METHOD ARTICLE

Procedure and datasets to compute links between genes and phenotypes defined by MeSH keywords [version 1; peer review: 2 approved with reservations]

Erinija Pranckeviciene

Department of Human and Medical Genetics, Faculty of Medicine, Vilnius University, Santariskiu str. 2, LT-0866 Vilnius, Lithuania

v1 **First published:** 19 Feb 2015, **4**:47 (https://doi.org/10.12688/f1000research.6140.1)

Latest published: 19 Feb 2015, **4**:47 (https://doi.org/10.12688/f1000research.6140.1)

Abstract

Algorithms mining relationships between genes and phenotypes can be classified into several overlapping categories based on how a phenotype is defined: by training genes known to be related to the phenotype; by keywords and algorithms designed to work with disease phenotypes. In this work an algorithm of linking phenotypes to Gene Ontology (GO) annotations is outlined, which does not require training genes and is based on algorithmic principles of Genes to Diseases (G2D) gene prioritization tool. In the outlined algorithm phenotypes are defined by terms of Medical Subject Headings (MeSH). GO annotations are linked to phenotypes through intermediate MeSH D terms of drugs and chemicals. This inference uses mathematical framework of fuzzy binary relationships based on fuzzy set theory. Strength of relationships between the terms is defined through frequency of co-occurrences of the pairs of terms in PubMed articles and a frequency of association between GO annotations and MeSH D terms in NCBI Gene gene2go and gene2pubmed datasets. Three plain tab-delimited datasets that are required by the algorithm are contributed to support computations. These datasets can be imported into a relational MySQL database. MySQL statements to create tables are provided. MySQL procedure implementing computations that are performed by outlined algorithm is listed. Plain tab-delimited format of contributed tables makes it easy to use this dataset in other applications.

Keywords

ontology , medical subject headings , MySQL , annotation, phentypes

Open Peer Review

Reviewer Status ? ?

	Invited Reviewers	
	1	**2**
version 1 published 19 Feb 2015	? report	? report

1 **Jason E. McDermott**, Pacific Northwest National Laboratory, Richland, USA

2 **Emidio Capriotti**, University of Bologna, Birmingham, Italy

Any reports and responses or comments on the article can be found at the end of the article.

Corresponding author: Erinija Pranckeviciene (erinija.pranckeviciene@mf.vu.lt)

Competing interests: Author declares no competing interests.

Grant information: The author(s) declared that no grants were involved in supporting this work.

How to cite this article: Pranckeviciene E. **Procedure and datasets to compute links between genes and phenotypes defined by MeSH keywords [version 1; peer review: 2 approved with reservations]** F1000Research 2015, **4**:47 (https://doi.org/10.12688/f1000research.6140.1)

First published: 19 Feb 2015, **4**:47 (https://doi.org/10.12688/f1000research.6140.1)

Introduction

Understanding molecular mechanisms underlying both normal cellular processes and disease-causing gene perturbations has numerous applications in clinical diagnostics, personal genomics and engineering[1–5]. Most of the genomic studies address two major questions: (i) *What genomic and molecular markers are associated with an observed phenotype?* (ii) *What molecular mechanisms lead to that phenotype in the studied organism?* Answering these questions and uncovering gene-phenotype relationships mostly relies on experimental research that has already generated very large amounts of high-throughput data stored in public databases[6–10]. New knowledge about genes and their functions is acquired all the time based on a constant gathering of genomic data. To date there are more than 1500 databases hosting various types of genomic and molecular biology data[11] acompanied by increasing number of research publications analyzing newly-generated data[12]. For this reason integrative algorithms to analyze high-throughput data by mining genomic databases and literature are in the focus of intensive research resulting in many publicly available bioinformatics tools for biologists and clinical researchers[6,13–19].

Biologists analyze lists of genes to dissect individual or collective gene involvement in the biological function that is being investigated, for example:

- functions of differentially expressed genes identified in a microarray or RNA-Seq experiment;
- relationships between a biological process of interest and target genes regulated by a transcription factor identified by ChIP-Seq experiment;
- causative relationships between functions of genes found in a chromosomal deletion or duplication identified in a patient and a clinical phenotype of the patient;
- identifying candidate genes from gene lists in literature and databases.

Finding meaningful relationships between genes in a large list and a phenotype by manually reviewing the literature and genomic databases is very laborious and time-consuming. Efforts to automate this process mostly have been directed towards the prioritization of human disease genes[20,21] and less for model organisms and general phenotypes[10]. Gene prioritization tools, that can be used to infer relationships between genes and phenotypes, differ from each other with respect to computational algorithms and data sources used in prioritization[21–23]. In computations, a definition of a phenotype will determine the rules by which the algorithm will mine available data resources to retrieve gene-phenotype links.

Phenotype definitions

The definition of a phenotype widely accepted in biology is *"the observable trait or the collection of traits of an organism resulting from the interaction of the genetic makeup of the organism and the environment"* meaning different things in different contexts[24,25]. In medicine the phenotype often refers to disease or abnormality[26]. In cellular contexts measurable cellular phenotypes are represented by features of cells such as the morphology (shape, size), the behavior (motility, growth), the developmental stage, the expression of specific genes and the rate of bio-chemical reactions[8,27].

Specific vocabularies of phenotype terminology are implemented as ontologies containing concepts, the relationships between the concepts and the definitions of both[28,29]. Specialized phenotype vocabularies are available for model organisms[30,31], life sciences[32–36] and human diseases[37,38]. Phenotype can also be defined as a subset of genes known to be functionally related to the phenotype of interest, usually used in gene prioritization algorithms[23,39]. However, if the phenotype of interest hasn't been well studied and does not have genes linked to it, then it is difficult or even impossible to use this approach.

Terms of Medical Subject Headings (MeSH) vocabulary can serve as appropriate phenotypic descriptions[38]. MeSH terms are curated and are assigned to the articles in PubMed to adequately reflect the content of each article since they are meaningfully associated with the biological processes that they denote. Phenotypes in Mammalian Phenotype Ontology (MPO) used in Mouse Genome Informatics (MGI) database[40] are also mapped to MeSH terms.

Approaches to infer gene phenotype links

Gene prioritization tools have to establish links between genes and phenotypes by use of some algorithm. Several overlapping categories of tools can be distinguished based on how a phenotype is defined: by training genes known to be related to the phenotype; by keywords and tools designed to work with disease phenotypes. Table 1 lists maintained prioritization tools from these categories that are frequently cited in GoogleScholar. Algorithms defining phenotypes by training genes in prioritization evaluate similarity between training genes and candidate genes. Supervised machine learning (most often kernel methods) are used in this category of tools[39,41]. Algorithms describing phenotypes by keywords usually use frequencies of gene-associated documents that have keyword matches. Majority of algorithms are designed to prioritize genes with respect to disease phenotypes defined by either the keywords or the training genes or by both.

Short summary of most representative tools

Endeavour. In Endeavour the phenotype is defined by the training genes. It builds a phenotype model using different sources of genomic information derived from the training genes. Endeavour data sources consist of gene annotations, gene sequences, expressed sequence tags over multiple conditions, protein-protein interaction data and known transcription factor binding sites. The program works with the genes of human, mouse, rat, fly and worm organisms.

Table 1. Frequently cited gene prioritization tools.

Tools	Web Link
Endeavour[39,41,42]	*www.esat.kuleuven.be/endeavour/*
G2D[43]	*www.ogic.ca/projects/g2d_2/*
ToppGene[44,45]	*toppgene.cchmc.org/*
GeneWanderer[46]	*compbio.charite.de/genewanderer/*
MimMiner[47]	*www.cmbi.ru.nl/MimMiner/*
PolySearch[48]	*wishart.biology.ualberta.ca/polysearch/*
SUSPECTS[49]	*www.genetics.med.ed.ac.uk/suspects/*
PhenoPred[50]	*www.phenopred.org*
CANDID[51]	*dsgweb.wustl.edu/hutz/*
PosMed[52]	*omicspace.riken.jp*
GeneProspector[53]	*www.hugenavigator.net*

It builds the model of the phenotype using information of the training genes in each of the genomic sources. It ranks the candidate genes according to how well they compare with the built model. Individual rankings in the Endeavour are combined by the order statistics[42]. In the table of the ranked genes the explanations are provided about the genes.

ToppGene. The candidate gene prioritization is one of the functions provided by the ToppGene tool. The user submits a set of training genes and a set of the test genes. The ToppGene first finds the significantly enriched annotations for the training genes in multiple data sources: GO annotations, literature, Interaction, Pathway, human and mouse phenotype data, TF binding sites, Cytobands, Co-expression Atlas, Drugs, microRNA and more. The candidate genes are ranked by the similarity of their functional annotations to the enriched annotations in the training genes. The similarity is computed as fuzzy-based measure[54] or Pearson correlation coefficient. The user can examine the genes and the enriched terms of the the training set.

