Prioritisation of candidate genes for psychiatric disorders: An *in silico* approach

Kerry Kalweit
Supervisor: Professor Nicola Mulder
Co-supervisor: Reinette Weidemann
Programme: Bioinformatics

Presented for BSc(Med)(Hons)
Faculty of Health Sciences
University of Cape Town

25 October 2013

Total Word Count: 8171
Abstract Word Count: 238

The financial assistance of the National Research Foundation (NRF) towards this research is hereby acknowledged. Opinions expressed and conclusions arrived at are those of the author and are not necessarily to be attributed to the NRF.

Table of Contents

Anti-Plagiarism Declaration

1. I know that plagiarism is wrong. Plagiarism is to use another's work and pretend that it is one's own.
2. I have used the American Psychological Association, 6th Edition convention for citation and referencing. Each contribution to, and quotation in, this literature review from the work(s) of other people has been attributed, and has been cited and referenced.
3. This literature review is my own work.
4. I have not allowed, and will not allow, anyone to copy my work with the intention of passing it off as his or her own work.
5. I acknowledge that copying someone else's assignment or essay, or part of it, is wrong, and declare that this is my own work.

_____

Kerry Leigh Kalweit
Presented for BSc(Med)(Hons) Bioinformatics
University of Cape Town
25 October 2013

List of abbreviations

API	Application Program Interface
AUC	Area under the curve
CD/CV	Common disease/common variant hypothesis
CD/RV	Common disease/rare variant hypothesis
DIG	Data Integrated Genetics
DO	Disease Ontology
DSM	Diagnostic and statistical manual of mental disorders
FDR	False discovery rate
ftp	File transfer protocol
GAD	Genetics Association Database
GRASE	General and Rapid Association Study Engine
GO	Gene ontology
GWAS	Genome-wide association study
html	Hypertext markup language
HTT	Huntingtin gene
ICD	International
ID(s)	Identifier(s)
JSON	JavaScript Object Notation
MeSH	Medical subject headings
NCBI	National Centre for Biotechnology Information
NCI	National Cancer Institute
NGS	Next-generation sequencing
OMIM	Online Mendelian Inheritance of Man
PCR	Polymerase chain reaction
PPI	Protein–protein interactions
ROC	Receiver-operating characteristic
rs	reference SNP
SNP	Single nucleotide polymorphism
TPR	True positive rate
UTR	Untranslated region
VCF	Variant call format
xml	Extensible markup language

List of figures and tables

Abstract

The application of genome-wide association studies and next-generation sequencing has had limited success in identifying causal genes for complex diseases. Bipolar disorder is one such disease whose aetiology has not been elucidated despite the application of these technologies. Candidate gene prioritisation offers a solution to limit the vast amount of possible candidate genes produced from the combination of data sources. Current prioritisation tools rely heavily on previous data and thus do not perform well for poorly characterised diseases such as bipolar disorder. Here we have developed Data Integrated Genetics, DIG, a new candidate gene prioritisation tool designed specifically for complex genetic diseases. Given a user-specified disease query, DIG initially data-mines literature, linkage, homolog and sequence data to create a pool of possible candidates. The tool filters out likely false positives by removing pseudogenes. A unique data integration method is used to rank the remaining list of genes. Additionally, ranking is validated by tissue expression and single nucleotide polymorphism annotation. DIG exhibited comparable performance to existing tools when evaluated with four complex diseases. Eight novel genes were identified when DIG was applied to bipolar disorder, of which the Huntingtin gene poses as an exciting avenue for new aetiology research. The ease of use and realistic number of possible candidates given in the DIG results make this tool highly useful for research application in the study of complex genetic diseases. DIG is freely available from http://www.cbio.uct.ac.za/DIG.

1 Introduction

Identifying causal genetic factors is an essential step in our understanding and subsequent prevention and treatment of complex disorders (Sun *et al*., 2009). The advent of genome-wide association studies (GWAS) and next generation sequencing (NGS) has offered us the ability to explore genomic variations in affected individuals. However, there are several factors that complicate the ability of these technologies to confirm if variants are false positives or genuinely causal genes (Oti *et al*., 2011). The vast amount of data produced by GWAS and NGS is one such factor. Intricate computational analysis serves as a solution to reduce the number of likely candidates for further investigation into the true differences between case and control patients.

1.1 What is a complex disorder?

Complex disorders are caused by a combination of genetic and environmental factors. These can occur as discrete traits, such as diabetes, as well as continuous traits, such as the psychiatric disorders bipolar disorder and schizophrenia. Complex disorders do not obey the standard Mendelian patterns of inheritance as they are polygenic, whereby a combination of variants in many genes contribute to the manifestation of disease (Pritchard, 2001).

There are currently two central hypotheses to explain the genetic contribution needed for a certain complex phenotype to manifest. The first, the "common disease/common variant" (CD/CV) hypothesis postulates that a disease phenotype is likely to result from the aggregate effect of variants present at high frequencies in human populations (Lander, 1996). Recently, the alternative "common disease/rare variant" (CD/RV) hypothesis was proposed, as studies on common variants have failed to explain a large portion of the heritability of complex disorders (Schork *et al*., 2009), deemed as the "missing heritability". The CD/RV hypothesis states that multiple rare variants with large effect sizes are the main determinants of heritability of the complex phenotypes (Bodmer and Bonilla, 2008). The missing heritability could, however, also be explained by common variants with small effect sizes, such that GWAS are underpowered to detect the slightly significant differences between case and control samples.

Minor individual contributions from a large set of genes, as well as the reduced probability of finding rare susceptibility genes, creates a challenge to determine true causal variants. Methods to

prioritise candidate genes, thereby limiting the number of investigated genes, are the first step for investigating the aetiology and genetic cause of complex diseases.

1.2 Bipolar disorder as a case study

By the year 2020, it is estimated that psychiatric disorders will account for 15% of the total burden of all diseases (Merikangas *et al*., 2009). Bipolar disorder (BPD) is one of the major contributors of this statistic. Unfortunately, very little knowledge is known about the aetiology of this disorder. BPD presents a niche in which to apply new techniques that aim to uncover genetic determinants of complex disorders.

According to the *Diagnostic and statistical manual of mental disorders* (DSM, 4th edition, 1994), BPD is diagnosed with the presence of one or more manic or mixed episodes, usually accompanied with one or more major depressive episodes. The lifetime prevalence of BPD varies from 0% to 2.1% among different populations (Merikangas *et al*., 2009). There is a strong genetic component to the susceptibility to BPD, with heritability estimated at 87% (Smoller and Finn, 2003). Concordance rate for monozygotic twins is around 70% (Cardno *et al*., 1999) and relative risk for siblings of affected individuals is estimated at 7.9 times that of the general population (Lichtenstein *et al*., 2009).

Large numbers of genetic studies, including several GWAS, have been carried out to identify BPD susceptibility loci (Goes *et al*., 2012; Greenwood *et al*., 2013; Meier *et al*., 2012). Functional neuroimaging studies in BPD have identified dysfunction in key neural circuits, including the amygdale, limbic nuclei, prefrontal cortex, the anterior cingulate, the medial thalamus, and regions of the basal ganglia (Carlson *et al*., 2006).

Despite the abundance of research, there has been no convincing insight into understanding the aetiology of BPD. New methods in finding susceptibility genes are required to propose novel genes to investigate for involvement in pathogenesis of this disorder. BPD is therefore an ideal disease model to test the functionality and application of a newly designed candidate gene prioritisation approach. If successful, the identification of novel candidate genes could aid geneticists in the elucidation of the biological mechanism responsible for the BPD phenotype.

1.3 Candidate gene prioritisation

Candidate gene prioritisation is the identification of putative genes that show the most promising role in disease aetiology from a large list of genes (Tranchevent *et al*., 2010). This approach greatly reduces the number of genes to consider in the study of a complex disorder and ranks them according to the likelihood of being involved in disease pathogenesis (Sun *et al*., 2009). Most current prioritisation strategies use the 'guilt-by-association' concept wherein novel candidate genes will be those that are similar to genes already known to be associated with the given disease phenotype (Tranchevent *et al*., 2010).

Bioinformatic approaches, which allow for computational analysis of relevant information from a variety of different data sources, are critical in effective candidate gene prioritisation. Numerous bioinformatic tools are available to data-mine high-throughput results from NGS, GWAS and other functional data sources. Using statistical and computational tools, data-mining of experiments and public web databases can be combined to give an unbiased candidate gene list for a specific complex disorder (Tiffin *et al*., 2006).

This *in silico* approach to candidate gene prioritisation was first implemented in the Genes to Diseases (G2D) tool (Perez-Iratxeta *et al*., 2005), the methodology of which has subsequently been modified and used in a range of other bioinformatic programs. All of these programs have specific inputs and advantages, as well as having certain fallbacks and bias.

1.4 Review of existing prioritisation tools

With the exception of Candid (Hutz *et al*., 2008) and Endeavour (Aerts *et al*., 2009), most previous candidate gene prioritisation tools have focused on integrating a very limited number of data resources. A selection of these tools is presented below, explaining the basis of each method. Tools were chosen to represent the broad range of data sources used, and are analysed with respect to their ability to find novel complex genes for poorly characterised diseases. G2D and PosMed (Yoshida, *et al*., 2009) are based on text-mining available data sources; Candid and Endeavour both use multiple data sources but with a genome-wide method versus a guilt-by-association approach, respectively, and Genewanderer (Kohler, *et al*., 2008) prioritises genes based on protein-protein interactions.

