META-ANALYSIS OF GENETIC ASSOCIATIONS USING KNOWLEDGE REPRESENTATION

by

Jesus Enrique Herrera-Galeano

A Dissertation
Submitted to the
Graduate Faculty
of
George Mason University
in Partial Fulfillment of
The Requirements for the Degree
of
Doctor of Philosophy
School of Systems Biology

Committee:

_____	Dr. Jeffrey Solka, Dissertation Director
_____	Dr. Iosif Vaisman, Committee Chair
_____	Dr. Patrick M. Gillevet, Committee Member
_____	Dr. David Hirschberg, Committee Member
_____	Dr. James D. Willett, Director, School of Systems Biology
_____	Dr. Timothy L. Born, Associate Dean for Student and Academic Affairs, College of Science
_____	Dr. Vikas Chandhoke, Dean, College of Science
Date: _____	Spring Semester 2013 George Mason University Fairfax, VA

Meta-Analysis of Genetic Associations Using Knowledge Representation

A Dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy at George Mason University

by

Jesus Enrique Herrera-Galeano
Master of Science
Johns Hopkins University, 2005

Director: Jeffrey Solka, Professor
School of Systems Biology

Spring Semester 2013
George Mason University
Fairfax, VA

DEDICATION

This is dedicated to my two wonderful children Kevin and Carolina and to my wife Heather.

ACKNOWLEDGEMENTS

To Kevin, Carolina and my wife Heather from whom I stole some many moments to work on this dissertation.

Many thanks to my advisor Dr. Solka for his excellent advice not only for this dissertation, but in live.

I want to thank the members of my committee for participating in this process for years and for the excellent suggestions and advice.

I want to thank my best friend since I was five, Juan Carlos Gonzalez Perez, not only my best friend, but my brother.

To Josh Moody an unconditional friend that has helped me in very hard times. I thank my mother Margarita Galeano who has always supported me. Thanks to Dr. Diane Becker whose initial support allowed me to start the Ph.D. program. Thanks to Diane St. Germain, her excellent organization and fabulous attitude facilitated finally completing the program. To the members of my new home, NMRC Frederick, and especially Dr. Vish Mokashi for his flexibility and support To Dr. Kim Bishop-Lilly for editing and reviewing the document.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF EQUATIONS

LIST OF ABBREVIATIONS

Candidate gene approach ..... CGA
Database management system ..... DBMS
Directed acyclic graph ..... DAG
Gene ontology ..... GO
Genome wide association studies ..... GWAS
Genotype phenotype database ..... dbGAP
Genotype-phenotype association database ..... GAD
Human phenotype ontology ..... HPO
Infectious disease ontology ..... IDO
International classification of disease version 10 ..... ICD10
Linkage disequilibrium ..... LD
Online mendelian inheritance in man ..... OMIM
Ontology of genetic associations ..... OGA
Polymerized chain reaction ..... PCR
Single nucleotide polymorphisms ..... SNP
Single strand conformational polymorphisms ..... SSCP
Systematized nomenclature of medicine ..... SNOMED

ABSTRACT

META-ANALYSIS OF GENETIC ASSOCIATIONS USING KNOWLEDGE REPRESENTATION

Jesus Enrique Herrera-Galeano, Ph.D.

George Mason University, 2013

Dissertation Director: Dr. Jeffrey Solka

Recent advances in genomic technology have resulted in the availability of an unprecedented amount of genetic data. However, despite the impressive resolution in genetic markers currently available, those who study complex diseases still are haunted by the "missing heritability problem," (Manolio et al., 2009) the problem of not being able to explain a large portion of the expected genetic heritability of a disease. Many efforts are currently being conducted to try to explain a larger portion of the heritability by finding combinations of genes or markers that affect the phenotype of interest. Here, we introduce a methodology to utilize structured knowledge of the phenotypes to find correlations among genes/markers. As a motivating example, we focused on answering questions such as: Is there a common gene related to groups of related phenotypes and is the meta-analysis of associations related by the ontology significant? This work presents the methodology and tools necessary to answer such questions. Here we present a new

application, the ontology of genetic associations (OGA). OGA is completely standalone and allows the user to (1) navigate the phenotype ontology and observe the corresponding gene associations, (2) find the genes common to two or more phenotypes, and (3) find an empirical *p* value to indicate the probability of arriving at the same findings by chance.

CHAPTER ONE: BACKGROUND

Historical context

Phenotypes are the visible manifestations of our genetic make-up, and genotypes are the underlying genetic characteristics that determine those manifestations. The description of phenotypes started within ancient Greece; Greek physicians such as Hippocrates (460-370 BC), Celsus (25 BC-50 AD), and Galen (130-201 AD) started describing physical characteristics of the human body (Groth et al., 2007). Mendel, the father of modern genetics, distinguished between the internal state ("genotype") and the external state ("phenotype") (Groth, et al., 2007). The rules published by Mendel became known as "mendelian inheritance", and they describe a single gene controlling or modifying a specific phenotype. An example of simple mendelian inheritance in humans is Phenylketonuria (PKU) (OMIM: 261600), a metabolic disorder that is the result of the convergence of two defective copies of the gene phenylalanine hydroxylase (PAH). PKU was discovered by Norwegian physician Asbjørn Følling in 1934 (Christ, 2003). PKU is a devastating phenotype when untreated, but the manifestation of the phenotype can be prevented with diet. Another example is Oculocutaneous albinism (OMIM: 203100), which is caused by a mutation on the tyrosinase gene (Chian & Wilgram, 1967). The observation that one gene controlled one phenotype fueled a quest that continues to the current day as researchers continue the search for the correlation or the relationship between genotype and phenotype.

The amount of biological data that has been generated in the past decade is astonishing. In a now classic image (Figure 1), the growth chart of NCBI's genbank, it is evident that the amount of sequencing data available has grown exponentially in recent years. The course of this growth is closely correlated with an increase in developments in DNA sequencing technology.

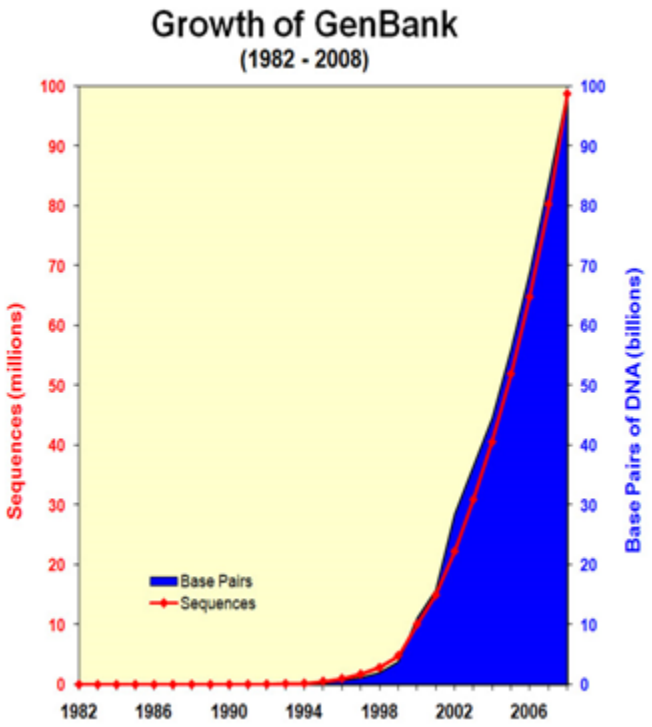

Figure 1 The growth of NCBI's Genbank
(http://www.ncbi.nlm.nih.gov/genbank/genbankstats.html)

Genetic epidemiology

The history of the field of Genetic Epidemiology is also coupled with advances in genotyping and sequencing technology. Improvements in the Polymerise Chain Reaction (PCR) technique in the 1980's (Mullis et al., 1986) allowed the genotyping of Short Tandem Repeats (STRs) (Nakamura et al., 1987). STRs are polymorphic (sometimes involving hundreds of alleles) and because they are not transcribed, STRs are believed to be neutral to selection (Jeffreys, Wilson, & Thein, 1985). These characteristics make them ideal genetic markers. It became possible to design primers for several hundred STRs scattered through the human genome, and, in this way, the first scan of the human genome came into existence (Hearne, Ghosh, & Todd, 1992).