GeneWanderer. In GeneWanderer the candidate genes are retrieved from the genomic region given the genomic coordinates. The phenotype is defined either by the disease keyword or by the list of the training genes known to be related to the phenotype. If the phenotype is defined by the keyword then the known genes associated with it are retrieved. The tool measures the distance between the candidate genes and the training genes in the protein-protein interaction network. The tool is specific to the human diseases.

PolySearch. PolySearch allows queries in the form of: *Given X find all Y*. X and Y can be diseases, tissues, cell compartments, gene/protein names, SNPs, mutations, drugs and metabolites. If the phenotype is defined by keywords, then PolySearch retrieves the documents matching all keywords in the Pubmed, OMIM, DrugBank, Swiss-Prot, Human Mutation Database, Genetic Association Database and Human Protein Database. The ranked list of requested biomedical entities that are associated with the text of the query is returned. The score of the entity is proportional to the number of document matches in the databases. User can browse the results and examine the matching publications and sentences.

PosMed. Positional PubMed is the semantic engine that ranks biomedical entities by the statistical significance of the associations with the provided keywords. The strength of the associations between biomedical entities and keywords is based on the number of the documents they share. The document categories comprise PubMed (PubMed titles, abstracts and MeSH terms), REACTOME (Pathway information from REACTOME), Protein-protein interaction (Protein-Protein Interactions in Human and Mouse from IntAct and Arabidopsis from AtPID), Gene ontology, Human disease ontology, Mammalian phenotype ontology, Microarray based co-expression data for Arabidopsis. Given the keyword defining the phenotype and the type of biomedical entity to score (either gene or metabolite or drug) the PosMed returns list of the scored entities linked to the phenotype, sorted according to the strength of the connection between them. The PosMed supports human, mouse, rat, arabidopsis and rice organisms. The user can browse through all documents of the established links.

G2D. In G2D the disease phenotype is defined by the OMIM identifier which is mapped to the associated MeSH terms of the diseases. The candidate genes are selected from the provided genomic region possibly containing a marker associated with the disease phenotype. G2D establishes a chain of evidence connecting the disease phenotype to the genes by forming the links between the terms in MeSH and GO annotations. The MeSH terms of the disease (category C) are linked to the MeSH terms of the chemicals and drugs (category D) which are linked to the Gene Ontology annotations. The connections between the terms are established by computing the normalized frequency of PubMed documents in which the MeSH terms (C and D) occur together. The connection between the protein GO annotations and the MeSH D terms are established by computing the normalized frequency of cooccurrence of GO and MeSH D term in papers supporting experimental evidence for the GO annotation. The GO annotation is weighted by a combined score which is used for ranking of the candidate genes.

This inference is illustrated in Figure 1 through the example of exploring candidate genes associated with cleft lip phenotype. Association between the cleft lip and the rs987525 variant from region 8q24.21 has been replicated independently in several different populations[55] but no associated gene was found. G2D suggests the MYC gene as candidate. The link between this gene and the cleft lip disease phenotype is inferred through the relationship between the terms **“Craniofacial abnormalities”** and **“Homeodomain proteins”** and the relationship between the later term and the GO annotation a **“Sequence specific DNA binding transcription factor activity”** of the MYC gene. The MYC gene is regulated by the CTCF transcription factor[56] which has a binding site at the genomic location of rs987525 leading to a possible hypothesis that this SNP marker might be linked to the cleft lip through a regulatory interaction with the MYC gene[57]. Another connection between the BMP4 gene and cleft lip OMIM phenotype is inferred through the relationship between the **“Cleft lip”** term and the term **“Bone Morphogenetic Protein 4”** which is related to the GO annotation **“BMP signalling pathway”**. The BMP4 gene harbors the rs1957860 marker variant which is known to be associated with cleft lip[58].

Computing relationships between genes and phenotypes

Gene prioritization algorithms produce lists of the best candidates which are most strongly related to the phenotype of interest according to the criteria set by the algorithm. Rankings are based on evidence scores of relationships computed by the prioritization algorithm for each candidate gene. For generation of meaningful hypothesis it is important to know what factors led to the obtained rankings and links established between genes and phenotypes. In methods relying on phenotype definitions by training genes a detailed examination of such evidence is difficult. In multipurpose systems such as PolySearch and PosMed provided evidence lacks specificity. Most comprehensive in this respect is G2D in which OMIM phenotype is translated into adequate MeSH term of the disease. In this study an attempt is made to develop means to support computations linking genes and phenotypes defined by the MeSH terms extending beyond the diseases and building upon the algorithmic principles of G2D[43,59].

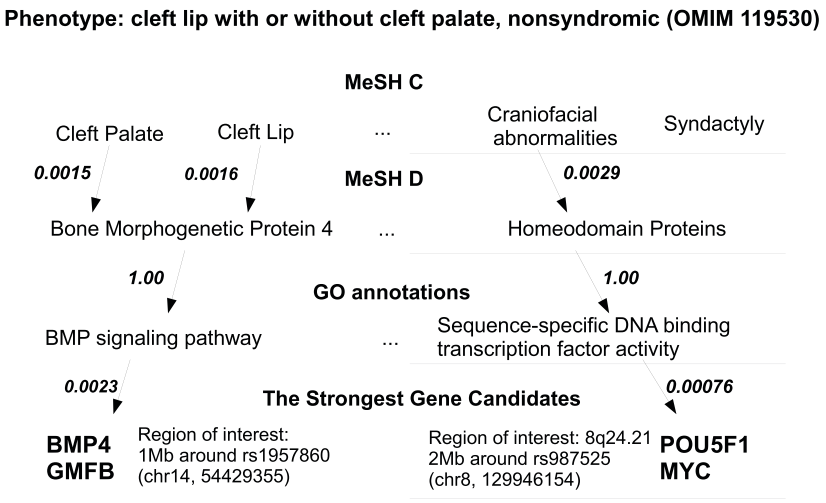

Phenotype: cleft lip with or without cleft palate, nonsyndromic (OMIM 119530)
MeSH C
Cleft Palate
Cleft Lip
...
Craniofacial abnormalities
Syndactyly
0.0015
0.0016
MeSH D
0.0029
Bone Morphogenetic Protein 4
...
Homeodomain Proteins
1.00
GO annotations
1.00
BMP signaling pathway
...
Sequence-specific DNA binding transcription factor activity
0.0023
The Strongest Gene Candidates
0.00076
BMP4
GMFB
Region of interest: 1Mb around rs1957860 (chr14, 54429355)
Region of interest: 8q24.21 2Mb around rs987525 (chr8, 129946154)
POU5F1
MYC

Figure 1. Connections computed by G2D in prioritization of genes with respect to the cleft lip phenotype.

Methods

It was shown in applications of Arrowsmith algorithm that biomedical knowledge can be discovered through finding hidden links between concepts in scientific literature. The concepts, co-occurring at high frequency in two disparate sets of literature articles, indicated meaningful links[60,61]. The link suggested that fish oil can reduce Raynaud's syndrome symptoms, later confirmed experimentally[62]. An inference leading to this result was "fish oil reduces blood viscosity, platelet aggregations and vascular re-activity which are increased in Raynaud's syndrome"[63]. In similar way algorithms, based on linking the concepts or entities in the collections of data, relate genes to phenotypes by using concept co-occurrences in literature and controlled vocabularies[43,64].

Links between phenotypes and gene GO annotations can be computed through intermediate links with chemicals as shown in G2D[59]. It is hypothesized that phenotype defined by the MeSH term can be meaningfully related to a subset of MeSH D terms denoting molecular entities of drugs and chemicals. Similarly, gene functions encoded by GO annotations can be meaningfully related to molecular entities denoted by MeSH D terms through related chemical processes affecting gene functions. Strengths of relationships can be derived from information in annotations of PubMed articles and NCBI datasets *gene2go* and *gene2pubmed*[65]. Figure 2 outlines the idea of the algorithm in which a phenotype and GO annotations are linked through chemicals.

Let us denote MeSH D terms pertaining to chemicals by d_j, $j = 1, \ldots, N$. A relationship $m(phenotype, chemical)$ between the phenotype defined by MeSH term g and the chemical defined by MeSH D term d_j is denoted by $m(g, d_j)$. Let us denote a relationship $m(chemical, GO\ annotation)$ between the MeSH D term d_j of chemical and GO annotation go_i, $i = 1, \ldots, M$ by $m(d_j, go_i)$. Values of the $m(g, d_j)$ and $m(d_j, go_i)$ relationships represent strengths of the connections between terms. The strengths of connections between the phenotype g and GO annotation go_i passing through the chemicals d_j, $j = 1, \ldots, N$ are computed as $w_{goi} = \max_j (m(g, d_j) \times m(d_j, go_i))$. These computed weights express the strength of association between the functional annotation go_i and the phenotype of interest. Table in a bottom panel of Figure 2 illustrates one possible way to order annotated genes by the magnitude of weights of their association to the phenotype of interest. Principles underlying the algorithm to compute strengths of relationships $m(phenotype, chemical)$ and $m(chemical, GO\ annotation)$ between phenotypes and functional gene annotations can be founded on fuzzy set theory (FST)[43,66]. Using mathematical framework of FST the relationships are defined as fuzzy binary relationships (FBRs) and can take a variety of forms[67]. A thorough explanation can be found in[68] on pages 69–84.