1.4.1 G2D

Perez-Iratxeta *et al*. (2005) are accredited for the design of Genes to Diseases (G2D) as the first formalised candidate gene prioritisation tool. This program prioritises genes related to a disease by text-mining OMIM (http://www.omim.org/), MEDLINE (http://www.ncbi.nlm.nih.gov/pubmed) and Ref-Seq (http://www.ncbi.nlm.nih.gov/refseq/) databases for a list of weighted MeSH C terms (disease category) and Gene Ontology (GO) terms. An OMIM identifier (ID) and candidate locus are used as inputs, with all genes within the locus ranked for involvement in disease aetiology.

Should users be interested in more than one loci, common within analysis of complex diseases, the search would have to be individually repeated for each additional loci, and results manually combined. The genomic range specified is limited to a certain range due to search methods being computational intensive. Should a disease not be specified by an OMIM ID, the authors advise the use of an ID for a similar disease. The use of functional annotation specified by GO is dependent on the quality and completeness of such annotation, and also shows bias against novel genes that may have little or no annotation. In addition, G2D does not appear to be maintained, since the last update occurred in March 2007. Many other tools also share this issue of not being maintained.

1.4.2 PosMed

The PosMed tool (Yoshida, *et al*., 2009) utilises the General and Rapid Association Study Engine (GRASE) as a method to text mine a set of more than 17 million medical and biological documents in relation to the user-defined disease query. GRASE returns a set of documents related to the disease and the genes found within these documents. Users are also required to specify a genomic region as a set of candidates. The set of documents is subsequently evaluated for statistical significance among the candidate genes based on co-citations in documents, protein–protein interactions (PPI), ortholog genes or co-expression data. Only genes that receive a p-value less than 0.01 are finally ranked according to their likelihood for involvement in the queried disease. This makes PosMed unique in that it returns a limited list of ranked candidate genes.

PosMed allows users to input either a disease name or a list of terms related to the disease. Bornigen *et al*. (2012) found that the set of keywords specified by the user strongly influences

results returned by the tool. Like G2D, PosMed is limited in analysing a single linked genomic locus at a time. Since PosMed relies solely on curated information, the methodology shows heavy bias toward well studied genes, as novel genes will not be linked to the disease by literature or other annotation.

1.4.3 Candid

Candid (Hutz *et al*., 2008) is one of the few existing tools that makes use of several different types of data sources, including literature, protein domains, gene conservation and expression information, PPI, linkage and association analysis and an option to add custom data. Options to search the whole genome or just specified loci are available. Input can be given by either keywords or training genes. This tool was specifically designed to analyse complex genetic diseases; however, users define their own weighting criteria for the various data sources, meaning that a thorough understanding of the queried disease is necessary in order to gain valuable and reliable output. Data sources not incorporated into this tool include homolog information, pseudogene analysis and enrichment within regulatory regions and miRNA target sites.

1.4.4 Endeavour

Endeavour (Aerts *et al*., 2009) is another tool that incorporates multiple data sources. This tool prioritises a set of user-specified candidate genes according to characteristics found over-represented in a training set, defined by the user, as compared to the complete genome. Gene function, gene expression, protein sequence, mutant phenotypes and PPI are all assessed.

For poorly characterised diseases, such as bipolar disorder, training genes may either not be available or show any strong overrepresentation of the same characteristics. Users effectively limit the candidate search space by having to provide the genes in which to search for an association. Since Endeavour relies solely on the guilt-by-association principle, candidates that differ in characteristics from the training genes will not be prioritised.

1.4.5 Genewanderer

Like Endeavour, Genewanderer (Kohler, *et al.*, 2008) requires users to specify a set of training genes. A chromosomal locus of interest is also required as an input. This tool builds a gene interaction network from the training genes using HPRD (http://www.hprd.org/), BIND (http://bind.ca), BioGrid (http://www.thebiogrid.org/), IntACT (http://www.ebi.ac.uk/intact/), DIP (http://dip.doe-mbi.ucla.edu) and STRING (http://www.string-db.org/). All genes within the locus are evaluated for proximity to this network. Users may choose between four different proximity measures; however, the random walk method performed best during benchmarking. This approach starts at a given training node and transitions to a randomly selected neighbour in order to calculate the probability of reaching a candidate node.

Genewanderer operates solely on PPI data, and thus can only be used for genes that have their interactions mapped, which may be rare for novel genes. Network-based prioritisation is greatly dependent on the quality of interaction data. Although multiple interaction databases are incorporated, they are still incomplete. Electronic annotations may also be a source of unreliable interactions and therefore produce false positives.

Existing tools, such as the ones described above, place high emphasis on previous knowledge. For poorly characterised diseases, this requirement limits the success of prioritisation.

1.5 Problem statement

Besides the restricted application of current tools to under-studied diseases, bioinformatic solutions have not been appropriately utilised. Candidate gene prioritisation has been used in several studies of complex disorders; however, most of this research is specific to the disease under study and not been made into a usable tool that can be applied to other complex genetic diseases. Sun *et al*. (2009) performed candidate prioritisation for schizophrenia by manually combining data from literature, linkage, association and expression sources. They then assessed 625 different weight matrices to evaluate the optimal weights at which to combine the sub-scores for each candidate gene that allowed for the highest ranking of a set of training genes. This research selected 502 genes for follow up study in a schizophrenia cohort. This exhaustive approach is not suitable for wide-scale use and has limited clinical validity since many possible

candidates remain. Other studies, similar to the above example, have essentially created a redundancy of bioinformatic work since the results can only be applied to the specific disorder studied.

Following from these arguments, there is a need for a new, generic tool to be designed that can be applied to any complex genetic disease. The tool should incorporate a maximum number of data resources as possible, thereby utilising the vast amount of genetic data that has been produced specifically within the last decade. This tool must attempt to avoid the problems identified in current tools that limit their success for complex disorders.

1.6 System considerations

The different approaches used in a variety of existing tools were analysed to identify the common advantages and pitfalls to consider in the design of a new prioritisation tool. These factors include the bias to well-studied genes, information that is available on homologs, dependence on database quality and validation of the ranking of genes.

Tools that require a training set (genes that have already been associated with the disease), rely on previous biological characterisation of the trait, and can create a bias in selecting candidate genes. These tools are limited in that it is assumed that the undiscovered disease genes will be consistent with what is already known about a disease and/or its genetic basis, which is not true, necessarily (Hutz *et al*., 2008). Such a method is particularly poor for studying diseases where few or no disease genes are associated with the phenotype or where the known disease genes account for only a small percentage of cases presenting with the disease, such as in phenotypes with reduced penetrance or variable expressivity (George *et al*., 2006).

The same bias is created when a tool relies heavily on scientific literature resources, as these do not include uncharacterised genes. Contrary to this, sequence-based tools reward genes for their putative involvement in disease regardless of how frequent research has been conducted to associate the gene to a particular phenotype (Adie *et al*., 2005).

A further consideration is the inclusion of homolog information, especially for diseases that are not well understood or whose mechanism of action has yet to be discovered. Data on animal

models of the disease is important, since data are collected under strictly controlled conditions not possible in human studies (Simmons, 2008).

Tools that prioritise genes based on functional annotation rely on the accuracy and content of the databases they gather information from. This means that data on a query disease must already reside in the database (Hutz *et al*., 2008). The experimental quality of associations and interactions documented in a database determines the accuracy of the gene prioritisation, especially for network-based prioritisation. Experimental evidence, such as from Yeast 2 Hybrid assays, and manual curation, such as used in the SwissProt database (http://www.uniprot.org/), is more reliable than inferred interactions or electronic curation. Higher weighting should be given to sources that are more reliable, whist at the same time candidates should not be penalised due to missing information.

Once all data sources have been integrated to give an overall score per gene, ranking of the genes needs to be evaluated for accuracy. Benchmark tests are needed to assess the tool's ability to find relationships between a given disease and the genes within a queried genome. Benchmarking must be performed with diseases that already have a list of genes confidently known to be associated with the phenotype (Perez-Iratxeta *et al*., 2005).

All of the above system considerations need to be incorporated into a new gene prioritisation tool to undertake the growing bioinformatic demands of analysing complex genetic diseases. Once a new tool has been designed and validated, it may be used to prioritise poorly-characterised diseases, the results of which can then be clinically tested on patient samples.

The burden of complex genetic diseases on public healthcare systems will only subside if we first understand the aetiology and biological mechanisms of these phenotypes. GWAS and NGS technologies are generating overwhelming amounts of data that need to be processed in order to extract significant results. Candidate gene prioritisation is the link between this data and the application of these findings to clinical genetic research.

This paper continues on to describe the development of Data Integrated Genetics (DIG), which was designed according to the system considerations discussed above. It is an innovative tool to integrate data for any human genetic disease from a variety of data sources. We aim to allow human geneticists to effectively use the vast amounts of freely-available data so as to focus their

experiments on the most probable causative genes, ultimately leading to the discovery of disease pathogenesis, treatment and prevention, whilst simultaneously saving time and funding.

1.7 Aim

To create a new web-based data integration tool for candidate gene prioritisation using BPD as a test case.

