 60 database links met our screening criteria (Table 2). The gene count associated with the databases range from 4 to 3,362. The Top 20 databases based on gene count is presented in Table 3.

| Organism | Gene Count |
|---|---|
| *Anopheles gambiae* str. PEST | 124 |
| *Arabidopsis thaliana* | 46 |
| *Ashbya gossypii* ATCC 10895 | 22 |
| *Bos taurus* | 326 |
| *Caenorhabditis elegans* | 101 |
| *Canis lupus familiaris* | 330 |
| *Danio rerio* | 321 |
| *Drosophila melanogaster* | 142 |
| *Gallus gallus* | 290 |
| *Homo sapiens* | 388 |
| *Kluyveromyces lactis* NRRL Y-1140 | 30 |
| *Macaca mulatta* | 6 |
| *Magnaporthe grisea* 70-15 | 36 |
| *Mus musculus* | 433 |
| *Neurospora crassa* | 36 |
| *Oryza sativa* Japonica Group | 44 |
| *Pan troglodytes* | 294 |
| *Plasmodium falciparum* 3D7 | 10 |
| *Rattus norvegicus* | 352 |
| *Saccharomyces cerevisiae* | 31 |
| *Schizosaccharomyces pombe* | 37 |

**Table 1.** Number of genes from organisms in bipolar disorder and substance abuse gene set.

**Binary-encoding the Availability of Database Links for Genes.** Each gene record in Entrez Gene was processed for evidence for the presence or absence of the selected 60 database links. The result of the search was encoded as a binary signature. The collection of signatures referred to as a binary matrix consisted of 989 unique binary signatures.

**Pattern Mining of Binary Matrix of Genes and Links.** The web resource for selecting genes and defining gene sets by binary signatures is available at http://compbio.jsums.edu/bpd. To demonstrate the potential utility of our approach to translational biomedical research, we queried the database availability matrix for genes with evidence of links to the NCBI PubChem BioAssay database (Binary Digit 3) that provides information on bioactivity screens of substances in the PubChem database. A total of 9 genes were retrieved including estrogen receptor 1 (ESR1) and 5-hydroxytryptamine (serotonin) receptor 1A (HTR1A) (Table 4). The database link profiles of these 9 genes were visualized using Matrix2png software [16] are presented in Figure 2.

| Digit for Binary Signature and Database |
|---|
| [1] AceView; [2] ArkDB; [3] BioAssay; [4]Books; [5] CCDS; [6] Conserved Domains; [7] Cytochrome P450 Homepage; [8] Ensembl; [9] EST; [10] Evidence Viewer; [11] FLYBASE; [12] Full text in PMC; [13] GeneDB; [14] Gene Expression Database at ZFIN; [15] Gene Expression Database (GXD) at MGI; [16] Genome; [17] GENSAT; [18] GEO Profiles; [19] GSS; [20] HbVar: A Database of Human Hemoglobin Variants and Thalassemias; [21] HGMD; [22] HGNC; [23] HomoloGene; [24] HPRD; [25] HuGE Navigator; [26] INRA; [27] Integrated X Chromosome Database (IXDB); [28] KEGG; [29] LinkOut; [30] Map Viewer; [31] MGC; [32] MGI; [33] MIPS; [34] ModelMaker; [35] Nucleotide; [36] OMIA; [37] OMIM; [38] Order cDNA clone; [39] PharmGKB; [40] Probe; [41] Protein; [42] PubMed; [43] PubMed (GeneRIF); [44] PubMed (OMIM); [45] RATMAP; [46] Reactome; [47] RGD; [48] SGD; [49] SNP; [50] SNP: Genotype; [51] SNP VarView; [52] TAIR; [53] Taxonomy; [54] TIGR; [55] UCSC; [56] UniGene; [57] UniSTS; [58] VectorBase; [59] WormBase; [60] ZFIN |