Definitions of relationships between MeSH terms and GO annotations

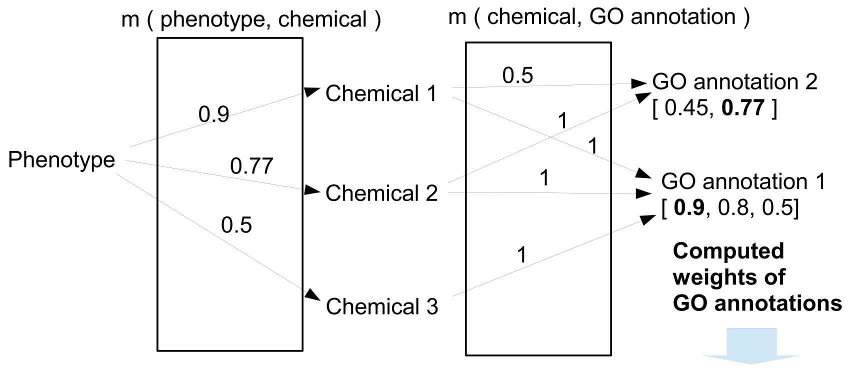
Let us denote phenotype MeSH terms as g_j, $j \in (1 \ldots NG)$ in which j refers to a particular MeSH term. Similarly, let us denote MeSH D terms by d_k, $k \in (1 \ldots ND)$. A subset of PubMed articles annotated by a specific g_j term is denoted by G_j. Similarly, a subset articles annotated by a particular term d_k is denoted by D_k. A fuzzy binary relation R_{GD} between two MeSH terms (g_j, d_k) is defined as:

$$\{[(g_j, d_k), m_{gd}(g_j, d_k)] \mid (g_j, d_k) \in G_{NG} \times D_{ND}\}, \qquad (1)$$

with membership function

$$m_{gd}(g_j, d_k) = \frac{|G_j \cap D_k|}{|D_k|}. \qquad (2)$$

The brackets $|\cdot|$ in Equation 2 denote the cardinality, a number of elements in a set, of an intersection $|G_j \cap D_k|$ of the two sets. The intersection represents a set of the articles annotated by both the g_j term and the d_k MeSH D term. FBR in Equation 1 is defined



Genes	Gene Ontology annotations			
Gene A	GO1 *w1=0.9*	GO2 *w2=0.77*		
Gene B	GO11 *w11=0.8*	GO12 *w12=0.5*	GO13 *w13=0.48*	GO14 *w14=0.2*
Gene C	GO101 *w101=0.79*	GO2 *w1=0.77*	GO103 *w103=0.15*	

Figure 2. Computation of relationships and weighting of GO annotations.

on all pairs of selected annotations in the universe of all articles annotated by those MeSH terms. The membership function in Equation 2 models a degree of inclusion of a narrower concept d_k (chemical) into a broader concept g_j (phenotype) $d_k \subseteq g_j$. The FBR of inclusion in a quantitative way defines a semantic relationship between meanings of the broader and narrower concepts[69,70].

Inclusion relationship between GO annotations and MeSH D terms is defined using a universe of genes instead of articles. Let us denote GO annotations by go_i, $i \in (1 \ldots NGO)$ in which i refers to a particular annotation. NGO is a total number of GO annotations of genes in *gene2go*. Let us denote by GO_i a subset of genes in *gene2go* annotated by a particular go_i. Let us denote by GD_k a subset of genes in *gene2pubmed* associated with articles, annotated by the MeSH D term $d_k \in (1 \ldots NGD)$, where NGD is total number of MeSH D terms associated with genes through articles. Fuzzy binary relation R_{DGO} between these terms is defined as:

$$\{[(d_k, go_i), m_{dgo}\,(d_k, go_i)] \mid (go_i, d_k) \in GO_{NGO} \times GD_{NGD}\}, \quad (3)$$

with membership function

$$m_{dgo}(d_k, go_i) = \frac{|GD_k \cap GO_i|}{|GO_i|}. \quad (4)$$

The degree of connection between GO annotation and MeSH D chemical in Equation 4 is determined by a number of genes sharing these two annotations over a number of genes annotated by that GO.

A relationship between GO annotation and phenotype defining MeSH term is computed by applying maximum composition operation $R_{GD} \circ R_{DGO}$ on fuzzy binary relations defined by Equation 1 and Equation 3 resulting in a following FBR:

$$\{[(g_j, go_i), \max(m_{gd}\,(g_j, d_k) * m_{dgo}(d_k, go_i))] \\ \mid go_i \in GO_{NGO}, d_k \in (D_{ND} \cap GD_{NGD}), g_j \in G_{NG}\}. \quad (5)$$

MySQL database and SQL procedures were created in order to experiment with and to support outlined inference[70]. The created datasets are limited to the annotated genes of human, mouse and fly organisms. The MeSH terms (mtree 2012) defining phenotypes are provided for the categories of Anatomy (A), Diseases (C), Drugs and chemicals (D) and Biological processes and phenomena (G). Information in the created datasets is as of September 2013.

Results

Datasets and procedure to compute links between genes and phenotypes

In this section three contributed data sets are described. These datasets and presented MySQL procedure support computations of links between phenotype and GO annotations outlined in a previous section. Datasets were created by using NCBI *E-utilities*[71] and custom scripts. The data sources (as of September 2013) used to create these datasets are described in Table 2. MeSH terms of category B are present but are not used to define phenotype in computation. The datasets are submitted in a format of tab delimited tables that can be imported into MySQL database or used as plain data. In this work a presented data management framework is based on MySQL.

Table 2. Data sources (content as of September 2013) and tables for computing gene phenotype relationships.

Data Source	Usage
mtree2012.bin	To have the full list of the MeSH terms with the corresponding category identifier. MeSH term categories A,B,C,D and G were used to retrieve the corresponding PubMed identifiers of the articles having co-occurring MeSH terms.
Articles and annotations retrieved by NCBI *E-utilities*	Article annotations were used to create a table of PMID counts for pairs of co-occurring MeSH annotations in articles.
gene2go	To collect annotated genes of the human, mouse and fly together with their GO annotations.
gene2pubmed	To retrieve MeSH D terms in the articles associated with the genes and to link GO annotations assigned to these genes with the retrieved MeSH D terms.
homolo.gene	To create table of homologous genes of the three organisms and their GO annotations.
Data table	**Size and description**
mesh_terms	9,725,157 rows store data pertaining to pairs of MeSH term of category A,B,C,D,G and MeSH D terms.
dterm_go	14,225,540 rows store data pertaining to pairs of MeSH D terms and GO annotations.
go_terms	20,266 rows store GO annotations of genes of human, mouse and fly organism.

Dataset 1. Table mesh_terms

http://dx.doi.org/10.5256/f1000research.6140.d43167

Table mesh_terms stores relationship information for pairs of MeSH terms. Each row corresponds to one pair of MeSH terms: a term of category A,C,D,G defining a phenotype (column mterm) and a term of category D defining a chemical (column dterm). Attributes of this relationship consist of number of articles in PubMed annotated by each of these MeSH terms separately (columns nm and nd contains number of articles annotated by mterm and dterm respectively), number of articles annotated by both terms (column inters), number of articles annotated by either term (column unio), computed strength of the relationship between the terms in the pair (column dscore) and comma separated list of PMID identifiers of the articles annotated by both terms. Column mid is an identifier of the row and dtid column is a key linking into dterm_go table. This table has 9,725,157 rows and 10 columns that are separated by tabs. Size of a plain table is 1.97GB. Compressed table takes 801MB. Information in this table is as of September 2013.

Dataset 2. Table dterm_go

http://dx.doi.org/10.5256/f1000research.6140.d43168

Table dterm_go stores relationship information for pairs of MeSH D terms and gene ontology annotations. Each row corresponds to one pair: a MeSH term of category D defining a chemical (column dterm) and a gene ontology annotation as a 10 character identifier of GO (column goterm). Attributes of this relationship consist of number of genes annotated by the GO annotation goterm in gene2go dataset of NCBI (column gogenes), number of genes associated to articles in gene2pubmed dataset in NCBI (column genetot) annotated by the dterm, number of genes having both annotations dterm and goterm (column genenum), list of comma separated Entrez identifiers of genes that make genenum (genes sharing both dterm and goterm annotations). Column id is a unique row identifier. Column dtid is a key linking to the table mesh_terms. This table has 14,225,540 rows and 9 columns that are separated by tabs. Size of a plain table is 1.31GB. Compressed table takes 379MB. Information in this table is as of September 2013.