1.7.1 Objectives

- Implementation:
    - To integrate as many freely-available data sources as possible into a usable, web-based tool
- Validation:
    - To validate the tool's performance with several complex disease datasets
    - To compare the tool's performance to existing prioritisation tools
- Application:
    - To test BPD as a case study to evaluate the tool in indentifying novel susceptibility genes

2 Methodology

In the design of DIG, a combination of different data sources and web tools were used in order to take advantage of the vast amount of freely-available data. DIG does not require the entry of any training disease genes, but instead the candidate search space is created from information that is automatically retrieved from databases for genes and chromosomal bands already known to be associated with the queried disease from literature, linkage and homolog data. DIG then integrates these three sources as well as a number of other data sources to evaluate the likelihood of each of the genes being involved in the query disease. The DIG script was written in Python 2.75 and the website was implemented in Cherrypy, with all major browsers supported. Supplementary figure 1 shows the total workflow of the program. The full python scripts are available from http://www.cbio.uct.ac.za/DIG/source. Please contact the author for access credentials.

2.1 Query expansion

Firstly, DIG uses the Disease Ontology (DO) (Schriml *et al*., 2012) to expand the search query entered by the user. The DO database combines disease names and identifiers from MeSH (Nelson *et al*., 2004), ICD (Ayme *et al*., 2010), OMIM and NCI (Sioutos *et al*., 2007) to create a directed acyclic graph for over 8000 inherited, developmental and acquired human diseases. This hierarchy maps the direct and indirect relationships between diseases.

DIG uses a GET Request (http://docs.python-requests.org/en/) to extract the DO unique ID for the user query term. The returned DO ID is then used in a second GET Request to access the REST Application Program Interface (API) used by the DO database (http://www.disease-ontology.org/search?adv_search=True&operator=AND&field-1=name&value-1='+keyword).
This returns a JSON packet containing all the metadata for the queried term, including parents, children, synonyms, definition, name and alternative IDs. Regular expressions are used to build lists of synonyms, parent and child terms from the JSON data.

2.2 Literature search

Pubmed literature is the first data source used to extract genes related to the user-defined disease. The Entrez EUtilities (Sayers, 2008) suite of tools is used to access the Pubmed API. EUtilities

use a fixed URL syntax to query the National Centre for Biotechnology Information (NCBI) databases. The user is required to supply their email address in order to use this service.

A search is created by concatenating all expanded keywords from DO into a string with the deliminator 'OR' to separate different keywords. A query array of terms is created by slicing this global concatenated query term into substrings of 150 characters or less. This is done to comply with the character limit set on the NCBI API.

From EUtilities, the ESearch tool is used to find all Pubmed articles matching the query string(s). ESearch iterates through all elements in the query array, retrieving a list of Pubmed article IDs matching the query. The Pubmed IDs are added to a set in order to automatically remove redundancy. Links to each of these articles is created by concatenating the Pubmed home URL with the Pubmed ID.

A mapping file 'gene2pubmed', available from the Entrez file transfer protocol (ftp) site (ftp://ftp.ncbi.nlm.nih.gov/gene/DATA/), is used to find all human genes mentioned within an article from the list of matching Pubmed IDs. All genes are added to the pool of possible candidate genes.

A literature score is calculated for each of the genes identified in the literature search. This overall literature score is calculated as shown by Equation 1.

$$litScore = \frac{\text{number of disease articles the gene appeared in}}{total\ disease\ articles\ mapped\ to\ human\ genes} \quad (1)$$

2.3 Linkage data

Linkage data is collected from a tool called aBandApart (van Vooren *et al*., 2007). This tool text-mines MEDLINE abstracts relating to cytogenetic bands for an overrepresentation of a queried biomedical concept. In this way, aBandApart is able to link large genomic aberrations to genetic diseases, syndromes and dysmorphology in human development. The relationship between band and concept is quantified by calculating a p-value for the observed number of papers that are associated to a band and that mention the query concept or one of its synonyms.

WebDriver, from the python module Selenium (https://pypi.python.org/pypi/selenium), designed to emulate an instance of the Moxilla Firefox browser, is used to submit a form to aBandApart. The main term returned from DO is used for the linkage query, unless this returns no linkage data, where upon the parent term is used.

XPath, from the lxml python module (https://pypi.python.org/pypi/lxml), is then used to extract the cytogenetic bands and their corresponding p-values from the returned extensible markup language (xml) data. The user is requested to supply a significance cut-off (default = 0.05) to then filter the returned bands.

A file of all human genes mapped to a cytogenetic band was downloaded from BioMart (http://www.ensembl.org/biomart/martview/2cfef8ac90eebc767b87136f3e3c8e03). This local file is used to extract all genes designated by the filtered bands. Genes from the same band are assigned the same p-value and score as per Equation 2. These genes are added to the candidate gene pool.

$$linkScore = 1 - Pvalue \tag{2}$$

2.4 Homolog data

Homolog data is extracted from a text-mining tool called Genie (Fontaine *et al*., 2011). Using the MEDLINE, Gene and HomoloGene databases from NCBI (http://www.ncbi.nlm.nih.gov/), Genie allows for the prioritisation of all genes from a given species related to a biomedical topic. This tool assesses the significance of the abstracts that contain genes associated to the target model organism for their relevance to the chosen biomedical concept. The false discovery rate (FDR) for each gene-to-topic relationship is then calculated.

The DIG user is asked to choose a model organism from a list in which to search for homolog data. They may select from the following organisms: *Mus musculus* (mouse), *Rattus norvegicus* (rat), *Drosophila melanogaster* (fruit fly), *Sus scrofa* (pig), *Macaca mulatta* (rhesus macaque), *Oryctolagus cuniculus* (rabbit) and *Pan troglodytes* (chimpanzee). The default choice is *Mus musculus*.

DIG posts an HTTP Request to Genie using the python module Mechanize (https://pypi.python.org/pypi/mechanize/). The script then sets the biomedical concept as the main term returned by DO and the target species as the taxonomic ID of the species chosen by the user. A regular expression is subsequently used to search the returned request data for the homolog gene identifiers and their corresponding FDRs.

The Entrez Gene identifiers of the model organisms are converted to their corresponding Ensembl Gene identifiers using local mapping files from BioMart. A second local BioMart file then maps these IDs to Ensembl human homologs. These are added to the candidate gene pool. A homolog score is calculated for each human gene as described by Equation 3.

$$homoScore = 1 - FDR \tag{3}$$

Model organism genes may occasionally map to multiple human homologs; these genes all receive the same FDR related to the model organism source.

2.5 Sequence data

Data from a previous tool, PROSPECR (Adie *et al.*, 2005), was used to obtain information about gene sequences. The authors found that disease genes share certain properties in gene structure as compared to genes not involved in disease. The disease characteristics include longer gene and protein length, larger number of exons, well conserved mouse and rat homologs, secretion of the gene product (validated by the presence of a signal peptide), longer 3' untranslated regions (UTR) and larger distance to the nearest neighbouring gene. Using this information, PROSPECR uses an alternating decision tree algorithm to classify genes as likely to be involved in hereditary disease or otherwise.

A file containing pre-computed PROSPECR scores for 22, 240 genes was downloaded (http://www.genetics.med.ed.ac.uk/prospectr/.) DIG searches this local file for all matches to the candidate gene pool, returning their PROSPECR scores or "None" if the gene is not found.

From all of the above sources, genes will have a minimum of one sub-score and a maximum of four. If a gene had only one data source of evidence, this single sub-score would be used as an

estimate of the total integrated score. Should a gene have more than one sub-score, a method to combine the sub-scores is necessary.

2.6 Data integration

The aim of data integration is to achieve a balance between limiting false positives and the avoidance of penalising genes with few sub-scores due to incomplete databases. Owing to the different quality and amount of data available from different information sources, an integrated score that assigns different weights to each sub-score is needed. However, this weighting will be dependent on the research conducted on each disease queried, and thus a standardised weighting will not be possible. It is thus necessary to calculate a weight matrix in response to the data retrieved for the user's disease.

If we assume that each sub-score is an unbiased, independent estimate of the gene's association to the disease, we can then apply a linear regression model to the data such that a higher integrated score represents a higher likelihood for a gene's involvement in disease aetiology. The relative contribution of each sub-score to the reliability of such a measure can be described by a weight matrix that accounts for the different types of data available for the disease. To prevent over-fitting of the data to this linear regression model, the sum of all weights must equal one (Equation 4).

$$w_{lit} + w_{link} + w_{homo} + w_{seq} = 1 \tag{4}$$

To maintain this logic for genes that contain missing or null sub-scores, the constant $\frac{1}{\sum w_i}$ is introduced. The denominator is the sum of weights for all non-null sub-scores. This constant is then multiplied by the product of each sub-score and its weight as demonstrated by Equation 5. For genes that have scores for all four values, this constant will equal one, with the integrated score simply calculated as the sum of the sub-scores multiplied by their respective weights. However, for genes that have missing values, this proportion will serve to readjust the weights of non-null sub-scores so that Equation 4 is still satisfied.

$$Integrated\ score = \frac{1}{\sum w_i}(w_{lit}s_1 + w_{link}s_2 + w_{homo}s_3 + w_{seq}s_4) \tag{5}$$

Variance of a sub-score is used as an approximation for the reliability of each data source as an estimator of whether a gene is causal for the disease or not. All genes in the candidate pool must have at least some relation to the disease in order to have been considered a candidate. Thus, a higher variance for non-null values of a sub-score is seen as decreased confidence in the data source, and hence a smaller weight is assigned for the sub-score to achieve the most accurate integrated score.