Heritability

The observation of the influence of genes on phenotypes prompts the question of the magnitude of such influence. The measure of the influence of the genetic component in a particular trait can also be thought of as the degree of resemblance between relatives (Falconer, 1993). The total phenotypic variance (Vp) of a particular trait can be caused by several factors; these factors can be divided into the Genetic Variance (Vg) and the Environmental Variance (Ve).

Equation 1 The components of the phenotypic variance

The influence of the genetic variance on the total phenotypic variance is also called the ***heritability*** of the trait (Falconer, 1993). An excellent example of the

estimation of the genetic component of the phenotypic variance in traits of interest is the estimation of the heritability for platelet aggregation traits (Bray et al., 2007).

Table 1 The heritability estimates for platelet aggregation measures

	White subjects			African American subjects		
Assay	h^2	SE	*P*-value*	h^2	SE	*P*-value*
Whole blood platelet aggregation						
ADP 10 μM	0.505	0.095	< 0.001	0.389	0.167	0.020
Collagen 1 μg mL^{-1}	0.014	0.076	0.850	0.017	0.136	0.903
Collagen 5 μg mL^{-1}	0.281	0.090	0.002	0.000	–	–
Arachidonic acid 0.5 mM	0.584	0.089	< 0.001	0.122	0.136	0.373
Thromboxane B_2	0.512	0.098	< 0.001	0.216	0.144	0.134
PFA closure time (epinephrine)	0.237	0.119	0.046	0.609	0.195	0.002
Platelet-rich plasma aggregation						
Collagen 0.5 μg mL^{-1}	0.128	0.157	0.414	0.529	0.239	0.027
Collagen 2 μg mL^{-1}	0.256	0.116	0.027	0.661	0.166	< 0.001
Collagen 5 μg mL^{-1}	0.179	0.102	0.080	0.787	0.201	< 0.001
Lag phase to collagen 2 μg mL^{-1}	0.504	0.107	< 0.001	0.470	0.176	0.008
Arachidonic acid 1.6 mM	0.383	0.101	< 0.001	0.146	0.212	0.490
Epinephrine 2 μM	0.363	0.088	< 0.001	0.758	0.157	< 0.001
Epinephrine 10 μM	0.357	0.097	< 0.001	0.711	0.173	< 0.001
ADP 2 μM	0.419	0.111	< 0.001	0.778	0.164	< 0.001
ADP 10 μM	0.424	0.000	< 0.001	0.729	0.000	< 0.001
Platelet count	0.596	0.098	< 0.001	0.670	0.139	< 0.001
Mean platelet volume	0.553	0.109	< 0.001	0.708	0.183	< 0.001
Hemoglobin	0.374	0.099	< 0.001	0.418	0.190	0.027
White blood cell count	0.422	0.095	< 0.001	0.420	0.166	0.011

ADP, adenosine diphosphate; PFA, platelet function analyzer. *Maximum likelihood estimates of heritability were obtained on logarithm-transformed data for normality, and the likelihood ratio test was used to test significance of heritability comparing a model in which heritability is estimated with one in which it is set to 0.

Table 1 was taken from Bray et al. (Bray, et al., 2007) and shows how the estimates of heritability can be 50% or higher. These estimates indicate that at least 50% of the phenotypic variance should be caused by the genetic component. In other words, it should be possible to find one or more genes that explain at least 50% of the phenotypic variance. However, these estimates refer to the calculation based on putative genetic factors without the identification of a particular gene. The contribution of a particular

gene to the overall phenotypic variance can also be calculated, an example derived from the association of the PEAR1 gene to platelet aggregation factors can be seen in Table 3.

Table 2 Heritability of several complex diseases

Disease	Number of loci	Proportion of heritability explained
Age-related macular degeneration	5	50%
Crohn's disease	32	20%
Systematic lupus erythematosus	6	15%
Type 2 diabetes	18	6%
HDL cholesterol	7	5.2%
Height	40	5%
Early onset myocardial infarction	9	2.8%
Fasting glucose	4	1.5%

Table 2 was taken from (Manolio, et al., 2009), it shows how five loci explain about 50% of the expected heritability for age-related macular degeneration. However, for the next disease in the list, Chron's disease, the portion of the heritability explained is just 20% (Barrett et al., 2008). For the third trait in the list, the portion of heritability explained is just 15% (Todd, 2006). For all others in the list, the portion of heritability explained is less than 6%, as follows: Type 2 diabetes 6% (Zeggini et al., 2008), HDL cholesterol 5.2% (Kathiresan et al., 2008), Height 6% (Weedon & Frayling, 2008), early

myocardial infarction 2.8% (Myocardial Infarction Genetics et al., 2009), and fasting glucose 1.5%

Table 3 The contribution of PEAR1 to the phenotypic variance of platelet aggregation measures

Variance components analysis by ethnicity N=1486

	Whites N=927					African Americans N=559				
	VP	VL	VL/VP	H2	VL/VG	VP	VL	VL/VP	H2	VL/VG
Pre-Aspirin										
Collagen 2µg/mL	856	4.9	0.57	0.21	2.85	1087	2.21	0.2	0.46	0.43
Collagen 5 µg/mL	297	2.9	0.97	0.26	3.7	469	16.1	3.4	0.64	5.3
Epinephrine 2 µM	1121	1.2	0.1	0.23	0.43	1295	1.79	0.13	0.59	0.22
Epinephrine 10 µM	713	5.4	0.75	0.32	2.34	1084	7.9	0.7	0.49	1.42
ADP 10 µM	188	0.49	0.25	0.36	0.69	268	5.55	2.07	0.43	4.8
Post Aspirin										
Collagen 2 µg/mL	167	1.63	0.98	0.47	2.0	252	8.19	3.25	0.59	5.5
Collagen 5 µg/mL	422	5.92	1.40	0.56	2.5	499.8	12.00	2.40	0.53	4.5
Epinephrine 2 µM	154	1.87	1.21	0.50	2.4	205	6.89	3.40	0.49	6.9
Epinephrine 10 µM	217	1.53	0.70	0.43	1.6	264	6.98	2.64	0.57	4.6
ADP 10 µM	165	1.06	0.64	0.51	1.25	183	1.14	0.62	0.37	1.67

VP (phenotypic variance); VL (locus contribution to the variance); VL/VP (proportion of the phenotypic variance explained by the locus × 100). H2 (heritability), VG (genetic variance): VL/VG (proportion of the genetic variance explained by the locus × 100).

Linkage

Linkage Analysis (LA) occupied the time and attention of human geneticists for most of the 1980's and 1990's. These efforts culminated in the one of the better known genetic epidemiology success stories, the discovery and mapping of Breast Cancer genes (OMIM:114480 http://www.ncbi.nlm.nih.gov/omim/114480) .

In 1993, several publications reported the linkage of breast cancer to the long arm of Chromosome 17 (17q). In particular, the efforts of Easton enabled the refinement of

the linkage region to the area between markers D17S588 and D17S250 (Easton, Bishop, Ford, & Crockford, 1993). This genetic mapping would later lead to the discovery of two genes, BRAC1 and BRAC2 (Miki, Katagiri, Kasumi, Yoshimoto, & Nakamura, 1996) (Gayther et al., 1995), and with them possibly the best known story about the identification of genetic predisposition, and direct clinical intervention based only on genetic background (Rebbeck et al., 2004; Zagouri et al., 2013).