Dataset 3. Table go_terms

http://dx.doi.org/10.5256/f1000research.6140.d43176

Table go_terms stores description of gene ontology annotations that are in table dterm_go. Each row contains a gene ontology annotation as 10 character identifier in GO (column goterm), textual description of this term (column description) and its category (column category) which can have one of the possible values: 'Process','Function' or 'Component'. Column gokey is a unique row identifier. This table has 20,266 rows and 4 columns that are separated by tabs. Size of a plain table is 1.23MB. Compressed table takes 270KB.

Dataset mesh_terms

The mesh_terms table stores associations between MeSH terms defining phenotype and MeSH D terms defining chemicals. Statements to create this table in MySQL database are presented in Table 3. Each row stores a pair of MeSH term (category A,C,D and G used to define phenotypes) and a MeSH D term defining

Table 3. MySQL statements to create *mesh_terms* table.

create_mesh_terms_table.sql

```
DROP TABLE IF EXISTS mesh_terms;
CREATE TABLE mesh_terms
(
   mid int(10) unsigned NOT NULL,
   mterm varchar(250) NOT NULL,
   dterm varchar(250) NOT NULL,
   dscore float NOT NULL,
   inters int(10) unsigned NOT NULL,
   nm int(10) unsigned NOT NULL,
   nd int(10) unsigned NOT NULL,
   unio int(10) unsigned NOT NULL,
   pmids text NOT NULL,
   dtid int(11) NOT NULL
);
LOAD DATA LOCAL INFILE 'mesh_terms' INTO TABLE mesh_terms
COLUMNS TERMINATED BY '\t' IGNORE 1 LINES;
```

a chemical and their relationship as defined by Equation 1 and Equation 2 with supporting information. Data in this table are based on annotated PubMed content as of September 2013. Meaning of columns in **mesh_terms** table is as follows:

- *mid* is unique identifier of a row;
- *mterm* is MeSH term in which spaces are replaced by underscores (for example Cell_Fusion);
- *dterm* is MeSH D term in which spaces are replaced by underscores (for example BMP4_Protein);
- *dscore* is a float number representing a strength of connection between *mterm* and *dterm* computed as in Equation 2;
- *nm* number of PubMed articles annotated by *mterm*;
- *nd* number of PubMed articles annotated by *dterm* corresponding to $|D_{dterm}|$ in Equation 2;
- *inters* number of PubMed articles annotated by both *mterm* and *dterm* corresponding to $|G_{mterm} \cap D_{dterm}|$ in Equation 2;
- *unio* number of PubMed articles annotated by either *mterm* or *dterm* or both;
- *pmids* comma separated list of PMID identifiers of PubMed articles that are the *inters* articles;
- *dtid* numerical key identifying *dterm* in another table **dterm_go**.

Dataset dterm_go

The **dterm_go** table stores associations between MeSH D terms defining chemicals and gene ontology annotations of genes. This table was created by custom scripts from NCBI gene2go and gene2pubmed datasets. These datasets can be found on NCBI ftp site *ftp://ftp.ncbi.nlm.nih.gov/gene/DATA/*.

The gene2go dataset stores pairs of gene and its GO annotation. Annotations and genes of human, mouse and fly were retrieved. The gene2pubmed dataset stores pairs of genes and PMIDs of articles associated to them. From this datset pairs of genes and MeSH D annotations of their associated articles were retrieved. These intermediate datasets were used to create the **dterm_go** table. Statements to create this table in MySQL are presented in Table 4. Meaning of the columns in the **dterm_go** table is as follows:

- *dterm* is MeSH D term in which spaces are replaced by underscores (for example BMP4_Protein);
- *goterm* is identifier of GO annotation in the Gene Ontology (for example GO:0000001 is identifier for annotation “mitochondrion inheritance”);
- *gscore* is a float number representing a strength of connection between *goterm* and *dterm* computed according to Equation 3 and Equation 4;
- *gogenes* number of genes (of mouse, human and fly) annotated by *goterm* as was recorded in gene2go dataset in NCBI ftp repository corresponding to $|GO_{goterm}|$ in Equation 4;

Table 4. MySQL statements to create *dterm_go* table.

create_dterm_go_table.sql

```
DROP TABLE IF EXISTS dterm_go;
CREATE TABLE dterm_go
(
   dterm varchar(250) NOT NULL,
   goterm varchar(12) NOT NULL,
   gscore decimal(15,4) unsigned NOT NULL,
   gogenes int(10) unsigned NOT NULL,
   genenum int(10) unsigned NOT NULL,
   genetot int(10) unsigned NOT NULL,
   genes text NOT NULL,
   id int(11) NOT NULL
   dtid int(11) NOT NULL
);
LOAD DATA LOCAL INFILE ‘dterm_go’ INTO TABLE dterm_go
COLUMNS TERMINATED BY ‘\t’ IGNORE 1 LINES;
CREATE INDEX dtid_gscore_ind ON dterm_go(dtid,gscore);
```

- *genenum* number of genes (of mouse, human and fly) sharing the *goterm* and *dterm* annotations corresponding to $|GD_{dterm} \cap GO_{goterm}|$ in Equation 4;
- *genetot* number of genes (of mouse, human and fly) associated with articles annotated by *dterm* recorder in gene2pubmed dataset in NCBI ftp repository;
- *genes* comma separated list of Entrez Gene identifiers of genes that form *genenum* genes;
- *id* unique identifier of the row;
- *dtid* numerical key identifying *dterm* in table **mesh_terms**.

Dataset go_terms

Table **go_terms** stores description information of gene ontology annotations. Statements to create this table in MySQL are presented in Table 5. Meaning of the columns in the **dterm_go** table is as follows:

- *gokey* is a unique row identifier;
- *goterm* is an identifier of GO annotation in the Gene Ontology (for example GO:0000001 is identifier for annotation “mitochondrion inheritance”);

Table 5. MySQL statements to create *go_terms* table.

create_go_terms_table.sql

```
DROP TABLE IF EXISTS go_terms;
CREATE TABLE go_terms
(
   gokey int(10) unsigned NOT NULL,
   goterm varchar(12) NOT NULL,
   description text NOT NULL,
   category varchar(15) NOT NULL
);
LOAD DATA LOCAL INFILE ‘go_terms’ INTO TABLE go_terms
COLUMNS TERMINATED BY ‘\t’ IGNORE 1 LINES;
CREATE INDEX goterm_ind ON go_terms(goterm(10));
```

- *description* is a description of GO annotation in the Gene Ontology in which spaces are replaced by underscores (for example mitochondrion_inheritance is description of GO:0000001 identifier);
- *category* is an indicator of a category of GO annotation and can take value of "Process", "Function" or "Component".

Procedure to compute links between phenotype and go annotations

Figure 2 and Equation 1, Equation 3, Equation 5 outline a possible way of establishing links between phenotypes defined by MeSH terms and GO annotations that pass through chemicals. These links can be computed by MySQL statements given that MySQL database tables were created as shown in Table 3, Table 4, Table 5. The suggested procedure is presented in Table 6. Three input parameters *queryterm, dfrac, gofrac* can be provided to the procedure ***mesh_to_go.sql***. This MySQL procedure has a computation and an output part. The parameter *gueryterm* provides a MeSH term that defines phenotype. 16914 MeSH terms from 2012 MeSH edition[72] can be queried in current implementation. These terms form pairs with 5908 MeSH D terms of chemicals. Textual fields of all MeSH terms and GO annotations have underscores instead of spaces between words that should be used in formulating queries.

Parameters *dfrac* and *gofrac* set thresholds on the corresponding *dscore* and *gscore* values. They can be used to filter terms in computation based on strengths of relationships between the phenotype and chemical and between the chemical and GO annotation (disallowing weaker relationships). Value of *dfrac* can vary in range of [0.0000041, 1] which is a range of *dscore* values. Value of *gofrac* can vary in interval of [0.0001, 1].

In computation presented in Table 6, a creation of t2 and t3 tables corresponds to performing a maximum composition operation defined by Equation 5. The table t2 contains all relationships between the phenotype in *queryterm* and GO annotations passing through all chemicals that have a connection to the *queryterm* phenotype. Fewer GO annotations with only maximum weight in relationship to the phenotype are selected into table t3. Statements in the output part create plain sectioned text file. Weighted GO annotations are listed in the first section. Second section identified by "list_of_all_links_go_dterms", lists all connections in the table t2.