The mean, $E\chi_i$, is calculated from all non-null values for a given sub-score, and then used to calculate the variance $(E\chi_i - s_i)^2$ of each gene from this expected value. The variance for a sub-score (σ_i^2) is equal to the sum of variances for all genes divided by the number of genes, j, as described by Equation 6. Since variance is inversely proportional to the weight of the sub-score, we assign it a relative weighting of $\frac{1}{\sigma_i^2}$.

$$Approximate\ variance_{sub-score} = \sigma_i^2 = \frac{\sum_{i=1}^{j}(E(\chi_i)-s_i)^2}{j} \quad (6)$$

In order to comply to Equation 4, where the sum of all non-null weights must equal to one, we calculate an absolute weight by diving the relative weight for the sub-score by the sum of relative weights for all sub-scores as shown in Equation 7.

$$w_i = \frac{\frac{1}{\sigma_i^2}}{\sum_{i=1}^{n}\frac{1}{\sigma_i^2}} \quad (7)$$

Variance of each sub-score will then be used to calculate the most appropriate weighting for each respective sub-score. The integrated score for each gene is then calculated with Equation 5.

2.7 Filtering

Once integration is completed, candidate genes are then filtered to extract false positives and to limit results to the most promising genes, generating output as quickly as possible whilst not losing valuable information.

2.7.1 Number of data sources

A sub-score of zero is still considered an observation and thus increases the confidence of a score as compared to a null sub-score. With more null sub-scores, the confidence for a gene decreases and the possibility for a gene to be significant purely by chance increases. For this reason, genes that have fewer than three sources of evidence are automatically filtered out by DIG. This is to eliminate the effects from false positives that arise from the three data sources used to build the candidate gene pool, since all genes should theoretically have a sequence score. If fewer than 15 candidates are returned, then the filter is automatically adjusted to include genes with two or more sources of evidence.

2.7.2 Pseudogene data

A second method to eliminate false positives is via pseudogene screening. A comprehensive list of pseudogenes was created by extracting and combining Ensembl IDs using BioMart's IG_C_pseudogene, IG_J_pseudogene, IG_V_pseudogene, polymorphic_pseudogene, processed_pseudogene, pseudogene, TR_J_pseudogene and TR_V_pseudogene datasets. This resulted in a total of 15 524 genes.

DIG reads this file to find any of the candidate genes in the list of known pseudogenes. Should a match be found, the gene is removed from the list of potential candidate genes.

2.8 Ranking

Ordered lists are created in which genes are ranked in decreasing order according to one of the sub-scores. Global ranking is performed using the integrated score as a sorting method. To account for genes that share the same score, the rankings of tied scores are replaced with the top rank of the tied genes so that these genes receive the same rank. Null sub-scores are ranked lowest.

2.9 Data validation

To further highlight false positives, two validation strategies are implemented. Tissue annotation is included in the output of DIG so that the user may critically determine if expression of the gene is plausible to be involved in the aetiology of the disease. Genes that are expressed in the same tissues that are differentially regulated in the disease are good candidates and boost confidence for the gene to be a possible candidate.

Single nucleotide polymorphisms (SNPs) from dbSNP (http://www.ncbi.nlm.nih.gov/SNP/) are also mapped to the top ranking candidate genes, along with the region within the gene that the SNP is found in, being either intronic, exonic, upstream, downsteam, or within the regulatory region. This was achieved by downloading a local version of dbSNP. The file was then annotated with SNPEffect (http://snpeff.sourceforge.net/) and dbNSFP (Liu *et al.*, 2013). Fields were extracted to make a new local file containing rs ID, gene name, SNP effect, SIFT (http://sift.jcvi.org/) and PolyPhen (http://genetics.bwh.harvard.edu/pph2_/dokuwiki/) scores for all ranked genes.

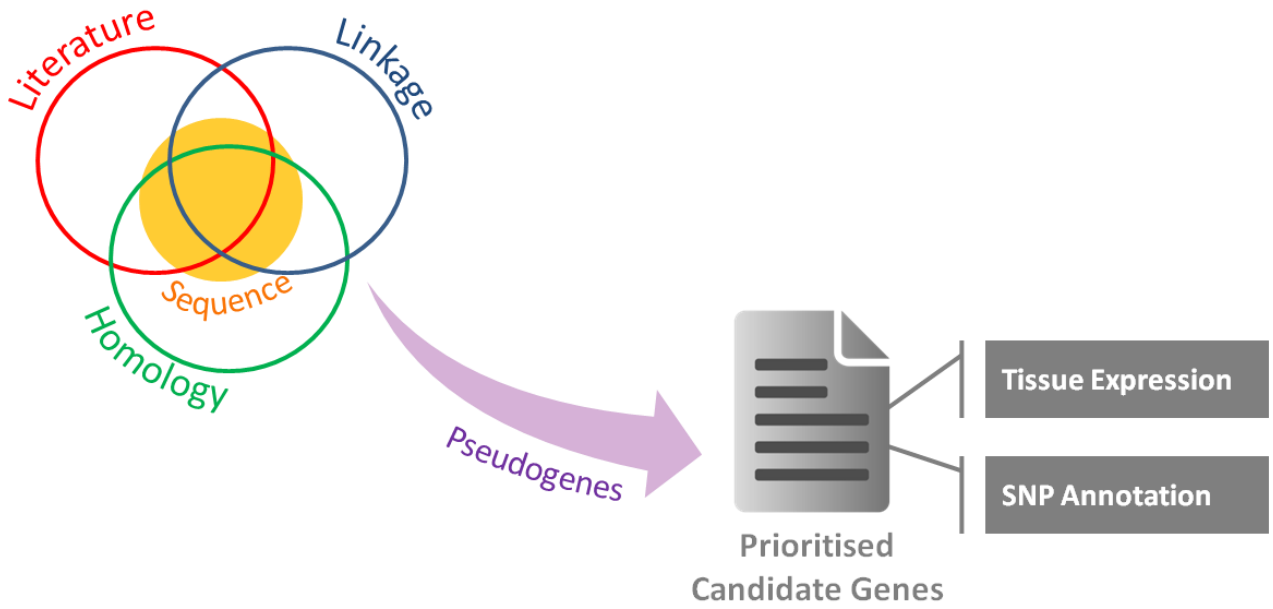

Figure 1. Diagram depicting the overall methodology employed by DIG to prioritise candidate genes. Literature, linkage and homolog data sources are used to collect a large pool of possible candidates which are subsequently further analysed and filtered before ranking.

2.10 Performance measures

2.10.1 Evaluation of DIG before and after variance matrix application

In order to evaluate the accuracy of DIG, four well-studied complex diseases, all of whom have known associated genes, were tested. Supplementary table 1 shows the training genes used for breast cancer, type 2 diabetes mellitus, asthma and Alzheimer's disease that were taken from Hutz *et al*. (2008). Linkage cut-off was set to 0.05 and *Mus musculus* used as the animal model for all diseases tested. Indicators used in Bornigen *et al*. (2012) were then used to assess the performance of DIG.

STATA12 was used to calculate the area under the curve (AUC) for receiver-operating characteristic (ROC) curves for each disease to evaluate global accuracy. Sensitivity of the tool, also known as the true positive rate (TPR), was calculated at 5 per cent, 10 per cent and 30 per cent for each disease. Response rate for a disease, defined as the number of training genes included in the DIG output, was also evaluated.

A one by four weight matrix with equal weights for each score was initially used. Performance indicators were calculated for all sub-scores and for the integrated score by comparing the DIG output of ranked genes to the list of training genes. Analysis of the diseases was repeated using the variance weight matrix as per Equation 6. Differences in performance measures were calculated for rankings made before and after the application of the new matrix.

2.10.2 Comparison of DIG to existing candidate gene prioritisation tools

Trachevent et al., (2010) created a decision tree that categorises existing candidate prioritisation tools according to input used and outputs produced. Using this tree, tools most similar to DIG were selected to compare performance statistics. Endeavour and PosMed were selected since both tools use keywords as an input to assign ranks to candidate genes. Other tools were excluded due to their use of training genes as an input or because genes were simply selected and not ranked according to likelihood of disease involvement.

2.11 Tool application to BPD

DIG was finally run for bipolar disorder. “Bipolar disorder” was entered as the query keyword and *Mus musculus* set as the animal model. The linkage cut-off value was set to a more lenient 0.1 since most positional studies show weak genetic signals for linkage data (Serretti and Mandelli, 2008).

Ranked genes were then evaluated from a comprehensive list of training genes for BPD attained from Serretti and Mandelli (2008) and Chen, *et al*. (2009). The top 41 genes were manually evaluated for involvement in BPD aetiology by reading associated literature and assessing pathways, Gene Ontology (GO) terms, protein-protein interactions and searching Genetics Association Database (http://geneticassociationdb.nih.gov/cgi-bin/index.cgi) for each candidate.