**Table 2.** Selected database links from Entrez Gene database.

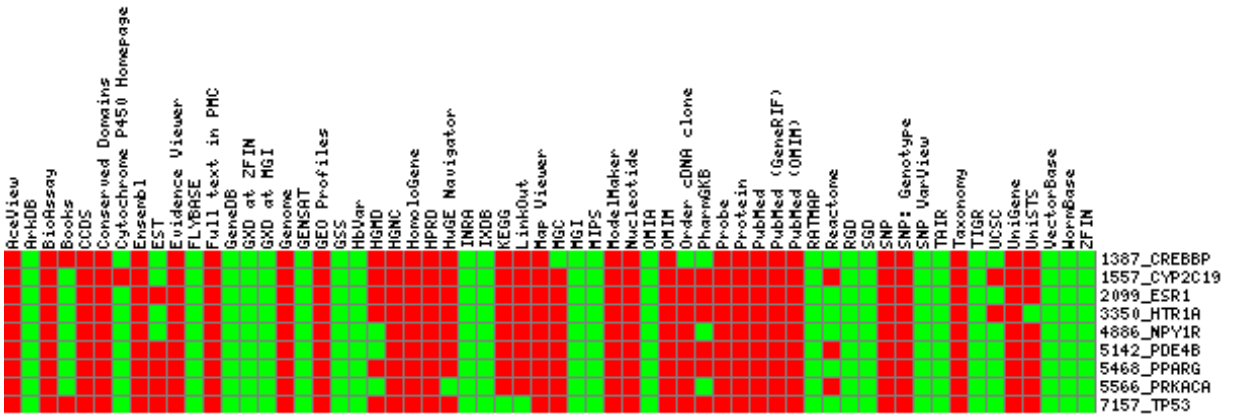



**Figure 2.** Visualization of profiles of database links profiles of genes mapped to PubMed citations on bipolar disorder and/or substance abuse. Column labels are databases while row labels are Entrez GeneID and Official Gene Symbol. Red Box: Presence of Database Link. Green Box: Absence of Database Link.

| Database | Gene count | Digit in Signature* |
|---|---|---|
| Map Viewer | 3362 | 30 |
| Taxonomy | 3352 | 53 |
| HomoloGene | 3307 | 23 |
| Nucleotide | 3285 | 35 |
| Protein | 3279 | 41 |
| Genome | 3203 | 16 |
| Conserved Domains | 3063 | 6 |
| LinkOut | 2808 | 29 |
| KEGG | 2780 | 28 |
| Evidence Viewer | 2664 | 10 |
| ModelMaker | 2664 | 34 |
| UniGene | 2503 | 56 |
| PubMed | 2349 | 42 |
| GEO Profiles | 2064 | 18 |
| SNP | 1765 | 49 |
| Full text in PMC | 1729 | 12 |
| Ensembl | 1579 | 8 |
| SNP: Genotype | 1411 | 50 |
| Probe | 1400 | 40 |
| UniSTS | 1325 | 57 |

**Table 3.** Top 20 databases with "Links" for Bipolar Disorder and Substance Abuse gene set. *Number refers to the position of the database in the binary signature.

| Entrez GeneID | Gene Symbol | Gene Description |
|---|---|---|
| 1387 | CREBBP | CREB binding protein |
| 1557 | CYP2C19 | cytochrome P450, family 2, subfamily C, polypeptide 19 |
| 2099 | ESR1 | estrogen receptor 1 |
| 3350 | HTR1A | 5-hydroxytryptamine (serotonin) receptor 1A |
| 4886 | NPY1R | neuropeptide Y receptor Y1 |
| 5142 | PDE4B | phosphodiesterase 4B, cAMP-specific (phosphodiesterase E4 dunce homolog, Drosophila) |
| 5468 | PPARG | peroxisome proliferator-activated receptor gamma |
| 5566 | PRKACA | protein kinase, cAMP-dependent, catalytic, alpha |
| 7157 | TP53 | tumor protein p53 |

**Table 4.** Genes from Bipolar Disorder and Substance Abuse gene set that have links to PubChem Bioassay database.

**Discussion**

A useful step to alleviate the burden of comorbidity of bipolar disorder with substance abuse is to evaluate availability of public domain data for candidate or prioritize genes. In this study, we have provided an approach that involves extracting genes mapped to literature, enriched the gene count with putative homologs and then developed an integration strategy based on availability database links from the Entrez Gene record. The investigation was not focused on the analysis of any genetic or polymorphism data or gene function associated with the genes but the modeling of the information associated with genes studied in bipolar disorder and/or substance abuse literature.

Binary-based models of complex biological information systems provides a mechanism to access and synthesize the wealth of multi-modal and multi-dimensional biological data recorded in complex databases such as Entrez Gene for a gene set of interest. When the binary values of variables are combined they form a binary signature for a label and a collection of these signatures becomes a binary matrix. Several advantages offered by the binary integration of high-throughput data include computational efficiency and noise resilience[17,18].

We have used MeSH curated PubMed citations on bipolar disorder and substance abuse to extract genes associated with the citations. Furthermore, we enriched the gene set from 519 to 3,399 by obtaining putative homologs. The enriched gene set will maximize the potential to extract novel information from the diverse databases linked to Entrez Gene. Inclusion of phenotype information on mouse improved the prioritization of human disease candidate genes[12]. In an era where animal models of human disease are increasingly sought, our gene set includes genes from model organisms for biomedical research (Table 1).

The integration strategy provided a more comprehensive view of the relationships in the gene set. For example, we were able to identify genes that have been studied in the context of chemical substance bioactivity (Table 4). Furthermore, the visualization of a 9 genes (Figure 2) with a PubChem BioAssay link revealed that are not curated in particular databases. For example, CREBBP, PDE4B and PRKACA lack representation in The Pharmacogenetics and Pharmacogenomics Knowledge Base (PharmaGKB)[19]. PRKACA is represented in the Reactome resource[20] but not in PharmaGKB. Thus our approach could be used to improve representation of genes in specialized biological databases.