The story of success of the mapping of BRAC1/2 is, unfortunately, unique. There are other susceptibility regions that have been identified and have contributed to the understating of the physiopathology of many phenotypes, but few with the impact of BRAC1. Hundreds if not thousands of linkage reports were published over the years addressing the 17q linkage to Breast Cancer, but most of those reports never led to any functional confirmation. The number of examples available of linkage results that have not resulted in meaningful impact on the phenotype are perhaps in the thousands. The following are just two illustrating examples: Alzheimer's disease to chromosome 10q (Bertram et al., 2000) and Type 2 diabetes to Chromosome 2q (Hanis et al., 1996). Janine Altmüller in 2001 best summarized these observations by stating “Positional cloning based on whole-genome screens in complex human disease has proved more difficult than originally had been envisioned…” (Altmuller, Palmer, Fischer, Scherb, & Wjst, 2001).

Calculation of linkage

There are several methods for calculating the linkage of a marker to a polymorphism of interest. The most recognized methods are the Haseman-Elston algorithm (Haseman & Elston, 1972) and the variance components method.

The Haseman-Elston algorithm

The Hesemam-Elston is a simple approach; it is a linear regression of the square difference of the sibpairs against the estimated portion of identity-by-descent (IBD). IBD is opposite to the identity-by-state (IBS). An allele that is shared between a pair of siblings is identical by descent, if they both received the same allele from the same parent. If the alleles shared between two siblings are the same (i.e. they have exactly the same DNA sequence), but the siblings received the allele from different parents (i.e. one received the maternal allele and the other received the paternal copy) these alleles are identical by state.

Equation 2 Haseman-Elston regression

Where:

X= The trait value on the first sibling

Y=The trait value on the second sibling

r= Coefficient of correlation between X and Y

π= The estimated portion of IBS

ε= Random error

Q= The proportion of phenotypic variance explained by the additive effects of the trait

The Haseman-Elston equation as presented above was taken from (Sham & Purcell, 2001)

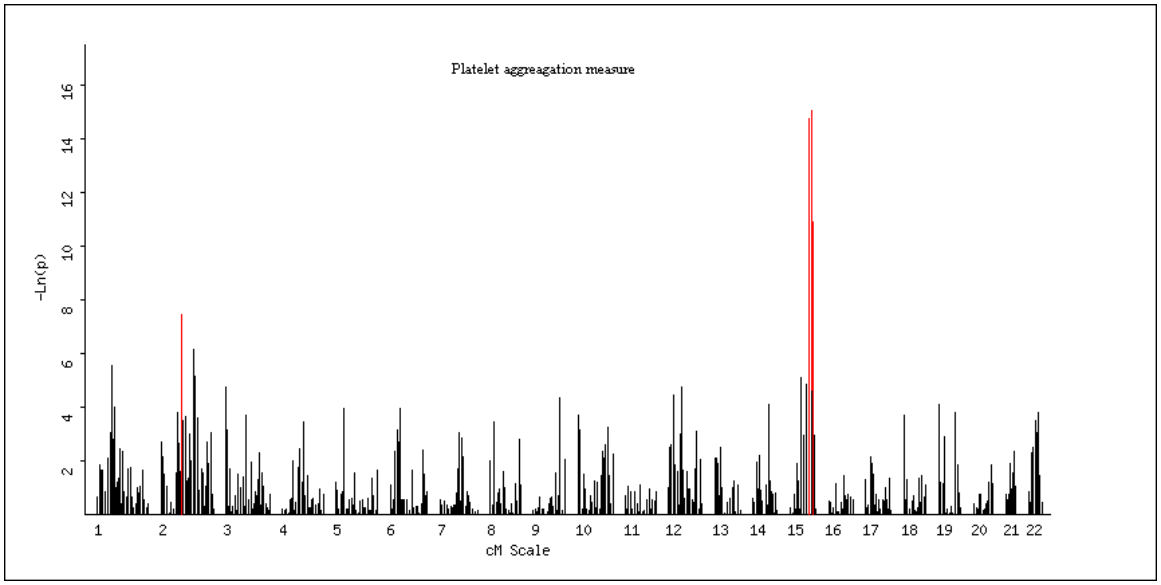

Figure 2 Haseman-Elston linkage results for platelet aggregation

Figure 2 shows the results of the Haseman-Elston regression for a Genome wide scan with 550 STRs performed as part of the GENESTAR study (Mathias et al., 2010).

Variance components

Just as the Haseman-Elston algorithm, the Variance components (VC) approach has gone through many improvements and modifications over the years (Amos, 1994). The main approach utilizes the following general model:

Equation 3 Genetic model on variance components

Where:

X_i = Trait value of the *i*th relative

μ = The overall trait mean

g_i = Unobserved major gene component

G_i = Random polygenic effect

β_k = The covariate effect uncorrelated with the genetic factors

Z_{ik} = The *k*th covariate measurement on the *i*th individual

ε = Random error

Equation 3 was taken from (Amos, 1994).

Candidate genes

The second major advance in genetic epidemiology came about in the mid 1990's, with the availability of sequencing data and the ability to compare sequences from different individuals. This advancement made possible the characterization of single nucleotide differences in the human genome. These changes became known as Single Nucleotide Polymorphisms (SNP) (Altshuler et al., 2000). Here, the approach is a little different from the one used in Linkage Analysis. Instead of trying to "query" the whole genome, the design was centered around identifying susceptibility loci based on previous knowledge of biochemical pathways, physiopathology of the disease, and any other source of information that could help pinpoint the possible location of a genetic causal variant. Around the late 1990's, this approach consisted of identifying small regions and looking for polymorphisms with methodologies such as Single Strand Conformational

Polymorphisms (SSCPs), then genotyping the available cohort with restriction enzymes (Suzuki, Orita, Shiraishi, Hayashi, & Sekiya, 1990).

The early 2000's brought a new technology as well. The introduction of the Illumina® Golden Gate array, for example, allowed the possibility of genotyping thousands of SNPs in a single assay (Cunningham et al., 2008). This development brought about a new strategy in genetic epidemiology, the candidate gene approach (CGA) (Kwon & Goate, 2000). The main idea of the CGA is to perform a hypothesis-based selection of genes. These genes are usually selected based on previous association to a particular phenotype, their role in biochemical pathways related to the phenotype, their function or cellular location, and many other criteria. After the domain experts agree to a set of "candidate genes," the next step would be to select a set of SNPs that "cover" the gene.

SNP selection

The question of what constitutes coverage of a gene by a set of SNPs was addressed by Harris in 2003. In his paper Harris proposes a formal definition of coverage, basically counting the number of times a base is within a radius of a selected SNP (Harris, Martin, Peden, & Rawlings, 2003).

The equation for the formal definition of SNP coverage as proposed by Harris (Harris, et al., 2003) is as follows:

Equation 4 Harris equation for SNP coverage

Where:

P_{ij} = the probability that base i is covered by SNP j

x_j = 1 if SNP j is included in the set of chosen SNPs and

x_j = 0 otherwise

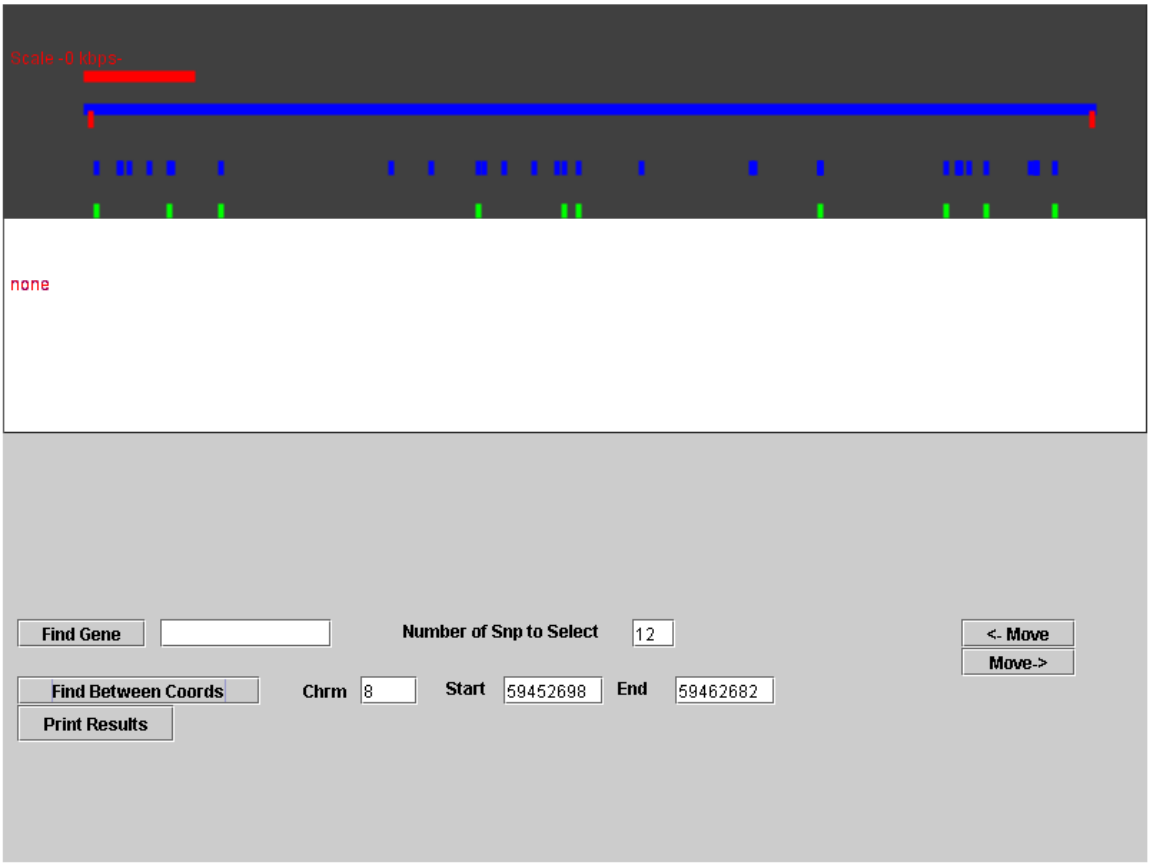

Figure 3 Snp selection application

Figure 3 shows a tool to select SNPs based on minimizing the variance of the distance by using a Metropolis Montecarlo Markov Model (MMCM) and a score including measures such as heterozygosity, Illumina score and others (Herrera-Galeano et al., 2008). The tool retrieves the SNP data automatically from the NCBI webserver. The only input required from the user is the chromosomal coordinates.

The question of how to select a subset of SNPs to best cover the selected genes became known as the "SNP selection problem" and many papers attempting to solve this problem were published. The SNP selection problem was proved to be NP-hard (Liu et al., 2005). The idea of selecting SNPs to cover a set of bases on the gene of interest is based on the expectation that, if a polymorphic SNP is genotyped close to the causal variant (the base change associated with the phenotype of interest), it is very likely that the typed SNP and the causal SNP will be in linkage. In other words, their genotypes are expected to form segments that are transmitted together, called haplotypes. The interesting part is that selecting SNPs that are perfectly spaced (so that the variance of the distance between them approaches the minimum) does not solve the problem, because in humans (as in all Eukaryotes with sexual reproduction) there are hot spots of recombination (Stumpf & Goldstein, 2003) which make some SNP genotype combinations more likely to occur than others. This phenomenon is known as Linkage Disequilibrium (LD).

A measure of LD is the capital D parameter, defined as follows:

Equation 5 Linkage Disequilibrium

Where:
D = Linkage disequilibrium

X_{11}= Frequency of the haplotype formed by allele 1 on locus 1 and allele 1 on locus 2

p_1= Frequency of allele 1 on locus 1

p_2=Frequency of allele 2 on locus 2

Figure 4 Haploview of the ITGB1 gene

Figure 4 shows the graphical representation of the LD for the SNP located in the integrin, beta 1 (fibronectin receptor, beta polypeptide, antigen CD29 includes MDF2, MSK12) or ITGB1 gene. The figure was obtained using the software Haploview (Barrett, 2009).

In order to take advantage of LD for genetic studies, some kind of characterization of the human haplotypes becomes necessary. The Hapmap Project was launched in 2002 to address this issue (Thorisson, Smith, Krishnan, & Stein, 2005).

Throughout the early 2000's, many studies based on the CGA received funding, and many SNP associations involving many phenotypes of interest were published. One out of many possible examples was the association of PEAR1 to platelet aggregation (Herrera-Galeano, et al., 2008). These findings were later corroborated by a meta-analysis of Genome Wide Association Studies (GWAS) which was published in Nature (Johnson et al., 2010).

Genetic association tests

One of the first measures of quality control for genetic association data is the Hardy–Weinberg equilibrium (HWE) test. The HWE principle states that in a population in equilibrium the frequency of the genotypes in the current generation should be predicted by the frequency of the alleles on the previous generation (Balding, 2006). Although a consistent deviation from the HWE on the overall dataset may be an indication of a problem with the genotyping or other experimental problems, the causal variant (or a variant in strong linkage with the causal variant) is not necessarily expected to be on HWE (Balding, 2006).

The genetic association test is relatively simple when the sample consists of unrelated individuals and the trait is quantitative; in this case a linear regression can be used (Lewis & Knight, 2012)

Genome wide association

In the mid-2000's, a technological advancement would take the field in a new direction. Illumina® released their high-density bead chip arrays, which possess the ability to genotype hundreds of thousands of variations for thousands of individuals

(Oliphant, Barker, Stuelpnagel, & Chee, 2002). This time, the technology approached a state in which almost all the possible variations in an individual's genotype could be queried. It is estimated that a SNP occurs every one kbp on average. With three billion bps, it is expected that the human genome will have around three million SNPs. With the Illumina® device, the number of SNPS queried (around 1 million for example in the Illumina® 1M bead chip) approaches the total number of variations. This level of genome interrogation has led to great expectations and overly optimistic predictions. It was expected that by this time we would have entered the “personalized medicine era." Many grandiose comments about GWAS associations were common-place in scientific publications (Du, Xie, Chang, Han, & Cao, 2012; Palotie, Widen, & Ripatti, 2012).

Population stratification

One of the major issues to control for with GWAS studies is population stratification (Balding, 2006). The problem is that the initial distribution of an allele among subpopulations can easily create a situation that can be taken as association. Figure 5 illustrates the issue.