3 Results

DIG utilises hypertext mark-up language (html) form submission to gain input about the queried disease from the user, followed by dynamic interaction to gain parameters for filter cut-off values before displaying the output. DIG is freely available from http://www.cbio.uct.ac.za/DIG/; no registration is required. A tutorial document is available on the site, detailing the input needed and answering common questions.

3.1 Implementation: The user's experience

Users are required to input the name of the disease they wish to query, a linkage significance cut-off value and also supply a valid email address as required for the Entrez ESearch tool. A list of seven radio buttons allows the user to select the animal model in which to search for homolog information. Once the user has entered all required input data and submitted the form, a loading page is displayed until job completion. Time taken for each job is dependent of the extent of research conducted on the disease. A minimum time of three minutes is expected, but can extend up to 10 minutes if analysing well-documented diseases.

Thereafter, a line graph displaying the distribution of integrated scores for all ranked genes is displayed. The user is asked to choose the number of genes to display in the results. The output page displays summary statistics for the gene prioritisation with a table showing ranked genes displayed below this.

For each gene, sub-scores for literature, linkage, homology and sequence data is shown alongside the relative rankings within each of the categories. Tissues in which the gene is expressed are listed. Any non-synonymous SNPs found in the gene are listed, with their corresponding gene region and effect. SIFT and Polyphen scores for each of these SNPs are also displayed if available. The number of data sources found for the gene is presented, having a maximum of 4. The integrated score for the candidate is highlighted at the end of the table row.

Links to Pubmed articles that were used to rank the gene in the literature sub-score are presented for easy navigation to relevant articles. This allows the user to perform quick evaluation of the results and to perform a follow up literature search. Links to the Ensembl gene sequence allow for

further analysis of the candidate and may help geneticists in their design of polymerase chain reaction (PCR) primers or clone constructs.

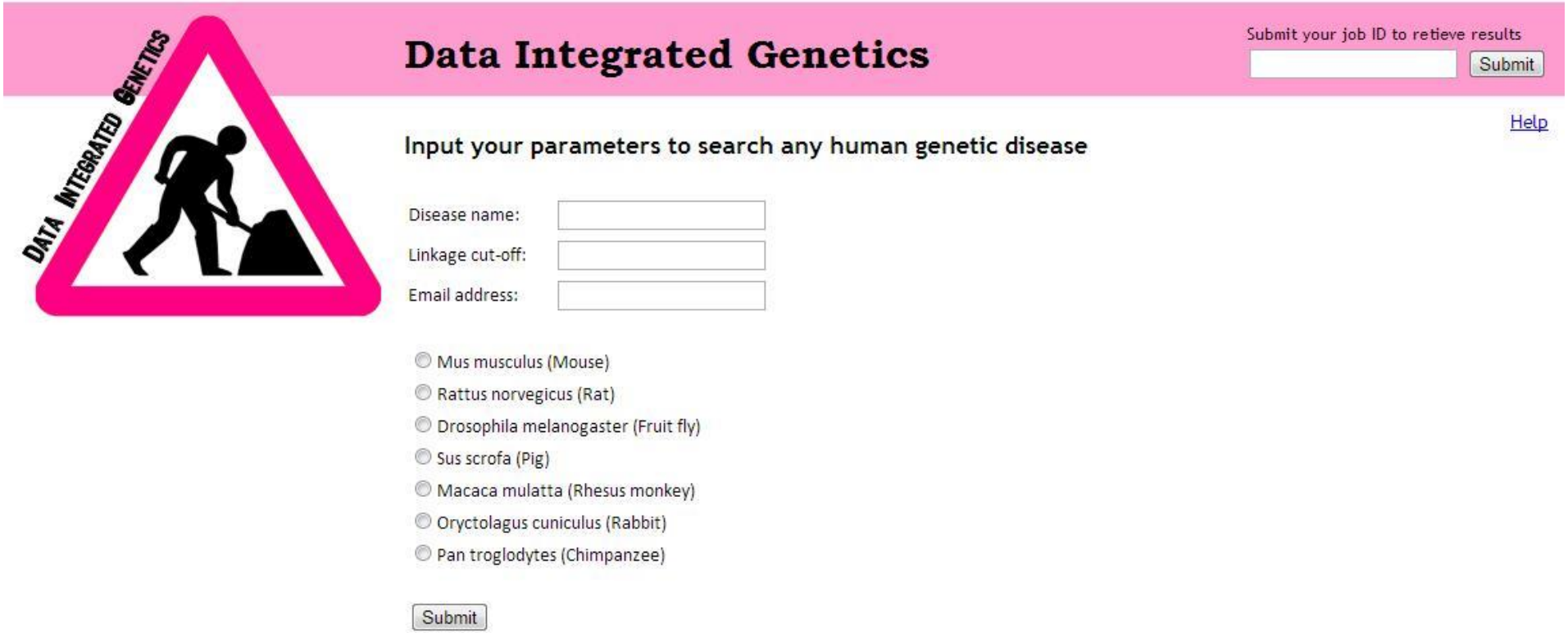

Figure 2. A print screen of the DIG form submission page. The four simple input requirements allow for easy use of the tool. At the top right corner, users may specify a job ID of past queries made within the previous 5 days. A link to the help document is also available.

3.2 Validation: Performance statistics

3.2.1 Accuracy of variance weight matrix

The variance weight matrix used in the DIG methodology was compared to the use of equal weighting for sub-scores by analysing performance indicators for four reasonably well characterised diseases. This was to evaluate the impact of using variance as a reliability measure on the accuracy of DIG in ranking candidate genes. Higher values for indicators indicate greater accuracy. TPR5, TPR10 and TPR30 represent the probability of finding true positive results (i.e., training genes) when only analysing the top 5%, top 10% and top 30% of the ranked genes respectively. AUC measures the probability that DIG will rank any one of the training genes higher than the rest of the genome. A minimum score of 0.5 for AUC means that tool ranks training genes randomly, whereas a maximum score of 1.0 indicates that training genes are always ranked first. Response rate describes the number of training genes ranked in the output results of the tool. For each training gene filtered out of the results, response rate decreases proportionally to the total number of training genes used. For example 42 breast cancer training genes were used, of which 30 were ranked by DIG, denoting a response rate of $30 \div 42 \times 100 = 71.4\%$.

A marginal improvement in performance was seen when using the variance weighting system as compared to the use of equal weights in most categories. However, the new matrix did not improve the response rate achieved by DIG. Table 1 shows the percentages achieved for performance indicators before and after application of the variance matrix to DIG.

Table 1. Comparison of performance indicators for DIG using equal weights for all sub-scores as compared to the use of variance weighting system described in the tool methodology.

Weighting	TPR5%	TPR10%	TPR30%	AUC	Response rate %
Equal weights	51.2	52.8	69.0	0.539	64.2
Variance weights	58.6	68.7	79.2	0.556	64.2
Difference	+7.4	+15.9	+10.2	+0.170	0

3.2.2 Comparison to existing tools

The performance of DIG was compared to Endeavour and PosMed. It was found that both DIG and PosMed returned far fewer candidate genes than Endeavour did. Table 2 shows that DIG outperformed both the existing tools in TPR5 and TPR10; however, Endeavour displayed the best performance in all other categories. Further analysis of these results is discussed later.

Table 2. Comparison of performance indicators for DIG, Endeavour and PosMed. Best performing tool for each indicator highlighted in bold.

Tool	TPR5%	TPR10%	TPR30%	AUC	Response rate %
DIG	**58.6**	**68.7**	79.2	0.556	64.2
PosMed	4.7	11.9	23.8	0.560	50.0
Endeavour	26.2	42.9	**90.5**	**0.820**	**100.0**

3.3 Application: Findings for novel BPD associations

The analysis of BPD resulted in 822 genes being ranked, with a maximum integrated score of 0.999 and minimum integrated score of 1.9e-04. Before filtering, 25 257 possible candidates were data-mined from literature, linkage and homology sources. Of these, 13 mouse homologs were found to be related to bipolar disorder. Equation 8 shows the variance weight matrix used for prioritisation. Of the 292 training genes used to evaluate the DIG results, 37 genes were returned in the output table.

$$[0.702, \quad 2.55e-04, \quad 0.298, \quad 1.35e-05] \tag{8}$$

The top 41 genes, representing the top 5% of ranked genes from the DIG results, were analysed and shown in Table 3. Of these genes, eight were found to have no previous association to BPD. In particular, the huntingtin gene (*HTT*) was a surprising result, ranking 8th in the DIG output. Enrichment analysis of the proteins encoded by these genes was performed. STRING showed no enrichment for PFAM or InterPro domains; however, protein interactions were highly enriched, with a p-value of 0.00. Figure 3 shows the 111 direct interactions that were observed among the top 41 proteins. Functional enrichment was also found in the three GO categories as well as known KEGG pathways. Table 4 lists the 5 most significant terms per functional category.

Table 3. Gene names of the top 41 genes for BPD as ranked by DIG. Previous association to BPD is also listed for each gene. “Training” indicates that the gene was used in the training gene set.