Table 6. MySQL procedure to compute links between phenotype and GO annotations.

mesh_to_go.sql

DROP PROCEDURE IF EXISTS **mesh_to_go**;
delimiter //
CREATE PROCEDURE **mesh_to_go**(in queryterm varchar(255), in dfrac float, in gofrac float)
 	
/*Increase memory limits if available for temporary tables*/
SET @@max_heap_table_size=1024*1024*1024*4;
SET @@tmp_table_size=1024*1024*1024*4;
 	
/* COMPUTATION */
DROP TABLE IF EXISTS t1,t2,t3;
CREATE TEMPORARY TABLE t1(INDEX dt_id (dtid)) ENGINE=MEMORY SELECT dtid,dscore,dterm, FROM **mesh_terms** WHERE mterm=queryterm and dscore>=dfrac;
 	
CREATE TEMPORARY TABLE t2 ENGINE=MEMORY SELECT TRUNCATE(a.gscore*b.dscore,9) AS ms, a.gscore, b.dscore, b.dterm, a.goterm FROM **dterm_go** AS a, t1 AS b WHERE a.dtid=b.dtid AND a.gscore>=gofrac;
 	
CREATE TEMPORARY TABLE t3 ENGINE=MEMORY SELECT MAX(ms) AS ms, goterm FROM t2 GROUP BY goterm ORDER BY ms DESC;
 	
/* OUTPUT*/
SELECT queryterm AS phenotype;
SELECT COUNT(*) AS list_of_max_go_terms FROM t3;
SELECT a.ms,a.goterm,b.description FROM t3 AS a, **go_terms** AS b WHERE a.goterm=b.goterm;
SELECT list_of_all_links_go_dterms;
SELECT a.ms, a.goterm, b.description, a.gscore, a.dterm, a.dscore
FROM t2 AS a, **go_terms** AS b WHERE a.goterm=b.goterm ORDER
BY ms DESC;
end //
delimiter ;

For example, a command line query for phenotype "Intellectual Disability" (user X and database Xdb) can be executed in a following way:

```
$mysql -u X -p  Xdb -e  "call             \\
   mesh_to_go( 'Intellectual_Disability',\\
   0.01,0.1);"  > id_out;
```

The first output file id_out section will have rows:

```
ms       goterm  description

0.071428596     GO:0014805   \\
     smooth_muscle_adaptation

0.071428596     GO:0045362   \\
    positive_regulation_of_ \\
    interleukin-1_biosynthetic_process

    ...
0.060150021     GO:0004908   \\
    interleukin-1_receptor_activity
    ...
```

These GO annotations are weighted by strength of their relationship to the "Intellectual Disability" phenotype. These weighted GO annotations can be used to rank genes as in Figure 2. Second section of output details relationships between the phenotype and chemicals and between the chemicals and GO annotations, for example considering information on GO:0045362:

```
ms              0.071428596
goterm          GO:0045362
description     positive_regulation_of_ \\
    interleukin-1_biosynthetic_process
gscore          1.0000
dterm           Interleukin-1_Receptor_ \\
    Accessory_Protein
dscore          0.0714286
```

A connection between "Intellectual Disability" phenotype and MeSH D term "Interleukin-1 Receptor" is quantified by *dscore* which equals to 0.0714286. Strength of connection between this chemical and "positive regulation of interleukin-1 biosynthetic process" GO term equals to 1. These two values determine the weight *ms* of this GO term in connection to "Intellectual Disability" (ID) phenotype. This computational procedure with respect to ID phenotype was previously explored[73].

Use case

Exploratory analysis of caner related genes in sequencing studies

In cancer genes accumulate a large number of mutations[74] and next generation sequencing screening may produce a vast number of genetic variants and genes. If a gene harboring a variant was not previously reported, then outlined computation can be used to explore connections of that gene to specific cancer based on a current available knowledge.

Among highly mutated genes identified by the whole genome and exome sequencing of breast tumors are PIK3CA, TP53, GATA3, CDH1, RB1, MLL3, MAP3K1 and CDKN1, which were previously observed in clinical breast cancer tumors[75]. Genes not previously observed in those tumors were TBX3, RUNX1, LDLRAP1, STNM2, MYH9, AGTR2, STMN2, SF3B1. Both sets of genes were explored in relation to "Breast Neoplasms" by ranking genes in whole human genome. The genes were ranked by magnitudes of weights of their annotations with respect to relationship to "Breast Neoplasms" as depicted in Figure 2 employing the outlined computational principle. Top genes from those sets appearing within the Top 5% of the ranked human genome and closer to this interval are listed in Table 7.

Cancer genes are well characterized and widely studied in literature. The genes, not previously reported as carrying clinically important mutations in studies of breast cancer in[75] had stronger links with cancer phenotype in question through their GO annotations. The genes from that study: RB1, GATA3, TP53 and CDH1 appeared in high ranking positions. Current exploration identifies "Tamoxifen" being strongly related to the breast cancer phenotype. Such link is logical because this chemical is used to treat hormone-sensitive tumors. Applied computational procedure through relationships between phenotypes and chemicals helps to explore contexts in which biological processes of interest take place. Unexpected links may be discovered that may help to formulate novel biological hypothesis.

Discussion

As of now, biologists still find it challenging to interpret large lists of poorly characterized genes with algorithms, which are somewhat limited in terms of how they define phenotypes. These lists may originate from a variety of sources, including microarray experiments, ChIP-Seq experiments identifying transcription factors' target genes, and scientific literature. The algorithm described here is useful in formulating biological hypotheses in situations in which little is known about the phenotype and the genes in question. The algorithm begins by linking lists of gene GO annotations to phenotypes (non-disease and disease) described by meaningful keywords through the MeSH and PubMed databases. The algorithm then deduces which of the links between the genes and phenotypes are strongest and presents the results in an organized manner. This is different from most of existing algorithms in terms of the methods used to define phenotypes of interest and infer their relationships with genes. To better understand how the outlined algorithm is unique, the existing algorithms are parsed into three categories overlapping at some extent and examined.

How algorithms define phenotypes and infer gene-phenotype relationships

The first category of existing algorithms only uses known phenotype-related genes, or training genes, while the second focuses solely on human disease phenotypes. The third category uses general keywords from literature to define phenotypes. All must deduce how genes and phenotypes are related by mining selected information sources, retrieving and integrating data from them, but there are some differences between them. Each will be discussed in turn.

Table 7. Genes appearing within Top 5% of the ranked whole genome with respect to "Breast Neoplasms".

Gene and description	Gene rank ratio (rank)	GO annotation and weight	MeSH D term and weight
RB1 retinoblastoma 1	0.0277 (5)	regulation of centromere complex assembly 0.647	BRCA2 Protein 0.647
GATA3 GATA binding protein 3	0.5370 (97)	type IV hypersensitivity 0.452	Receptors, Estrogen 0.452
TP53 tumor protein p53	0.7031 (127)	positive regulation of cell aging 0.417	BRCA2 Protein 0.647
RUNX1 runt-related transcription factor 1	0.9134 (165)	positive regulation of progesterone secretion 0.456	Receptors, Progesterone 0.456
CDH1 cadherin 1 type 1 E-cadherin (epithelial)	1.8324 (331)	regulation of water loss via skin 0.327	Tamoxifen 0.573
PIK3CA phosphatidylinositol-4 5-bisphosphate 3-kinase catalytic subunit alpha	2.2199 (401)	negative regulation of anoikis 0.307	Receptor, erbB-2 0.555
MAP3K1 mitogen-activated protein kinase kinase kinase 1 E3 ubiquitin protein ligase	3.1831 (575)	positive regulation of viral transcription 0.275	BRCA1 Protein 0.584
AGTR2 angiotensin II receptor type	4.4674 (807)	positive regulation of nitric-oxide synthase activity 0.232	Tamoxifen 0.573
TBX3 T-box	6.9475 (1255 outside Top 5%)	sinoatrial node cell development 0.191	Tamoxifen 0.573

Algorithms based on the use of training genes known to be related to the phenotype of interest, as in the Endeavour[39] and ToppGene[45] tools, make prioritizations on the basis of pattern classification[76]. Training genes extract phenotype-defining information from various data sources based on how similar the phenotype is to the training genes, and then build a model of a phenotype based on this extraction[42]. In other words, the model represents gene features that are most characteristic of that specific phenotype. The candidate genes are ranked by how similar their features are to the features of the model. For example, Endeavour relies on genomic data fusion from multiple information sources[41]. This tool may be very useful if the properties of the training genes clearly define the phenotype properties of interest in the organisms being investigated. Knowing these properties, one can characterize candidate genes by comparing them to the training genes. Genes that have similar characteristics to the training genes may also play an important role in previously unknown phenotype expressions. This principle of discovery is known as "guilt by association"[77]. Although very useful in detecting similarities between candidates and training genes, the integration of data from multiple sources has limitations. The main limitation is in existing schemes of combining the information from different sources to rank the candidates[41]. First, the prioritization algorithms using training genes generally differ with respect to the data sources they use[23]. Different information sources of training genes lead to different models and similarity metrics. Second, some data sources do not have complete information on some genes, so if the phenotype in question has not been sufficiently studied and there are no genes known to be associated with it, then the training genes approach is not effective. Third, the training genes might represent a heterogeneous group biasing phenotype definition in some way. For example, the data fusion scheme relies on the independence of information sources about gene properties, but in practice they are not entirely independent. Protein-protein interaction databases, the gene interaction databases and gene ontology refer to scientific publications as supporting evidence for the information they store, and might even be derived from the literature. While it should do so, the scheme does not always account for these possible interactions and overlaps between sources.