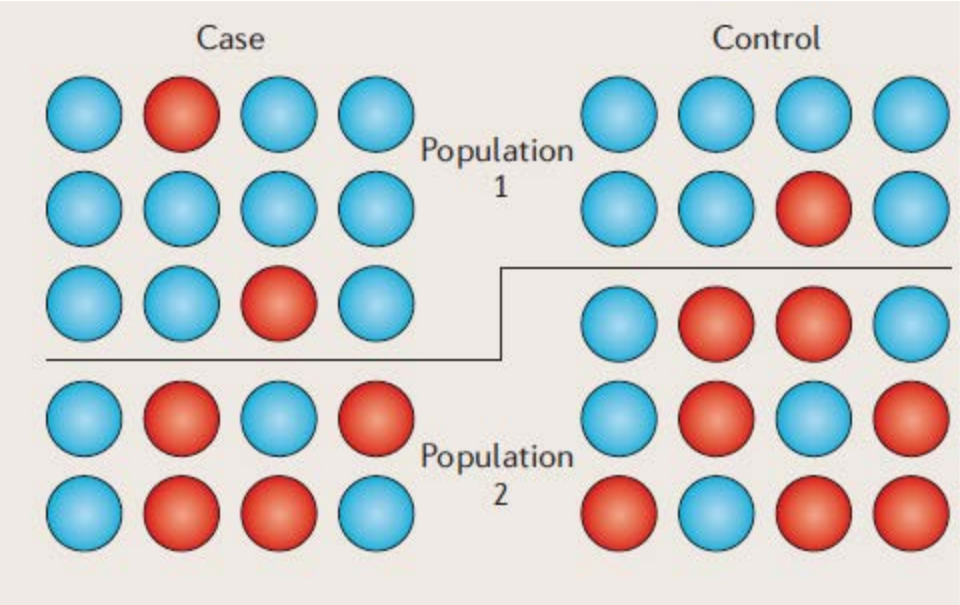

Figure 5 Population stratification

Population stratification may be the cause of many spurious associations reported for GWAS studies. Several methods have been proposed in order to try to control for this issue. One of the most commonly used approaches is Principal Component Analysis (PCA) (Price, Zaitlen, Reich, & Patterson, 2010). This method is based on the idea of finding the main factors in the PCA (for the marker data) and using those factors as way to correct for relatedness in the association test. One of the most popular tools to correct for stratification is EIGENSTRAT (Price et al., 2006). Paradoxically, family studies were at some point considered inferior to the studies using unrelated individuals, because they had to adjust for pedigree dependency; as it turns out that initial correction seems to control for population stratification and now using families can be seen as a strategy to avoid this issue.

Some examples of GWAS success are easy to find. For example, PPARG, a gene that encodes a drug target for diabetes, was found through a GWAS study (Manolio, et al., 2009). However, even as GWAS approaches the interrogation of all the SNPs in the human genome, the accomplishments of this technology again seem to be falling flat. This time the problem is that the "common variant, common effect" premise seems not to have been proven true. This is evident when results such as the association to height are analyzed. Twenty SNPs shown to be associated to height were compiled by Weedon. (Weedon & Frayling, 2008), but the total genetic effect explained by all of these SNPs was a mere 5% (Visscher, 2008), a very disappointing result for a trait such as height, which is widely recognized as having a large heritability estimate. This type of outcome has been repeated over and over. Even in the best case, in studies with very significant associations, the percentage of heritability explained seems not to be higher than 10%. These results do not mean that the associations are not real; it seems rather that most of the heritability must be found in combinations of variants (epistatic effect) that are not necessarily individually associated with the phenotype (Manolio, et al., 2009).

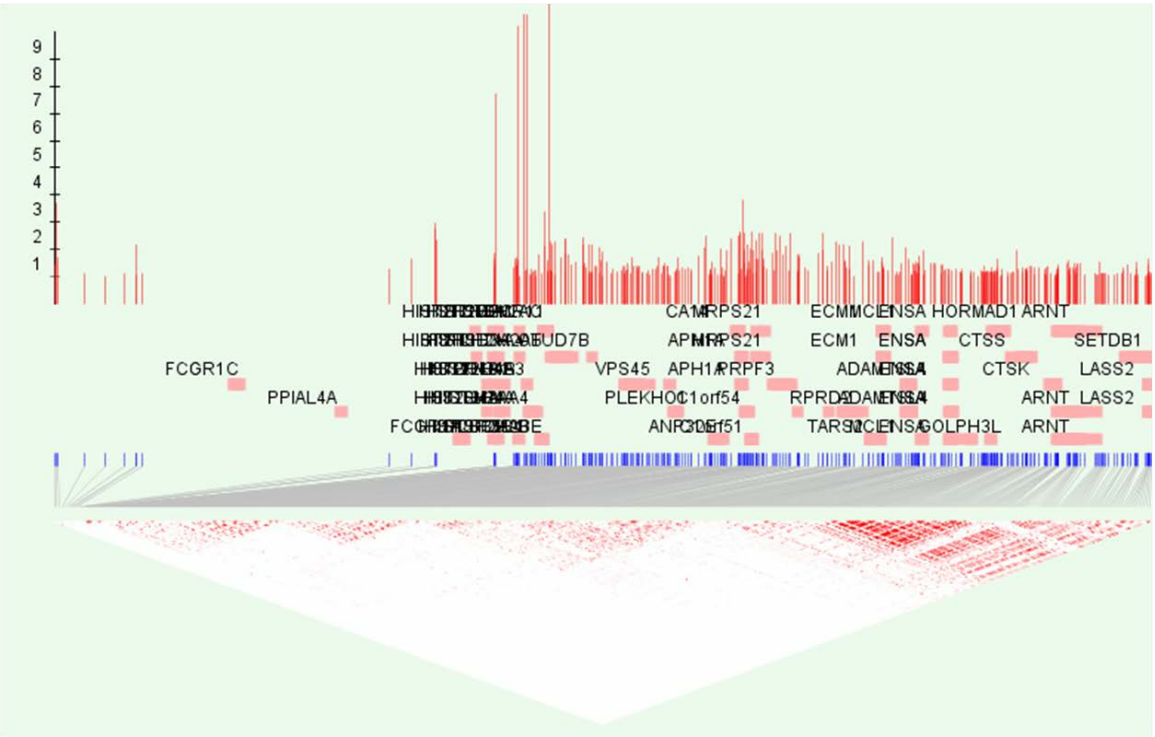

Figure 6 GWAS association results General Well Being

Figure 5 is a graphical representation of the GWAS results for General Well Being (GWB), a measure of mental fitness. The plot is divided in three sections; the top panel shows the negative logarithm of the association test *p* value, the middle panel shows the relative location of the genes in the area, and the bottom panel shows the pairwise LD measure for all the SNPs tested. The plot shows a variant with an association *p* value $< 10^{-10}$, which easily bypasses the correction for multiple testing. However, despite the strong *p* value, the only way to actually validate this finding would be replication. Replication would imply finding the same variant associated with the same trait, which is a major challenge, because most studies focus on a handful of phenotypes and it is rare to find a study that has obtained a measure that matches well with GWB.

Knowing that it was hard to find another study that could match the GWB trait, the study attempted to instead find other traits related to mental fitness that could be associated to the same gene. That pursuit was the basis for this dissertation work; the idea that related phenotypes could be used as method for validation of association findings.

Epistatic combination

It is clear that the field needs a new direction. Some papers are pointing to rare phenotypes, while others are pointing to just a better definition of phenotypes. Currently, some are conducting interrogation of the epistatic combination of SNPs by brute force methods using the cloud and techniques such as map/reduce (Wang, Wang, Tan, Wong, & Agrawal, 2011). Another approach is the combination of random variants in a model with serial evaluation of the contribution of each variant to the phenotype, this pseudo-Bayesian approach follows an updating route resembling of a "simulating annealing". This approach was coined "random handfuls" (Province & Borecki, 2008) . Others have applied machined learning algorithms to the problem. In one approach, the Random Forrest (RF) algorithm is applied first to detect a predictive subset of SNPs, and then the MARS is used to identify the interaction patterns among the selected SNPs (H. Y. Lin et al., 2012).

Ontologies

Independently of the direction that the field takes or the number of computations that will eventually be possible as predicted by Moore's law, the amount of data that has been generated demands a new way to view and analyze the data. A new approach to the massive amount of data will need to allow computer automated exploration of the data

and its internal relationships. This task can be accomplished by converting the available data into structured knowledge. Such an effort is now taking place in Biomedical Informatics with efforts such as Systematized Nomenclature Of Medicine Clinical Terms (SNOMED) (Allones, Taboada, Martinez, Lozano, & Sobrido, 2013), the International Classification of Disease version 10 (ICD-10) (Bedard, Lowry, & Sibbald, 2012), and other more specific efforts such as the Infection Disease Ontology (IDO) (Y. Lin, Xiang, & He, 2011). Another example of a successful effort to utilize structured knowledge in biology is gene ontology (GO) (Ashburner et al., 2000). GO is perhaps the best known ontology among the biomedical community. Many analysis tools have been spawned from GO; one example is the Gene Ontology for Functional Analysis (GOFFA), a tool for the analysis of gene expression experiments (Sun, Fang, Chen, Perkins, & Tong, 2006). GObar (Lee, Katari, & Sachidanandam, 2005), GOrilla (Eden, Navon, Steinfeld, Lipson, & Yakhini, 2009) and GOEAST(Zheng & Wang, 2008) are just a handful of additional examples among the many tools associated with GO.