Rank	Gene symbol	Gene name	Association
1	SLC6A4	Solute carrier family 6 (neurotransmitter transporter, serotonin), member 4	Training
2	BDNF	Brain-derived neurotrophic factor	Training
3	HTR1A	5-hydroxytryptamine (serotonin) receptor 1A	Yes
4	DISC1	Disrupted in schizophrenia 1	Training
5	DRD4	Dopamine receptor D4	Training
6	SLC6A3	Solute carrier family 6 (neurotransmitter transporter, dopamine), member 3	Training
7	DTNBP1	Dystrobrevin binding protein 1	Training
8	HTT	Huntingtin	No
9	GSK3B	Glycogen synthase kinase 3 beta	Yes
10	COMT	Catechol-O-methyltransferase	Yes
11	HTR2A	5-hydroxytryptamine (serotonin) receptor 2A	Yes
12	TPH2	Tryptophan hydroxylase 2	Training
13	TPH1	Tryptophan hydroxylase 1	Yes
14	ANK3	Ankyrin 3, node of Ranvier (ankyrin G)	Yes
15	CYP2D6	Cytochrome P450, family 2, subfamily D, polypeptide 6	No
16	DRD3	Dopamine receptor D3	Yes
17	P2RX7	Purinergic receptor P2X, ligand-gated ion channel, 7	Yes
18	ESR1	Estrogen receptor 1	No
19	HTR1B	5-hydroxytryptamine (serotonin) receptor 1B	Yes
20	CNR1	Cannabinoid receptor 1 (brain)	Yes
21	AKT1	V-akt murine thymoma viral oncogene homolog 1	Yes
22	IMPA2	Inositol(myo)-1(or 4)-monophosphatase 2	No
23	S100B	S100 calcium binding protein B	Yes
24	DRD1	Dopamine receptor D1	Yes
25	ARNTL	Aryl hydrocarbon receptor nuclear translocator-like	Yes
26	TH	Tyrosine hydroxylase	No
27	CHRNA7	Cholinergic receptor, nicotinic, alpha 7	Yes
28	XBP1	X-box binding protein 1	Yes
29	CRY2	Cryptochrome 2 (photolyase-like)	Yes
30	DGKH	Diacylglycerol kinase	Yes
31	NDUFV2	NADH dehydrogenase (ubiquinone) flavoprotein 2	Yes
32	CRH	Corticotropin releasing hormone	Yes
33	GRM7	Glutamate receptor, metabotropic 7	Yes
34	TAAR6	Trace amine associated receptor 6	Yes
35	DAO	D-amino-acid oxidase	Yes
36	OPRM1	Opioid receptor, mu 1	No
37	OXT	Oxytocin, prepropeptide	No
38	ADRBK2	Adrenergic, beta, receptor kinase 2	Yes
39	GABRB2	Gamma-aminobutyric acid (GABA) A receptor, beta 2	Yes
40	ADRA2A	Adrenergic, alpha-2A-, receptor	No
41	CCL2	Chemokine (C-C motif) ligand 2	Yes

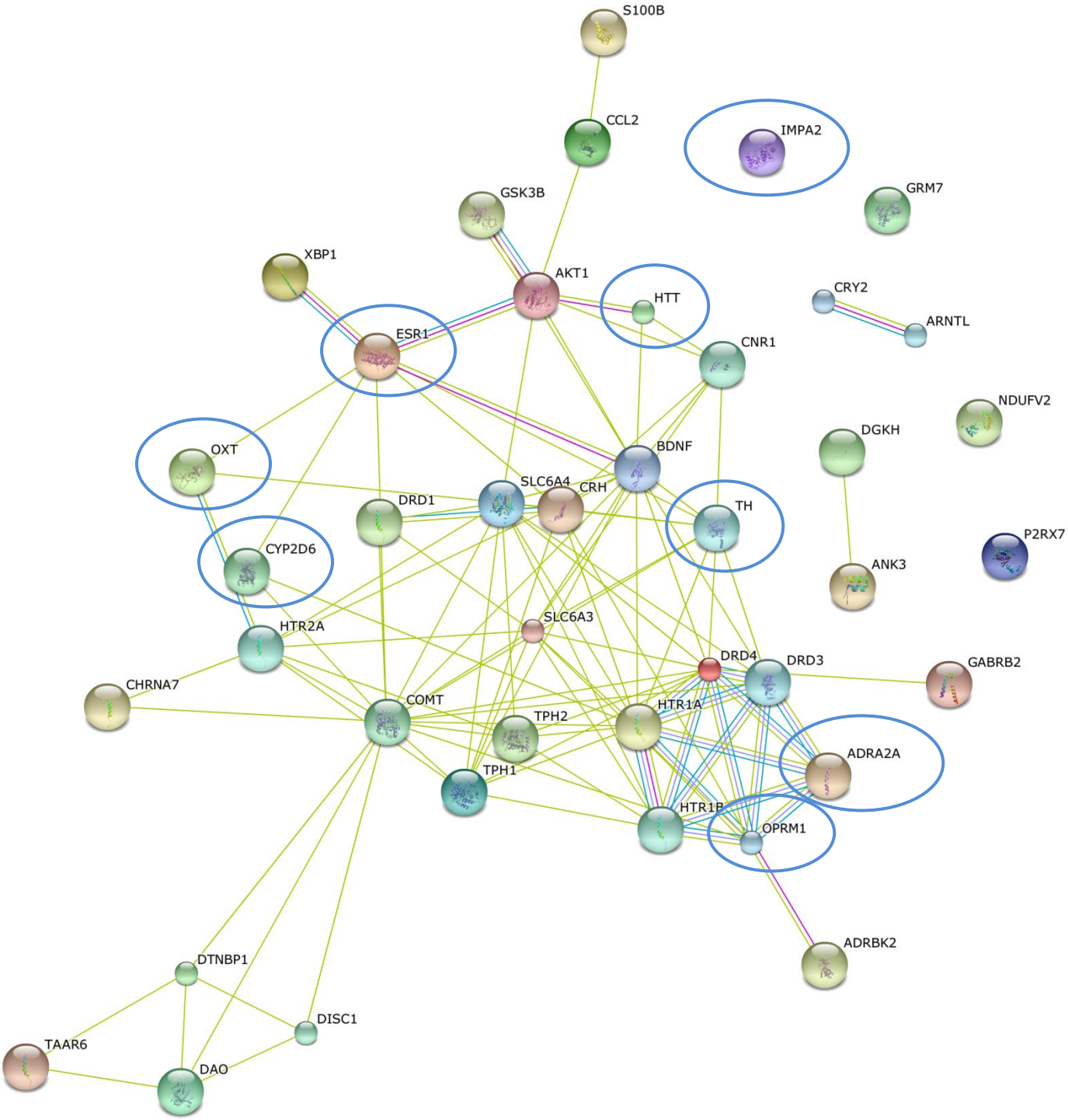

Figure 3. Protein-protein interactions between the top 41 ranked genes from the DIG analysis of BPD as mapped by STRING. Proteins circled in blue did not show previous association to BPD.

Table 4. The top 5 significantly enriched terms for several functional annotation categories when analysing the top 41 proteins for BPD as ranked by DIG.

Category	ID	Term
GO Biological process	GO:0007611	learning or memory
	GO:0050890	cognition
	GO:0007613	memory
	GO:0051952	regulation of amine transport
	GO:0007610	behaviour
GO molecular function	GO:0043176	amine binding
	GO:0008144	drug binding
	GO:0035240	dopamine binding
	GO:0008227	G-protein coupled amine receptor activity
	GO:0043178	alcohol binding
GO cellular compartment	GO:0033267	axon part
	GO:0043005	neuron projection
	GO:0044463	cell projection part
	GO:0043195	terminal button
	GO:0030424	axon
KEGG Pathway	hsa04080	Neuroactive ligand-receptor interaction
	hsa04710	Circadian rhythm - mammal
	hsa04020	Calcium signalling pathway
	hsa00380	Tryptophan metabolism
	hsa04062	Chemokine signalling pathway

4 Discussion

4.1 Prioritisation approach

This study describes the development of DIG, a novel candidate gene prioritisation tool for complex disorders. DIG applied a new approach to the ranking of candidate genes, which has not been used in existing candidate gene prioritisation tools. In this tool, literature, linkage, homology and sequence information is used to gather data on genes. Pseudogene information is used as a fifth source in order to filter candidates. DIG is able to both select and rank genes from a candidate search space that is not dependent on user-defined parameters.

4.1.1 Overcoming current issues

A major issue when analysing complex diseases with other available tools, is the requirement of training genes as an input. DIG does not rely on this guilt-by-association method to compare possible candidates, but rather overcomes this obstacle by analysing the variance of data for all genes that have any association to the disease, from various data sources. DIG then assumes that true susceptibility genes will score higher in disease-specific annotation than the rest of the genome. To account for genes in which annotation is missing, the integrated score is calculated such that genes are not penalised for incomplete information. This allows for the identification of novel genes that may be dissimilar to known training genes as well as supporting the discovery of poorly-characterised novel candidates.

The performance of Posmed, GeneDistiller and Candid have shown dependence on the input entered by the user (Bornigen *et al*., 2012). DIG removes the need for the user to specify the weight matrix, unlike Candid. DIG also limits user-bias by the use of query expansion with DOID rather than requesting the user to list multiple query terms.

4.1.2 Limiting false positives

To the investigators' knowledge, DIG is the first tool to include all linked loci for a disease in order to expand the candidate search space. This allows the analysis of all linkage data at once and also removes the need for the user to specify loci as an input. However, false positives are

frequent when extracting all genes from an associated chromosomal band, since only one variant within the region may be contributing towards the disease phenotype. Another source of false positives originates from text-mining which depends on co-occurrence of two terms, but cannot explain the relationship between them. For example if an article was specifying a negative disease-gene association. Articles that include more than one disease may also match a gene to the incorrect disease. To compensate for this, DIG uses stringent filters in order to limit candidates in a way that is not biased against poorly-studied genes.