Many tools specific to human diseases use phenotype definitions from databases[47,59]. Because human disease phenotypes have been extensively studied and are well represented by OMIM[78], and because they contain structured information suited to uncovering meaningful links between human diseases and genes, it is relatively easy to associate genes with said phenotypes. However, phenotypes other than diseases and phenotypes in Mammalian Phenotype Ontology[37] are not yet represented by well-structured and information-rich resources[33,34].

The third category of algorithms, which use general keywords from literature to define phenotypes, are exemplified by tools such as PosMed, PolySearch, GeneProspector and CANDID[48,51–53]. These tools rely on finding matching documents in MEDLINE or locally-created databases, and then associating genes with the matching documents. General purpose discovery-oriented systems such as iHOP[79], Anni2.0[80], Arrowsmith[60,61] and PosMed[52], use conceptual networks. Users can browse through the network and create textual profiles describing genes, proteins, or other biomedical concepts. However, once again, there will be genes and processes that are not well represented in literature and there is little information about them that can be retrieved.

Thus, while the obvious advantage offered by specialized gene information databases is that specific information can be extracted very quickly, complementing the literature with more information sources for gene prioritization is advantageous in allowing potential use of algorithms that offer novel interpretations of existing information. G2D[43,59] is the only existing method which provides underlying ideas for the algorithmic approach outlined here, which prioritizes genes with respect to human disease phenotypes. However, the scope of the application of the developed algorithm to link genes with phenotypes[70] outlined here is distinct from G2D by contributing the following:

- The outlined algorithm establishes meaningful links between genes and phenotypes, and enables prioritization, beyond human disease phenotypes by using concepts of MeSH vocabulary from the categories A, D, and G.
- The proposed algorithm can be applied beyond human organisms as the annotated genes of the entire genome for human (Homo sapiens), mouse (Mus musculus) and fly (Drosphila melanogaster) are used. In contrast, G2D focuses on human genes.
- The data in the *gene2go* and *gene2pubmed* NCBI databases[65] are used to link GO annotations to MeSH terms of Drugs and Chemicals describing the molecular entities. In contrast, G2D works with the RefSeq database for this purpose[43].
- The outlined algorithm similarly to G2D utilizes fuzzy binary relationships between concepts, based on mathematical operations of fuzzy set theory[67,68], to infer gene-phenotype links. G2D uses a similarity relationship[43] while this algorithm uses an inclusion relationship[70].

The outlined algorithm is an attempt to remedy some of the challenges presented by information shortages and the way existing algorithms described above are configured to define phenotypes and determine relationships. It has important advantages compared to the other gene prioritization algorithms, in addition to G2D, reviewed extensively in Introduction section.

Conclusions

The approach to link genes and phenotypes outlined in this work represents one out of existing possible approaches. Contributed datasets opens possibility to experimentation and development of other applications. These datasets, although in need of updating comprise co-occurrences of selected categories of MeSH terms in PubMed and co-occurrences of MeSH D terms with GO annotations created from NCBI Gene datasets. Availability of such offline data saves time of a researcher who may want to explore and apply text and data mining algorithms to analyze relationships between concepts.

Existing tools provide limited explanations for reasons for phenotype gene association. Using the outlined approach, evidence supporting the obtained strongest links can be easily examined. As a result of inference, the MeSH D terms which are most strongly related to both the candidate genes through their GO annotations and phenotype are identified. This is useful as it reveals the physical background domains related to the candidate genes gleaned from associated articles without reading their full text. The availability of this background information opens up the possibility of identifying and examining unique aspects of the functions of the studied genes.

However, a single information source cannot account for all aspects of gene relations to phenotypes even if MeSH vocabulary contains information about the processes, phenomena and phenotypes studied in literature. And while functional gene annotations are also associated with scientific publications, there will be genes and processes that are not well represented in literature, as stated earlier. In this situation inferring links between genes and phenotypes might be more effective using other information sources, as genes can also be characterized by their interactions with other molecular entities, by their sequences and by the information about the protein domains of the products. These gene properties can be retrieved computationally from other specialized databases.

Data availability

F1000Research: Dataset 1. Table mesh_terms, 10.5256/f1000research.6140.d43167[82]

F1000Research: Dataset 2. Table dterm_go, 10.5256/f1000research.6140.d43168[83]

F1000Research: Dataset 3. Table go_terms, 10.5256/f1000research.6140.d43176[84]

Competing interests

Author declares no competing interests.

Grant information

The author(s) declared that no grants were involved in supporting this work.

Acknowledgements

Part of this work was performed during completion of graduate training at department of Cellular and Molecular Medicine (CMM), University of Ottawa. Author gratefully acknowledges University of Ottawa for their admission scholarship and thanks Dr. C. Perez-Iratxeta from CMM and affiliated Ottawa Hospital Research Institute for her advise on G2D and related discussions. Author thanks Ms. Lindsey Li from Carleton University, Ottawa for her help in improving English expressions.

References

1. Shendure J, Aiden EL: **The expanding scope of DNA sequencing.** *Nat Biotechnol.* 2012; **30**(11): 1084–1094.
PubMed Abstract | Publisher Full Text | Free Full Text

2. Dudley JT, Karczewski KJ: **Exploring personal genomics**. Oxford University Press, 2013.
Publisher Full Text

3. Fernald GH, Capriotti E, Daneshjou R, *et al.*: **Bioinformatics challenges for personalized medicine.** *Bioinformatics.* 2011; **27**(13): 1741–1748.
PubMed Abstract | Publisher Full Text | Free Full Text

4. Lee BR, Cho S, Song Y, *et al.*: **Emerging tools for synthetic genome design.** *Mol Cells.* 2013; **35**(5): 359–370.
PubMed Abstract | Publisher Full Text | Free Full Text

5. Esvelt KM, Wang HH: **Genome-scale engineering for systems and synthetic biology.** *Mol Syst Biol.* 2013; **9**: 641.
PubMed Abstract | Publisher Full Text | Free Full Text

6. de la Iglesia D, Garcia-Remesal DM, de la Calle G, *et al.*: **The impact of computer science in molecular medicine: enabling high-throughput.** *Curr Top Med Chem.* 2013; **13**(5): 526–75.
PubMed Abstract | Publisher Full Text

7. Hawkins RD, Hon GC, Ren B: **Next-generation genomics: an integrative approach.** *Nat Rev Genet.* 2010; **11**(7): 476–486.
PubMed Abstract | Publisher Full Text | Free Full Text

8. Carpenter AE, Sabatini DM: **Systematic genome-wide screens of gene function.** *Nat Rev Genet.* 2004; **5**(1): 11–22.
PubMed Abstract | Publisher Full Text

9. Dunham I, Kundaje A, Aldred SF, *et al.*: **An integrated encyclopedia of DNA elements in the human genome.** *Nature.* 2012; **489**(7414): 57–74.
PubMed Abstract | Publisher Full Text | Free Full Text

10. Aerts S, Vilain S, Hu S, *et al.*: **Integrating computational biology and forward genetics in Drosophila.** *PLoS Genet.* 2009; **5**(1): e1000351.
PubMed Abstract | Publisher Full Text | Free Full Text

11. Fernandez-Suarez XM, Galperin MY: **The 2013 Nucleic Acids Research Database Issue and the online molecular biology database collection.** *Nucleic Acids Res.* 2013; **41**(Database issue): D1–D7.
Publisher Full Text

12. Manconi A, Vargiu E, Armano G, *et al.*: **Literature retrieval and mining in bioinformatics: state of the art and challenges.** *Adv Bioinformatics.* 2012; **2012**: 573846.
PubMed Abstract | Publisher Full Text | Free Full Text

13. Kersey P, Apweiler R: **Linking publication, gene and protein data.** *Nat Cell Biol.* 2006; **8**(11): 1183–1189.
PubMed Abstract | Publisher Full Text

14. Turenne N, Tiys E, Ivanisenko V, *et al.*: **Finding biomarkers in non-model species: literature mining of transcription factors involved in bovine embryo development.** *BioData Min.* 2012; **5**(1): 12.
PubMed Abstract | Publisher Full Text | Free Full Text

15. Andronis C, Sharma A, Virvilis V, *et al.*: **Literature mining, ontologies and information visualization for drug repurposing.** *Brief Bioinformatics.* 2011; **12**(4): 357–368.
PubMed Abstract | Publisher Full Text

16. Zhu Q, Lajiness MS, Ding Y, *et al.*: **WENDI: A tool for finding non-obvious relationships between compounds and biological properties, genes, diseases and scholarly publications.** *J Cheminform.* 2010; **2**: 6.
PubMed Abstract | Publisher Full Text | Free Full Text

17. Rebholz-Schuhmann D, Arregui M, Gaudan S, *et al.*: **Text processing through Web services: calling Whatizit.** *Bioinformatics.* 2008; **24**(2): 296–298.
PubMed Abstract | Publisher Full Text

18. Krallinger M, Leitner F, Valencia A: **Analysis of biological processes and diseases using text mining approaches.** *Methods Mol Biol.* 2010; **593**: 341–382.
PubMed Abstract | Publisher Full Text