An ontology is basically defined as "collections of formal, machine-processable and human-interpretable representations of entities, and the relations among those entities" (Musen et al., 2012). Structured knowledge allows the interrogation of the data based on the relationships of otherwise unrelated concepts. A clear indication of the value of ontologies in biomedical research can be seen in the creation of the National Center for Biomedical Ontology (NCBO) (Musen, et al., 2012). NCBO is an interdisciplinary consortium supported by National Center for Biomedical Computing (NCBC), part of the National Institutes of Health (NIH). The mission of NCBO is to

integrate all available biomedical ontologies and make them available to researchers. NCBO currently lists 1,358 ontological sources on its website (http://bioportal.bioontology.org/projects). Among the most prominent projects currently in NCBO are: Gene ontology consortium (GOC) (Ashburner, et al., 2000), the European Molecular Biology Open Software Suite (EMBOSS) (Lamprecht, Naujokat, Margaria, & Steffen, 2011), FlyBase (Tweedie et al., 2009), the Human Phenotype Ontology (HOP) (Robinson & Mundlos, 2010), the Virus Pathogen Resource (ViPR) (Pickett et al., 2012) and many others. In addition to the substantial body of ontological sources, NCBO has also given origin to many invaluable tools for the bioinformatics community. One of these tools is BIOPORTAL, a repository that allows access to the resources mentioned above through web services (Musen et al., 2008; Noy et al., 2009; Whetzel et al., 2011). Another important tool related to NCBO is OBO-EDIT, a tool to navigate and visualize ontologies in OBO format (Day-Richter, Harris, Haendel, Gene Ontology, & Lewis, 2007).

A commonly used example of the relevance of ontologies in biomedical research is the set of questions related to the relationship between pathogens and diseases, including questions such as: "What diseases are caused by retroviruses?" This question can be answered by reviewing the literature about these viruses, but this type of effort is time consuming and a way to answer this question automatically with a computer driven query would be far more desirable. However, obtaining such an automated answer requires a structure representing the hierarchical relationships among viruses, along with the same type of structure for the diseases and a representation of the causal relationships

among the two structures. In the same fashion, ontologies such as the Human Phenotype Ontology (HOP) (Robinson & Mundlos, 2010) allow the examination of data and the generation of hypotheses by automated means.

A common question in genetic epidemiology is as follows: Is there a related phenotype also associated with a particular gene of interest? Normally, the answer to this question implies some tedious research through the literature, but with the availability of the HOP and the recent release of a database of genotype-phenotype associations (GAD) (Becker, Barnes, Bright, & Wang, 2004) (Zhang et al., 2010), such a question can now be answered through automated means. The automation of such tasks will not only save scientists' time, but it also will allow the unveiling of many unexpected relationships among phenotypes, genes, and their associations.

Recently, the interest in integrating phenotype data with the overall genetic analysis has grown significantly. New resources such as the PhenomicDB are continuously been published. PhenomicDB is a cross-species genotype-phenotype database which collects data from different sources such as OMIM, MDG, and others. (Groth, et al., 2007). The existence of this resource demonstrates the interest in connecting phenotype and genotype data. However, the PhenomicDB is based on the clustering of text and the advantage of the ontology is not utilized. There have been multiple efforts to establish a systematic way to look at phenotype-genotype relationships. Chen used Natural Language Processing (NLP) to extract phenotype data from literature (Chen & Friedman, 2004) . In a similar fashion, PhenoGO attempts to

utilize NLP to assign phenotypic context to the Gene Ontology (GO) (Lussier, Borlawsky, Rappaport, Liu, & Friedman, 2006).

All these efforts point to the importance of a systematic way to establish the relationship between phenotype and genotype. In addition, Groth utilized text clustering of phenotype data to predict gene annotation. The result of this study indicates that it is possible to utilize the intrinsic nature of phenotypes to infer new gene function (Groth, Weiss, Pohlenz, & Leser, 2008).

Genotype phenotype associations

The importance of integrating disease data to its genetic makeup has been recognized for several decades. The best example of this observation is several decades of the existence of the Online Mendelian Inheritance in Man (OMIM). OMIM was created by Dr. Victor Mckusick, who is considered the father of medical genetics (McKusick, 2006). OMIM has grown dramatically over the decades just as the field has done the same; it has also been updated and transformed many time over the years (Amberger, Bocchini, Scott, & Hamosh, 2009; Boyadjiev & Jabs, 2000; Hamosh, Scott, Amberger, Bocchini, & McKusick, 2005; Pearson et al., 1994). OMIM continues being one of the most used and respected sources of genetic information in relation to disease (Amberger, Bocchini, & Hamosh, 2011; Baxevanis, 2012).

Phenopedia and Genopedia (PAG) is another important effort in the same realm of applications trying to link disease to its genetics background. PAG is a text mining application that uses machine learning and MeSH terms to identify links between phenotypes and genotypes by examining pubmed (Yu, Clyne, Khoury, & Gwinn, 2010)

Neither OMIM nor PAG are ontologies. However, an ontology based approach was recently published, Neurocarta (Portales-Casamar et al., 2013). Neurocarta is an ontology of disease-gene relationships and it is part of NCBO. The backbone of Neurocarta is formed by the Disease ontology (Schriml et al., 2012), the Mammalian Phenotype ontology (Smith & Eppig, 2012), and HOP. Neurocarta retrieves gene evidence from OMIM, pubmed, the Rat Genome Database (RGD) (Laulederkind et al., 2012) and other more specific sources related to neurological diseases such as SFARI (a database for autism research) (Banerjee-Basu & Packer, 2010). Neurocarta is manually curated and it is geared towards neurodevelopmental disorders. Currently, Neurocarta has 30,000 links between 7,000 genes and 2,000 phenotypes (Portales-Casamar, et al., 2013).