The binary matrix can be accessed for user-defined queries and analysis through a searchable interface available at http://compbio.jsums.edu/bpd.

## Conclusion

Integrative biomedical translational research requires pattern mining strategies that facilitate the discovery of novel relationships from disparate datasets. We have presented a binary-based integration strategy that have captured and integrated the availability of database links in records of genes relevant bipolar disorder and substance abuse.


## Acknowledgements

Research Centers in Minority Institutions (RCMI) - Center for Environmental Health at Jackson State University (NIH-NCRR G122R13459-09); Langston University; Oklahoma IdeA Network of Biomedical Research Excellence (OK-INBRE); OK-INBRE- (NIH P20RR016478-07; and National Center for Integrative Biomedical Informatics (NIH U54DA021519); Carnegie Mellon University, Pittsburgh Supercomputing Center (PSC) (NIH #5 T36 GM008789)


## References


1. Burmeister M, McInnis MG, Zollner S. Psychiatric genetics: progress amid controversy. Nat Rev Genet 2008;9:527-540.
2. Zandi PP, Zollner S, Avramopoulos D et al. Family-based SNP association study on 8q24 in bipolar disorder. Am J Med Genet B Neuropsychiatr Genet 2008;147B:612-618.
3. Kessler RC, Merikangas KR, Wang PS. Prevalence, comorbidity, and service utilization for mood disorders in the United States at the beginning of the twenty-first century. Annu Rev Clin Psychol 2007;3:137-158.
4. Kessler RC. The global burden of anxiety and mood disorders: putting the European Study of the Epidemiology of Mental Disorders (ESEMeD) findings into perspective. J Clin Psychiatry 2007;68 Suppl 2:10-19.
5. Frye MA, Salloum IM. Bipolar disorder and comorbid alcoholism: prevalence rate and treatment considerations. Bipolar Disord 2006;8:677-685.
6. McEachin RC, Keller BJ, Zandi PP et al. Prioritizing Disease Genes by Analysis of Common Elements (PDG-ACE). AMIA Annu Symp Proc 2007;1068.
7. McEachin RC, Keller BJ, Saunders EF et al. Modeling gene-by-environment interaction in comorbid depression with alcohol use disorders via an integrated bioinformatics approach. BioData Min 2008;1:2.
8. Le-Niculescu H, McFarland MJ, Ogden CA et al. Phenomic, convergent functional genomic, and biomarker studies in a stress-reactive genetic animal model of bipolar disorder and co-morbid alcoholism. Am J Med Genet B Neuropsychiatr Genet 2008;147B:134-166.
9. Peterson K. Biomarkers for alcohol use and abuse--a summary. Alcohol Res Health 2004;28:30-37.
10. Ross J, Berrettini W, Coryell W et al. Genome-wide parametric linkage analyses of 644 bipolar pedigrees suggest susceptibility loci at chromosomes 16 and 20. Psychiatr Genet 2008;18:191-198.
11. Burgun A, Bodenreider O. Accessing and integrating data and knowledge for biomedical research. Yearb Med Inform 2008;91-101.
12. Chen J, Xu H, Aronow BJ et al. Improved human disease candidate gene prioritization using mouse phenotype. BMC Bioinformatics 2007;8:392.
13. Polychronakos C. New applications of microarray data analysis: integrating genetics with 'Omics'. Organized by the Cambridge Healthtech Institute, 15-17 August 2007, Washington DC, USA. Pharmacogenomics 2008;9:15-17.
14. States DJ, Ade AS, Wright ZC et al. MiSearch Adaptive PubMed Search Tool. Bioinformatics 2008.
15. Wheeler DL, Barrett T, Benson DA et al. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res 2008;36:D13-D21.
16. Pavlidis P, Noble WS. Matrix2png: a utility for visualizing matrix data. Bioinformatics 2003;19:295-296.
17. Bowers PM, O'Connor BD, Cokus SJ et al. Utilizing logical relationships in genomic data to decipher cellular processes. FEBS J 2005;272:5110-5118.
18. Shmulevich I, Zhang W. Binary analysis and optimization-based normalization of gene expression data. Bioinformatics 2002;18:555-565.
19. Hernandez-Boussard T, Whirl-Carrillo M, Hebert JM et al. The pharmacogenetics and pharmacogenomics knowledge base: accentuating the knowledge. Nucleic Acids Res 2008;36:D913-D918.
20. Vastrik I, D'Eustachio P, Schmidt E et al. Reactome: a knowledge base of biologic pathways and processes. Genome Biol 2007;8:R39.