19. Brazas MD, Yim D, Yeung W, *et al.*: **A decade of Web Server updates at the Bioinformatics Links Directory: 2003–2012.** *Nucleic Acids Res.* 2012; **40**(Web Server issue): W3–W12.
PubMed Abstract | Publisher Full Text | Free Full Text

20. Masoudi-Nejad A, Meshkin A, Haji-Eghrari B, *et al.*: **Candidate gene prioritization.** *Mol Genet Genomics.* 2012; **287**(9): 679–698.
PubMed Abstract | Publisher Full Text

21. Piro RM, Di Cunto F: **Computational approaches to disease-gene prediction: rationale, classification and successes.** *FEBS J.* 2012; **279**(5): 678–696.
PubMed Abstract | Publisher Full Text

22. Capriotti E, Nehrt NL, Kann MG, *et al.*: **Bioinformatics for personal genome interpretation.** *Brief Bioinform.* 2012; **13**(4): 495–512.
PubMed Abstract | Publisher Full Text | Free Full Text

23. Tranchevent LC, Capdevila FB, Nitsch D, *et al.*: **A guide to web tools to prioritize candidate genes.** *Brief Bioinform.* 2011; **12**(1): 22–32.
PubMed Abstract | Publisher Full Text

24. Mahner M, Kary M: **What exactly are genomes, genotypes and phenotypes? And what about phenomes?** *J Theor Biol.* 1997; **186**(1): 55–63.
PubMed Abstract | Publisher Full Text

25. Marian AJ: **Challenges in medical applications of whole exome/genome sequencing discoveries.** *Trends Cardiovasc Med.* 2012; **22**(8): 219–223.
PubMed Abstract | Publisher Full Text | Free Full Text

26. Kohler S, Doelken SC, Rath A, *et al.*: **Ontological phenotype standards for neurogenetics.** *Hum Mutat.* 2012; **33**(9): 1333–1339.
PubMed Abstract | Publisher Full Text

27. Fuchs F, Pau G, Kranz D, *et al.*: **Clus-tering phenotype populations by genome-wide RNAi and multiparametric imaging.** *Mol Syst Biol.* 2010; **6**: 370.
PubMed Abstract | Publisher Full Text | Free Full Text

28. Hoehndorf R, Dumontier M, Gkoutos GV: **Evaluation of research in biomedical ontologies.** *Brief Bioinform.* 2012; **14**(6): 696–712.
PubMed Abstract | Publisher Full Text | Free Full Text

29. Hoehndorf R, Harris MA, Herre H, *et al.*: **Semantic integration of physiology phenotypes with an application to the Cellular Phenotype Ontology.** *Bioinformatics.* 2012; **28**(13): 1783–1789.
PubMed Abstract | Publisher Full Text | Free Full Text

30. Gkoutos GV, Green EC, Mallon AM, *et al.*: **Using ontologies to describe mouse phenotypes.** *Genome Biol.* 2005; **6**(1): R8.
PubMed Abstract | Publisher Full Text | Free Full Text

31. Flybase: **Links to the model organism projects at the flybase web portal**. 2013.

32. Smith B, Ashburner M, Rosse C, *et al.*: **The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration.** *Nat Biotechnol.* 2007; **25**(11): 1251–1255.
PubMed Abstract | Publisher Full Text | Free Full Text

33. Groth P, Kalev I, Kirov I, *et al.*: **Phenoclustering: online mining of cross-species phenotypes.** *Bioinformatics.* 2010; **26**(15): 1924–1925.
PubMed Abstract | Publisher Full Text | Free Full Text

34. Houle D, Govindaraju DR, Omholt S: **Phenomics: the next challenge.** *Nat Rev Genet.* 2010; **11**(12): 855–866.
PubMed Abstract | Publisher Full Text

35. Webb AJ, Thorisson GA, Brookes AJ: **An informatics project and online "Knowledge Centre" supporting modern genotype-to-phenotype research.** *Hum Mutat.* 2011; **32**(5): 543–550.
PubMed Abstract | Publisher Full Text

36. Butte AJ, Kohane IS: **Creation and implications of a phenome-genome network.** *Nat Biotechnol.* 2006; **24**(1): 55–62.
PubMed Abstract | Publisher Full Text | Free Full Text

37. Köhler S, Schulz MH, Krawitz P, *et al.*: **Clinical diagnostics in human genetics with semantic similarity searches in ontologies.** *Am J Hum Genet.* 2009; **85**(4): 457–464.
PubMed Abstract | Publisher Full Text | Free Full Text

38. Schofield PN, Sundberg JP, Hoehndorf R, *et al.*: **New approaches to the representation and analysis of phenotype knowledge in human diseases and their animal models.** *Brief Funct Genomics.* 2011; **10**(5): 258–265.
PubMed Abstract | Publisher Full Text | Free Full Text

39. Tranchevent LC, Barriot R, Yu S, *et al.*: **ENDEAVOUR update: a web resource for gene prioritization in multiple species.** *Nucleic Acids Res.* 2008; **36**(Web Server issue): W377–384.
PubMed Abstract | Publisher Full Text | Free Full Text

40. Bult CJ, Eppig JT, Blake JA, *et al.*: **The mouse genome database: genotypes, phenotypes, and models of human disease.** *Nucleic Acids Res.* 2013; **41**(Database issue): D885–891.
PubMed Abstract | Publisher Full Text | Free Full Text

41. Aerts S, Lambrechts D, Maity S, *et al.*: **Gene prioritization through genomic data fusion.** *Nat Biotechnol.* 2006; **24**(5): 537–544.
PubMed Abstract | Publisher Full Text

42. Tranchevent L: **Gene prioritization through genomic data fusion**. PhD thesis, Faculty of Engineering, K.U.Leuven (Leuven, Belgium), 2011.

43. Perez-Iratxeta C, Bork P, Andrade MA, *et al.*: **Association of genes to genetically inherited diseases using data mining.** *Nat Genet.* 2002; **31**(3): 316–319.
PubMed Abstract | Publisher Full Text

44. Chen J, Xu H, Aronow BJ, *et al.*: **Improved human disease candidate gene prioritization using mouse phenotype.** *BMC Bioinformatics.* 2007; **8**: 392.
PubMed Abstract | Publisher Full Text | Free Full Text

45. Chen J, Bardes EE, Aronow BJ, *et al.*: **ToppGene Suite for gene list enrichment analysis and candidate gene prioritization.** *Nucleic Acids Res.* 2009; **37**(Web Server issue): W305–311.
PubMed Abstract | Publisher Full Text | Free Full Text

46. Köhler S, Bauer S, Horn D, *et al.*: **Walking the interactome for prioritization of candidate disease genes.** *Am J Hu Genet.* 2008; **82**(4): 949–958.
PubMed Abstract | Publisher Full Text | Free Full Text

47. van Driel MA, Bruggeman J, Vriend G, *et al.*: **A text-mining analysis of the human phenome.** *Eur J Hum Genet.* 2006; **14**(5): 535–542.
PubMed Abstract | Publisher Full Text

48. Cheng D, Knox C, Young N, *et al.*: **PolySearch: a web-based text mining system for extracting relationships between human diseases, genes, mutations, drugs and metabolites.** *Nucleic Acids Res.* 2008; **36**(Web Server issue): W399–405.
PubMed Abstract | Publisher Full Text | Free Full Text

49. Adie EA, Adams RR, Evans KL, *et al.*: **SUSPECTS: enabling fast and effective prioritization of positional candidates.** *Bioinformatics.* 2006; **22**(6): 773–774.
PubMed Abstract | Publisher Full Text

50. Radivojac P, Peng K, Clark WT, *et al.*: **An integrated approach to inferring gene-disease associations in humans.** *Proteins.* 2008; **72**(3): 1030–1037.
PubMed Abstract | Publisher Full Text | Free Full Text

51. Hutz JE, Kraja AT, McLeod HL, *et al.*: **CANDID: a flexible method for prioritizing candidate genes for complex human traits.** *Genet Epidemiol.* 2008; **32**(8): 779–790.
PubMed Abstract | Publisher Full Text

52. Yoshida Y, Makita Y, Heida N, *et al.*: **PosMed (Positional Medline): prioritizing genes with an artificial neural network comprising medical documents to accelerate positional cloning.** *Nucleic Acids Res.* 2009; **37**(Web Server issue): W147–152.
PubMed Abstract | Publisher Full Text | Free Full Text

53. Yu W, Wulf A, Liu T, *et al.*: **Gene Prospector: an evidence gateway for evaluating potential susceptibility genes and interacting risk factors for human diseases.** *BMC Bioinformatics.* 2008; **9**: 528.
PubMed Abstract | Publisher Full Text | Free Full Text

54. Popescu M, Keller JM, Mitchell JA, *et al.*: **Fuzzy measures on the Gene Ontology for gene product similarity.** *IEEE/ACM Trans Comput Biol Bioinform.* 2006; **3**(3): 263–274.
PubMed Abstract | Publisher Full Text