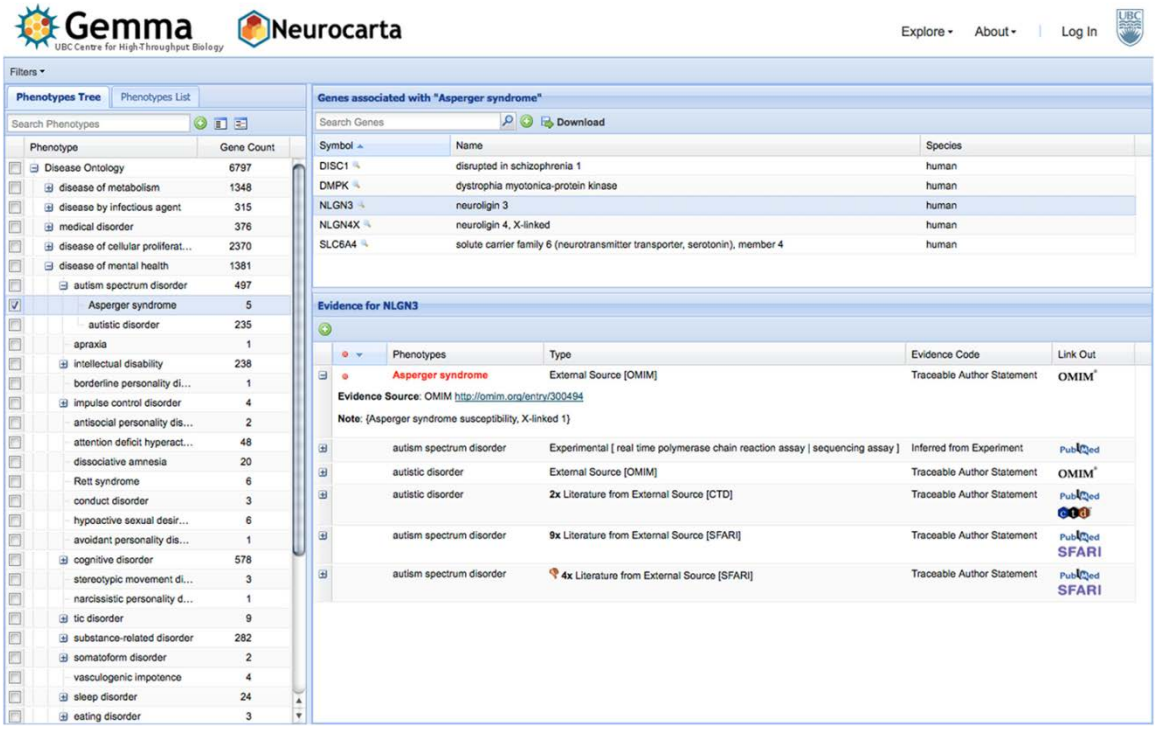

Figure 7 Neurocarta's website

CHAPTER TWO: THE PROBLEM AND HYPOTHESIS

At the time of this writing, Neurocarta is the only publically available application that allows scientists to easily observe the association results related to a central phenotype and easily follow associations to other phenotypes related by structured knowledge. However, Neurocarta is focused on neurological diseases and limits the search space by manual curation. Currently, answering a question similar to the one posted before: “is there a genetic association result for a phenotype related to the central phenotype of interest?” requires tedious examination of the scientific literature or repetitive examination of association databases. PhenoGo is an application that tries to solve a similar problem. However, it uses NLP to assign phenotypic data to GO. In that sense, the relationships available in the ontology are those of the genes and not those of the phenotypes. Here, we propose the utilization of HOP, which will allow the utilization of phenotype relationships.

In addition to allowing the scientific community access to the aforementioned types of queries, the proposed work will also allow researchers to find sets of genes common to clusters of phenotypes. A third possibility that will be facilitated by the proposed application is the combination of *p* values from existing association results for closely related phenotypes. This is equivalent to a meta-analysis of association tests. As a simplified example, consider Diabetes Type I and Diabetes Type II. They are not the

same disease and meta-analysis of associations related to these two diseases may be thought as potentially unproductive. But, indeed, a significant association of results pertinent to both diseases could prove important to the physio-pathology of Diabetes as the root disease. Notice that the *a priori* generation of hypotheses for each root disease is unnecessary given that the proposed application will perform this type of analysis automatically, hence providing hypothesis-free analysis of the data.

This type of analysis could also help bypass the multiple-testing problem, given that weak or borderline significant results could be combined for a common root phenotype and therefore provide significant results for associations that would not be discovered when examined for each phenotype individually. In addition, interesting results obtained through the GAD and HOP mapping could be validated by obtaining raw data for interesting phenotypes from the phenotype genotype database (dbGAP) (Mailman et al., 2007; Walker, Starks, West, & Fullerton, 2011). This could also constitute yet another use for this application, a way to prioritize phenotypes in dbGAP for additional analysis and/or meta-analysis.

The first requirement to make this work possible is the existence of a suitable ontology. Such a requirement is fulfilled by the Human Phenotype Ontology (Robinson et al,2010). The second requirement is the availability of the phenotype and genotype association data. This requirement is fulfilled by the mission of GAD. The first concrete effort in this work will be to "map" the GAD phenotypes. This task could be very cumbersome if approached manually. Instead, we propose the use of overlapping pattern matching sets. In this approach, we take all the words composing the GAD description

and find all the sets matching those words individually in the concepts in HOP. The intersection between the matching sets will be considered the matching association to the HOP concepts.

The ontological mapping, the linking of association results in GAD with the ontological concepts in HOP, would be an important contribution to the scientific community. This effort would open the door to a new class of analysis like the one proposed here, and we anticipate that this work could be welcomed by known journals for publication.

The working hypothesis of this dissertation is that the examination of phenotype relationships in an ontology that holds association data for SNP/Genes will allow the identification of groups of phenotypes, i.e. phenoclusters, which will be linked to a group of genes associated with a large percentage of the phenotypes. In other words, closely related phenotypes will be associated with the same genes**.** If this hypothesis is proven correct, this will show that a group of genes is responsible for the underlying physiopathology that creates phenotypes of the same type or common to the same root, and at the same time point to other genes that may be the modifiers or the ones responsible for the differences between the phenotypes. A classification of this type could be of great importance for prioritization of pharmaceutical targets and to the understanding of what makes up a phenotype in relation to the underlying biological processes.

To establish if a group of genes associated to a phenocluster is found by chance, we propose the calculation of an empirical *p* value for each group of genes found under a particular ontological concept or selected root. To produce an empirical *p* value, the matches between HOP and GAD are permuted randomly. Thereafter, the number of genes associated to a root concept is counted and the permutation is repeated until the number of genes found is equal or greater to the number found with the actual matching. This process generates the number of trials necessary to find the same number of genes associated with a phenocluster by chance, i.e. an empirical *p* value.

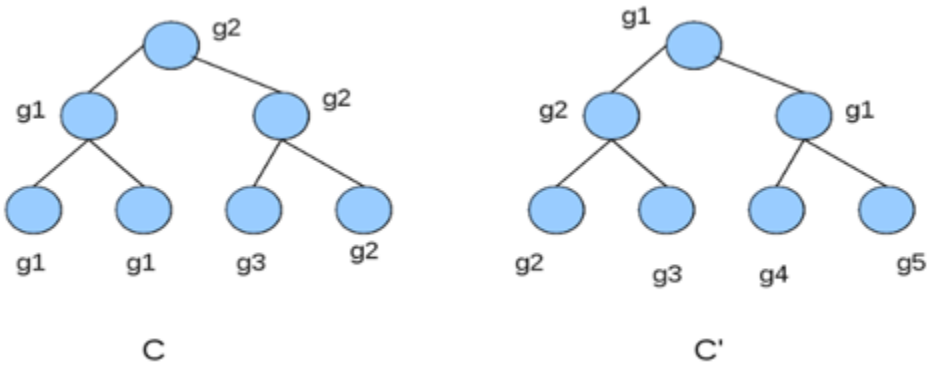

Figure 8 Depiction of a random permutation of gene labels
Let each node in this graphical representation be an ontological concept and each marking next to it a gene associated with the concept using GAD. The left graph would represent the matching between concept and gene actually found through the pattern matching. The one on the right represents a random matching obtained by permuting the links between concept and gene. The empirical p value is obtained by counting the number of C' configurations that need to be obtained in order to find a link count equal to the one in C. More concretely, the actual matching (C, on the left) has three concepts associated with g1, but the permutation (C', on the right) has only two concepts associated with g1. This means that this permutation is added to the count, but the simulation must continue, and a new permutation must be generated. This process continues until a permutation (C') has three concepts associated with g1; this means the same configuration has been found and the simulation can stop.

Equation 6 Empirical p value

CHAPTER THREE: METHODS

The data sources

The Human Phenotype Ontology is publicly available and was downloaded from (http://www.human-phenotype-ontology.org/index.php/downloads.html) (Robinson & Mundlos, 2010). The Genetic Association Database was downloaded from (http://geneticassociationdb.nih.gov/all.xls.zip) (Zhang, et al., 2010). We used the OBO-Edit to navigate the ontology (http://www.oboedit.org/) and to produce the ontology images presented in this work (Day-Richter, et al., 2007).

An OBO-format parser was implemented to parse the ontology and to produce a table that could be placed into a relational database. We focused on conserving the IS-A relationship to allow the traversal of the ontology. The ontology is represented as a Directed Acyclic Graph (DAG) in order to allow for multiple inheritance of the concepts. GAP was also parsed and a minimal representation of the data was extracted to create a table in the database. All scripts were written in PERL and the DBMS of choice was SQLITE. The graphical user interface (GUI) was named "Ontology of Genetic Associations" (OGA) and was fully implemented in JAVA using the SWING toolkit.