An issue not related to the performance of a tool, but rather its application, is that most tools do not sufficiently limit possible candidate genes. Geneticists are faced with a long list of genes from which they need to arbitrarily determine a cut-off value to then assay in case samples. DIG's filter system limits the number of total genes ranked, but without bias towards well-studied genes.

4.1.3 Further advantages of DIG

DIG requires the user to input basic knowledge of their disease as an input. The tool uses the best performing parameters as default settings which is convenient for non-expert users; however, more advanced users are still able to fine tune the p-values at which to look for linkage data. Users are advised to use a less stringent linkage cut-off value for highly heterogenic diseases or for diseases that have decreased penetrance, and therefore have weak gene signals.

Additional annotation such as tissue expression, SIFT and Polyphen scores allow users to critically evaluate the ranked genes. Furthermore, DIG interacts with databases directly, thereby remaining as up to date as possible. Local gene and association mappings will be regularly updated to maintain the accuracy of the system. DIG does not require registration or payment to make use of the tool, emphasising ease of access for the user. Although DIG has been evaluated with complex genetic diseases, this tool can also be applied to Mendelian or heterogenic diseases.

4.2 Relative performance as a new tool

Analysis of the test datasets showed that performance for asthma was worse than that for other test datasets. This is thought to be because asthma is relatively poorly studied compared to breast

cancer, type 2 diabetes and Alzheimer's disease. Further tests of additional poorly-documented diseases may elucidate flaws in the DIG methodology that would need to be addressed.

DIG demonstrates comparable performance when compared to PosMed, but Endeavour remains far more accurate according to TPR30, AUC and response rate. Further investigation shows that Endeavour does not filter any genes whatsoever, but simply ranks all possible candidates. This means response rate could not be anything less than 100% since no training genes would be excluded. TPR30 is also assumed to be inflated due to the vast amount of candidates, and thus much higher number of genes included in the top 30%. As a result, the performance measures for Endeavour are therefore better. Different performance indicators are perhaps needed to do an unbiased assessment of tools that filter candidate genes.

4.3 Novel BPD candidates

Of the 8 novel genes previously not associated to BPD, *HTT* was further investigated. The gene had 6 Pubmed articles linked to it, with a literature score of 1.127e-03. The high sequence score of 0.597 is most likely due to the elongated CAG trinucleotide repeat expansions in exon-1 that are known to cause Huntington's disease (Bano *et al*., 2011).

One study reported an increased prevalence of manic symptoms in pre-symptomatic mutation carriers for Huntington's disease as compared to non-carriers, but these symptoms did not completely fulfil DSM diagnostic criteria for BPD (Julien *et al*., 2007). Perlis *et al*. (2010) found that some individuals diagnosed with BPD also carried incompletely penetrant expanded *HTT* alleles.

The involvement of *HTT* in BPD may be explained by a hypothesis that implicates a decrease in synaptic plasticity as the cause for the disease phenotype. Synaptic plasticity is defined as the variation in the ability of neurons to transmit a response (Hughes, 1958). Differences in the number of neurotransmitters within the synapse and the number of receptors on the post-synaptic membrane change plasticity of the neuron. Receptors are transiently stabilised on the synaptic membrane by interactions with scaffold proteins before being recycled to the intracellular compartment by endocytosis (Gerrow and Triller, 2010). Regulations of receptor-scaffold and scaffold–scaffold interactions are thus able to control synapse function.

Previous BPD studies have found decreased expression of the synapsin family of proteins in the cingulate cortex and hippocampus (Eastwood and Harrison, 2001). These proteins are responsible for binding synaptic vesicles to the cytoskeleton, preventing their transport to the presynaptic membrane and subsequent neurotransmitter release. According to STRING, the putative function of *HTT* is listed as playing a role in microtubule-mediated transport or vesicle function. Should *HTT* have altered functioning or be down-regulated either due to CAG expansions or other mechanisms, similar effects on synaptic function could occur as noted in the above research. *HTT* has been shown to be most highly expressed in the cerebellar cortex, the neocortex, the striatum, and the hippocampal formation in the brain (Sayer *et al*., 2005). Additionally, this study shows that *HTT* interacts with *BNDF* (see figure 3), a major regulator of synaptic transmission and synaptic plasticity (http://www.string-db.org/). This evidence supports the notion that *HTT* may be a novel gene to test for association to BPD.

4.4 Limitations

The methodology behind DIG contains intrinsic errors that may be corrected in future designs. The reliance of webscraping puts the functionality of DIG at risk, should any of the websites it utilises change or be removed. For example, the source for linkage information, aBandApart, is currently not maintained. It was last updated in 2006; however, its use was still implemented for proof of concept in including all linked loci. It should be noted that since approximately 2001, linkage studies have rapidly declined due to the advent of GWAS and later NGS, thus it was assumed that linkage data would not be excessively outdated.

The local file of Prospectr scores used to construct sequence information is also incomplete. A new, updated version was not compiled due to time constraints, but the effect is lessened due to the small weighting usually attributed to the sequence sub-score.

The uneven coverage of the scientific literature, upon which DIG relies heavily, raises concerns about the ability of DIG to rank genes that have yet to be characterized in the scientific literature; however, the results from BPD suggest otherwise. It must also be noted that Gene2pubmed, dbSNP and homolog mappings are not dynamically queried, and will therefore need continual updates and maintenance.

4.5 Future work

DIG has been designed in such a way that additional sources may be easily added to the tool to further increase its data-mining capabilities. Future work will aim to add gene expression data, association data, PPI and functional annotation including GO terms, transcription factor binding sites, pathway and protein domain enrichment to the DIG methodology. Supplementary figure 1 demonstrates how these modules may be added to the DIG workflow. The implementation of search algorithms rather than webscraping is also advised for future revisions. To improve user functionality, the ability to export results from the website and increased speed of querying by parallel processing is also recommended.

4.6 Conclusion

We have introduced a new tool for candidate gene prioritisation, which combines literature, linkage, homolog, sequence and pseudogene data, allowing us to rank genes most likely involved in the aetiology of complex human genetic diseases. Validation with tissue expression and single nucleotide polymorphisms allow users to critically examine the ranked candidates. Performance indicators show that DIG is comparable to existing methods and is able to find novel disease-gene associations. In this study, the Huntingtin gene was shown to be a novel candidate gene to search for association to bipolar disorder.

Data Integrated Genetics has wide application to the genetic research of complex diseases. Results may be followed up by genotyping and sequencing patient samples or even the creation of custom microarrays to evaluate a large number of variants in case and control cohorts. DIG represents an opportunity for geneticists to apply the vast amount of available biological data to a clinical research setting, ultimately facilitating research that may find possible treatment and alleviation of the burden of these diseases.

5 Acknowledgements

I would like to thank the National Research Foundation (NRF) for funding this year of my studies and to the members of the department of Computational Biology at the University of Cape Town for all their assistance. Computations were performed using facilities provided by the University of Cape Town's ICTS High Performance Computing team: http://hpc.uct.ac.za. Special appreciation goes to Mr Ayton Meintjies for his programming mentorship and endless patience; to Ms Reinette Weidemann for her feedback on the manuscript; and to Professor Nicola Mulder for her excellent supervision.

6 References

Adie, E. *et al*. (2005) Speeding disease gene discovery by sequence based candidate prioritization. *BMC Bioinformatics*, 14, 55.

Adie, E. *et al*. (2006) SUSPECTS: Enabling fast and effective prioritization of positional candidates. *Bioinformatics*, 22, 773-4.

Aerts, S. *et al*. (2009) Integrating Computational Biology and Forward Genetics in Drosophila. *PLoS Genet*, 5, e1000351.

Ayme, S. *et al*. (2010) WHO International Classification of Diseases (ICD) Revision Process: incorporating rare diseases into the classification scheme: state of art. *Orphanet J Rare Dis*, 5, P1.

Bano, D. *et al*. (2011) Neurodegenerative processes in Huntington's disease. *Cell Death Dis*, 2, e228.

Bodmer, W. and Bonilla, C. (2008) Common and rare variants in multifactorial susceptibility to common diseases. *Nat Genet*, 40, 695–701.

Bornigen, D. *et al*. (2012) An unbiased evaluation of gene prioritization tools. *Bioinformatics*, 28, 3081-3088

Cardno A, *et al*. (1999) Heritability estimates for psychotic disorders: the Maudsley twin psychosis series. *Arch Gen Psychiatry* ,56, 162-8. DOI:10.1001/archpsyc.56.2.162.

Carlson, P. *et al*. (2006). Neural circuitry and neuroplasticity in mood disorders: Insights for novel therapeutic targets. *NeuroRx*, 3, 22-41.

Chen, J. *et al*. (2009) ToppGene Suite for gene list enrichment analysis. *Nucleic Acids Res*, 37, W305-11.

DSM-IV-TR. (1994) Diagnostic and Statistical Manual of Mental Disorders. Cited 2013 August 24. Available: http://www.psychiatryonline.org/book.aspx?bookid=22.