55. Nikopensius T, Ambrozaityte L, Ludwig KU, *et al.*: **Replication of novel susceptibility locus for nonsyndromic cleft lip with or without cleft palate on chromosome 8q24 in Estonian and Lithuanian patients.** *Am J Med Genet A.* 2009; **149A**(11): 2551–2553.
PubMed Abstract | Publisher Full Text

56. Qi CF, Martensson A, Mattioli M, *et al.*: **CTCF functions as a critical regulator of cell-cycle arrest and death after ligation of the B cell receptor on immature B cells.** *Proc Natl Acad Sci U S A.* 2003; **100**(2): 633–638.
PubMed Abstract | Publisher Full Text | Free Full Text

57. Schaub MA, Boyle AP, Kundaje A, *et al.*: **Linking disease associations with regulatory information in the human genome.** *Genome Res.* 2012; **22**(9): 1748–1759.
PubMed Abstract | Publisher Full Text | Free Full Text

58. Suzuki S, Marazita ML, Cooper ME, *et al.*: **Mutations in *BMP4* are associated with subepithelial, microform, and overt cleft lip.** *Am J Hum Genet.* 2009; **84**(3): 406–411.
PubMed Abstract | Publisher Full Text | Free Full Text

59. Perez-Iratxeta C, Wjst M, Bork P, *et al.*: **G2D: a tool for mining genes associated with disease.** *BMC Genet.* 2005; **6**: 45.
PubMed Abstract | Publisher Full Text | Free Full Text

60. Smalheiser NR, Swanson DR: **Using ARROWSMITH: a computer-assisted approach to formulating and assessing scientific hypotheses.** *Comput Methods Programs Biomed.* 1998; **57**(3): 149–153.
PubMed Abstract | Publisher Full Text

61. Smalheiser NR, Torvik VI, Zhou W: **Arrowsmith two-node search interface: a tutorial on finding meaningful links between two disparate sets of articles in MEDLINE.** *Comput Methods Programs Biomed.* 2009; **94**(2): 190–197.
PubMed Abstract | Publisher Full Text | Free Full Text

62. Swanson DR: **Fish oil Raynaud's syndrome, and undiscovered public knowledge.** *Perspect Biol Med.* 1986; **30**(1): 7–18.
PubMed Abstract

63. Shatkay H, Craven M: **Mining the Biomedical Literature**. MIT Press. 2012.
Reference Source

64. Hristovski D, Peterlin B, Mitchell JA, *et al.*: **Using literature-based discovery to identify disease candidate genes.** *Int J Med Inform.* 2005; **74**(2–4): 289–298.
PubMed Abstract | Publisher Full Text

65. Maglott D, Ostell J, Pruitt KD, *et al.*: **Entrez Gene: gene-centered information at NCBI.** *Nucleic Acids Res.* 2011; **39**(Database issue): D52–57.
PubMed Abstract | Publisher Full Text | Free Full Text

66. Perez-Iratxeta C, Keer HS, Bork P, *et al.*: **Computing fuzzy associations for the analysis of biological literature.** *BioTechniques.* 2002; **32**(6): 1380–1382.
PubMed Abstract

67. Zimmermann HJ: **Fuzzy set theory.** *Wiley Interdisciplinary Reviews: Computational Statistics.* 2010; **2**(3): 317–332.
Publisher Full Text

68. Zimmermann HJ: **Fuzzy Set Theory and its applications**. Kluwer Academic Publishers. 1996.
Reference Source

69. Miyamoto S: **Information retrieval based on fuzzy associations.** *Fuzzy sets and systems.* 1990; 38.
Publisher Full Text

70. Pranckeviciene E: **Bioinformatics tools for the analysis of gene-phenotype relationships coupled with a next generation ChIP-sequencing data processing pipeline.** *PhD thesis, Faculty of Medicine, Ottawa University.* 2015.
Reference Source

71. Sayers E: **The e-utilities in-depth: Parameters, syntax and more**. 2009.
Reference Source

72. **NCBI Medical Subject Headings.** *Mesh browser.* 2012.

73. Pranckeviciene E, Pranculis A, Preiksaitiene E, *et al.*: **Computational pipeline to analyze genomic variants with respect to clinical phenotypes by mining literature. Study of genomic regions related to intellectual disability.** *European Journal of Human Genetics.* 2014; **22**(Supplement 1): P16.48–M,p314.
Reference Source

74. Roukos DH: **Integrated clinical genomics: new horizon for diagnostic and biomarker discoveries in cancer.** *Expert Rev Mol Diagn.* 2013; **13**(1): 1–4.
PubMed Abstract | Publisher Full Text

75. Ellis MJ, Ding L, Shen D, *et al.*: **Whole-genome analysis informs breast cancer response to aromatase inhibition.** *Nature.* 2012; **486**(7403): 353–360.
PubMed Abstract | Publisher Full Text | Free Full Text

76. De Bie T, Tranchevent LC, van Oeffelen LM, *et al.*: **Kernel-based data fusion for gene prioritization.** *Bioinformatics.* 2007; **23**(13): i125–132.
PubMed Abstract | Publisher Full Text

77. Wang PI, Marcotte EM: **It's the machine that matters: Predicting gene function and phenotype from protein networks.** *J Proteomics.* 2010; **73**(11): 2277–2289.
PubMed Abstract | Publisher Full Text | Free Full Text

78. Hamosh A, Scott AF, Amberger JS, *et al.*: **Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders.** *Nucleic Acids Res.* 2005; **33**(Database issue): D514–517.
PubMed Abstract | Publisher Full Text | Free Full Text

79. Fernandez JM, Hoffmann R, Valencia A, *et al.*: **iHOP web services.** *Nucleic Acids Res.* 2007; **35**(Web Server issue): W21–26.
PubMed Abstract | Publisher Full Text | Free Full Text

80. Jelier R, Schuemie MJ, Veldhoven A, *et al.*: **Anni 2.0: a multipurpose textmining tool for the life sciences.** *Genome Biol.* 2008; **9**(6): R96.
PubMed Abstract | Publisher Full Text | Free Full Text

81. Valentini G, Paccanaro A, Caniza H, *et al.*: **An extensive analysis of disease-gene associations using network integration and fast kernel-based gene prioritization methods.** *Artif Intell Med.* 2014; **61**(2): 63–78.
PubMed Abstract | Publisher Full Text | Free Full Text

82 Pranckeviciene E: **Dataset 1 in "Procedure and datasets to compute links between genes and phenotypes defined by MeSH keywods".** *F1000Research.* 2014.
Data Source

83. Pranckeviciene E: **Dataset 2 in "Procedure and datasets to compute links between genes and phenotypes defned by MeSH keywords".** *F1000Research.* 2014.
Data Source

84. Pranckeviciene E: **Dataset 3 in "Procedure and datasets to compute links between genes and phenotypes defned by MeSH keywords".** *F1000Research.* 2014.
Data Source

Open Peer Review

Current Peer Review Status: ? ?

Version 1

Reviewer Report 27 April 2015

https://doi.org/10.5256/f1000research.6579.r7738

? **Emidio Capriotti**
Department of Biological, Geological and Environmental Sciences, University of Bologna, Bologna, Italy

The article addresses one important issue in the field but misses the following important aspects that should be discussed:

1. Most of the available annotation database such as GO, MESH are biased toward specific terms. Thus, the accuracy of the method could biased toward specific concepts and terms. The author should analyze the distribution of the m scores and associate the statistical significance calculated with respect to a background distribution.

2. For the definition of phenotypes the Human Phenotype Ontology (HPO) database is the reference. The author should cite this database in their paper and use it as benchmark set for the predictions.

3) A comparison with other methods should be provided.

Minor:

The description of similar methods previously developed is too long.

Competing Interests: No competing interests were disclosed.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Reviewer Report 03 March 2015

https://doi.org/10.5256/f1000research.6579.r7736

?

Jason E. McDermott
Department of Computational Biology, Pacific Northwest National Laboratory, Richland, WA, 99352, USA

Major concerns:

1. The process of ranking individual genes by their relationships as depicted in Figure 2 and in the Use Case (page 9) was very confusing and it was not apparent how the process of associating functions with phenotype could be mapped to genes of interest to the researcher. Greater care needs to be taken to describe this process (which is a somewhat complicated one) since this seems to be the main application of the method described.

2. The last five sentences in the Abstract provide detail that is not necessary here but should be included in the main text instead.

3. It would be very helpful to have a comparison of the results given by the method described and another method (likely one of those described in the paper) in terms of functions identified for the use case and ranking of genes given. This would not be a performance evaluation (since it is difficult to tell what the 'right' answer would be in this case) but would provide a nice comparison with previous methods.

Minor concerns:

1. The word "cancer" is misspelled in the title for the Use Case section.

2. In the use case it is unclear where the list of "Genes not previously observed in these tumors" comes from. I would assume that it was genes that were associated with cancer in the referenced work, but hadn't been previously associated? This needs to be made clear.

Competing Interests: No competing interests were disclosed.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.