Eastwood, S. and Harrison, P. (2001) Synaptic pathology in the anterior cingulate cortex in schizophrenia and mood disorders: A review and a Western blot study of synaptophysin, GAP-43 and the complexins. *Brain Res Bull*, 55, 569–578.

Fontaine, J. *et al*. (2011) Genie: literature-based gene prioritization at multi genomic scale. *Nucleic Acids Res*, 39, W455–W461.

George, R. *et al*. (2006) Analysis of protein sequence and interaction data for candidate disease gene prediction. *Nucleic Acids Res*, 34, e130.

Gerrow, K. and Triller, A. (2010) Synaptic stability and plasticity in a floating world. *Curr Opin Neurobiol*, 20, 631-639.

Goes, F. *et al*. (2012) Genome-wide association of mood-incongruent psychotic bipolar disorder. *Translational Psychiatry*, 2, e180.

Greenwood, T. *et al*. (2013) Genome-wide association study of irritable vs. elated mania suggests genetic differences between clinical subtypes of bipolar disorder. *PlosOne*, 8, e53804.

Hughes, R. (1958) Post-tetanic Potentiation. Physiol Rev, 38, 91–113.

Hutz, J. *et al*. (2008) Candid: A Flexible Method for Prioritizing Candidate Genes for Complex Human Traits. *Genet Epidemiol*, 32, 779-90.

Julien, C. *et al*. (2007) Psychiatric disorders in pre- clinical Huntington's disease. *J Neurol Neurosurg Psychiatry*, 78, 939–943.

Kohler, S. *et al*. (2008) Walking the Interactome for Prioritization of Candidate Disease Genes. *Am J Hum Genet*, 82, 949-58.

Lander, E. (1996) The new genomics: global views of biology. *Science*, 274, 536-539.

Lichtenstein, P. *et al*. (2009) Common genetic determinants of schizophrenia and bipolar disorder in Swedish families: a population-based study. *Lancet*, 373, 234-239.

Liu, X. *et al*. (2013) dbNSFP v2.0: A Database of Human Nonsynonymous SNVs and Their Functional Predictions and Annotations. *Hum Mutat*, 34, E2393-E2402.

Meier, S. *et al*. (2012) Genome-wide significant association between a 'negative mood delusions' dimension in bipolar disorder and genetic variation on chromosome 3q26.1. *Translational Psychiatry*, 2, e165.

Merikangas, K. *et al*. (2009) Epidemiology of mental disorders in children and adolescents. *Dialogues Clin*

Neurosci, 11, 7-20."
Nelson, S. *et al*. (2004) The MeSH translation maintenance system: structure, interface design, and implementation. In: Fieschi,M. *et al*. (eds), Proceedings of the 11th World Congress on Medical Informatics. IOS Press, Amsterdam, San Francisco, CA, 67–69.
Oti, M. *et al*. (2011) Web Tools for the Prioritization of Candidate Disease Genes. *Methods Mol Biol*, 760, 189-206.
Perez-Iratxeta, C. *et al*. (2005) G2D: A tool for mining genes associated with disease. *BMC Genet*, 6, 45.
Perlis, R. *et al*. (2010) Prevalence of incompletely penetrant Huntington's disease alleles among individuals with major depressive disorder. *Am J Psychiatry*, 167, 574-9.
Pritchard, J. (2001) Are rare variants responsible for susceptibility to complex diseases? *Am J Hum Genet*, 69, 124–137.
Sayer, J. *et al*. (2005) Interaction of the nuclear matrix protein NAKAP with HypA and huntingtin: implications for nuclear toxicity in Huntington's disease pathogenesis. *Neuromolecular Med*, 7, 297-310.
Sayers E. (2008) E-utilities Quick Start. Entrez Programming Utilities Help. Bethesda (MD): National Center
for Biotechnology Information (US).
Schork, N. *et al*. (2009) Common vs. rare allele hypotheses for complex diseases. *Curr Opin Genet Dev*, 19,
212–219.
Schriml, L. *et al*. (2012) Disease Ontology: a backbone for disease semantic integration. *Nucleic Acids Res*,
40, D940–D946.
Serretti, A. and Mandelli, L. (2008) The genetics of bipolar disorder: Genome ‘hot regions', genes, new potential candidates and future direction. *Mol Psychiatry*, 13, 742-771.
Simmons, D. (2008) The use of animal models in studying genetic disease: transgenesis and induced mutation. *Nature Education*, 1, 1-2.
Sioutos, N. *et al*. (2007) NCI Thesaurus: a semantic model integrating cancer-related clinical and molecular information. *J Biomed Inform*, 40, 30–43.
Smoller, J. and Finn, C. (2003) Family, twin, and adoption studies of bipolar disorder. *Am J Med Genet Part*
C, 1, 48–58.
Sun, J. *et al*. (2009) A multi-dimensional evidence-based candidate gene prioritization approach for complex diseases–schizophrenia as a case. *Bioinformatics*, 25, 2595–2602.
Tiffin, N. *et al*. (2006) Computational disease gene identification: A concert of methods prioritizes type 2 diabetes and obesity candidate genes. *Nucleic Acids Res*, 34, 3067–3081.
Tranchevent, L. *et al*. (2010) A guide to web tools to prioritize candidate genes. *Brief Bioinform*, 12, 22-32.
van Vooren, S. *et al*. (2007) Mapping biomedical concepts onto the human genome by mining literature on chromosomal aberrations. *Nucleic Acids Res*, 35, 2533–2543.
Yoshida, Y. *et al*. (2009) PosMed (Positional Medline): prioritizing genes with an artificial neural network comprising medical documents to accelerate positional cloning. *Nucleic Acids Res*, 37, W147-52.

7 Supplementary material

Supplementary table 1. Training genes used for the evaluation of the DIG variance weight matrix using four complex diseases.

Disease name input	Training genes			
Breast cancer	ENSG00000183765	ENSG00000233209	ENSG00000072571	ENSG00000012048
	ENSG00000093010	ENSG00000206237	ENSG00000174775	ENSG00000139618
	ENSG00000140465	ENSG00000231939	ENSG00000169083	ENSG00000118046
	ENSG00000138061	ENSG00000206302	ENSG00000095015	ENSG00000105329
	ENSG00000148795	ENSG00000179344	ENSG00000149311	ENSG00000141510
	ENSG00000137869	ENSG00000225824	ENSG00000110628	ENSG00000171940
	ENSG00000141736	ENSG00000231286	ENSG00000121879	ENSG00000083093
	ENSG00000142208	ENSG00000236884	ENSG00000176907	ENSG00000124151
	ENSG00000091831	ENSG00000206240	ENSG00000171862	ENSG00000136492
	ENSG00000066468	ENSG00000228080	ENSG00000138376	ENSG00000064012
	ENSG00000100697	ENSG00000196126	ENSG00000051180	ENSG00000170836
	ENSG00000123908	ENSG00000206306	ENSG00000137077	ENSG00000156735
	ENSG00000175324	ENSG00000229074	ENSG00000107562	ENSG00000023287
	ENSG00000184451			
Type 2 diabetes	ENSG00000142330	ENSG00000231939	ENSG00000187486	ENSG00000181856
	ENSG00000120915	ENSG00000206302	ENSG00000166035	ENSG00000006071
	ENSG00000105221	ENSG00000179344	ENSG00000198763	ENSG00000135100
	ENSG00000115159	ENSG00000225824	ENSG00000198786	ENSG00000108753
	ENSG00000236418	ENSG00000231286	ENSG00000162992	ENSG00000148737
	ENSG00000232062	ENSG00000231679	ENSG00000106331	ENSG00000008196
	ENSG00000206305	ENSG00000196101	ENSG00000197594	ENSG00000109501
	ENSG00000196735	ENSG00000230463	ENSG00000132170	ENSG00000132570
	ENSG00000225890	ENSG00000227357	ENSG00000154415	ENSG00000120833
	ENSG00000228284	ENSG00000227826	ENSG00000104918	ENSG00000181092
	ENSG00000233209	ENSG00000231021	ENSG00000196396	ENSG00000121653
	ENSG00000206237	ENSG00000169047	ENSG00000163581	
Asthma	ENSG00000090376	ENSG00000230413	ENSG00000169194	ENSG00000172156
	ENSG00000183134	ENSG00000235346	ENSG00000170745	ENSG00000166888
	ENSG00000164265	ENSG00000206506	ENSG00000187258	ENSG00000149021
	ENSG00000169252	ENSG00000204632	ENSG00000005381	ENSG00000146070
	ENSG00000181019	ENSG00000264751	ENSG00000171195	ENSG00000149451
	ENSG00000233095	ENSG00000150540	ENSG00000136147	ENSG00000121691
	ENSG00000237216	ENSG00000077238	ENSG00000168229	ENSG00000172057
	ENSG00000235680	ENSG00000113302		
Alzheimer's disease	ENSG00000175899	ENSG00000123384	ENSG00000206439	ENSG00000228321
	ENSG00000107331	ENSG00000164867	ENSG00000228978	ENSG00000228849
	ENSG00000010704	ENSG00000122861	ENSG00000204490	ENSG00000232810
	ENSG00000130203	ENSG00000143801	ENSG00000230108	ENSG00000112715
	ENSG00000142192	ENSG00000091513	ENSG00000223952	