The database

The final database consists of four tables; the entity relationship diagram can be seen below.

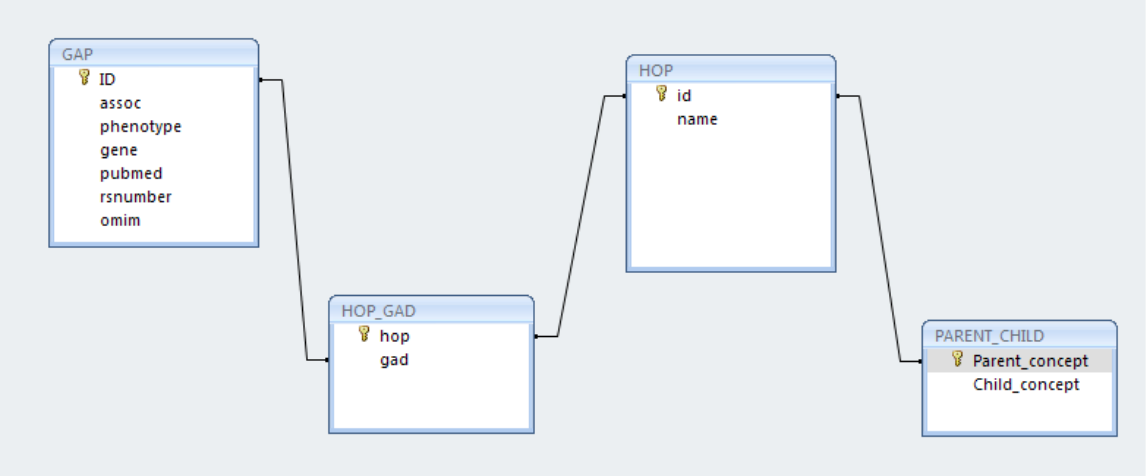

Figure 9 The entity relationship diagram for OGA

The table GAP is a summary of the GAD dataset; the HOP_GAD table stores the links or matching between GAD and HOP. The HOP table is simply a list of ids of names extracted from HOP and the PARENT_CHILD table is the backbone of the ontology, this table allows navigation in OGA.

Dependencies

The application is fully implemented in JAVA. Everything required to run the application is included in the download archive. The only dependency that the application has is the Java Run Environment (JRE) version 1.6 or more recent.

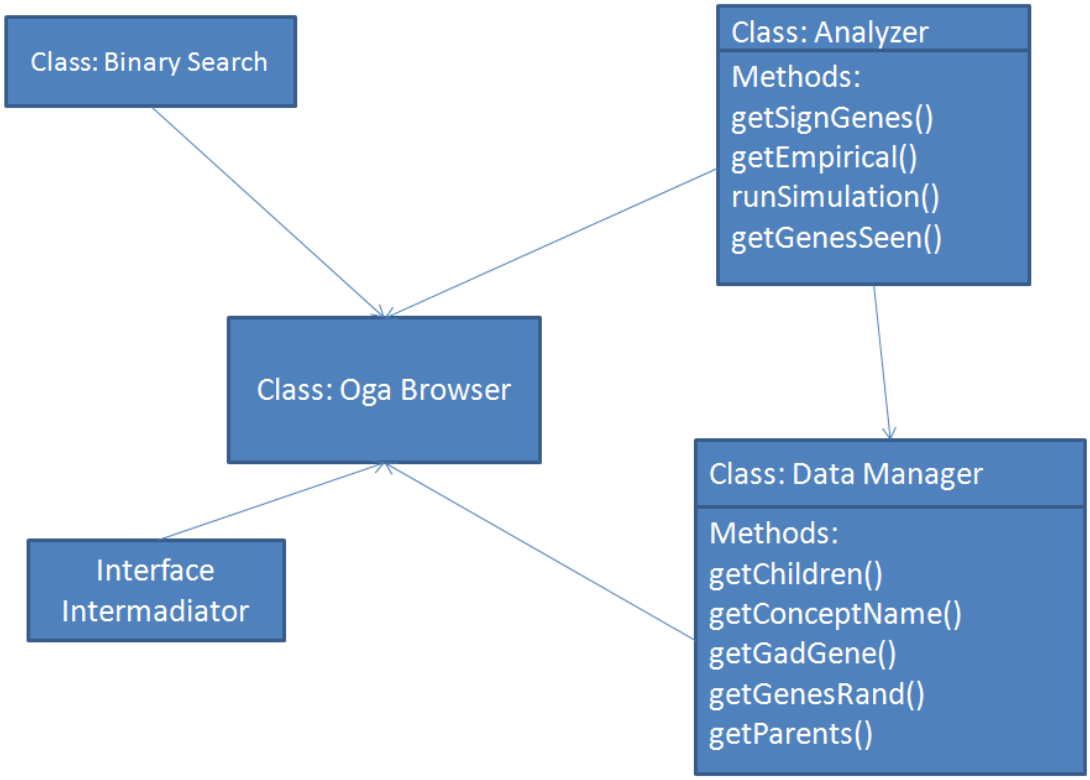

Figure 10 Simplified UML class diagram for OGA

Recursive methods

OGA takes advantage of the ontology by comparing not only a particular phenotype to another, but the complete lineage under each phenotype selected. In order to implement this functionality several recursive methods were implemented. All methods that are related to data are implemented in the class “DataManager.java”; this class implements the method “getGadsLineage.” This method is a recursion that traverses the GAD and collects the list of GAD associations related to each concept.

The simulation

Several iterations of the simulation are performed in order to calculate the empirical *p* value. At each iteration, the intersection of gene sets is calculated (a gene set is the group of genes associated with a particular phenotype lineage). The intersection calculation during the simulation very much resembles the actual intersection calculation for the regular OGA analysis; the only difference is that genes are drawn randomly rather than following the actual gene-phenotype links. The simulation is implemented in the class "Analyzer.java" and the method getGenesRand() implanted in the class "DataManager.java". The method is implemented by utilizing a SQLite query: "SELECT Gene FROM GAP ORDER BY RAND limit gene_set_size".

CHAPTER FOUR: PRELIMINARY RESULTS

First, we performed a non-systematic placing of the GAD associations on the corresponding concept in the ontology. By a "non-systematic placing" we mean that the matching between ontology concepts and associations was accomplished one at a time by performing individual queries in both databases, GAD and HOP. We arbitrarily selected a phenotype root (HOP ID HP:0000708, Behavioural/Psychiatric abnormality) and found three arbitrary children of the selected root concept, Depression (HP:0000716), Schizophrenia (HP:0100753), and Bulimia (HP:0100739). We found 674 genes associated with the concept "Schizophrenia" in GAD, 147 genes associated with "Depression," and 16 genes associated with Bulimia (Figure 5). The interesting part of these findings is that 12 of the 16 genes found in Bulimia are shared by Schizophrenia and 93 of the genes associated with Schizophrenia are also associated with Depression. Furthermore, 9 genes are actually associated with all three phenotypes: BDNF, CLOCK , CNR1, GHRL, HTR1B, HTR2A, HTR2C, SLC6A4, TPH1. These findings present a very plausible set of genes. HTR2A is 5-hydroxytryptamine (serotonin) receptor 2A, it would be easy to explain the role of a serotonin receptor in these disorders. CLOCK, however, is a less studied gene involved in the circadian cycle; its commonality among all these disorders therefore presents a more interesting finding, *a posteriori* there seems to be some support for this observation (Tortorella, Monteleone, Martiadis, Perris, &

Maj, 2007). Another interesting finding is CNR1, the cannabinoid receptor 1, this gene is also associated with all three disorders.

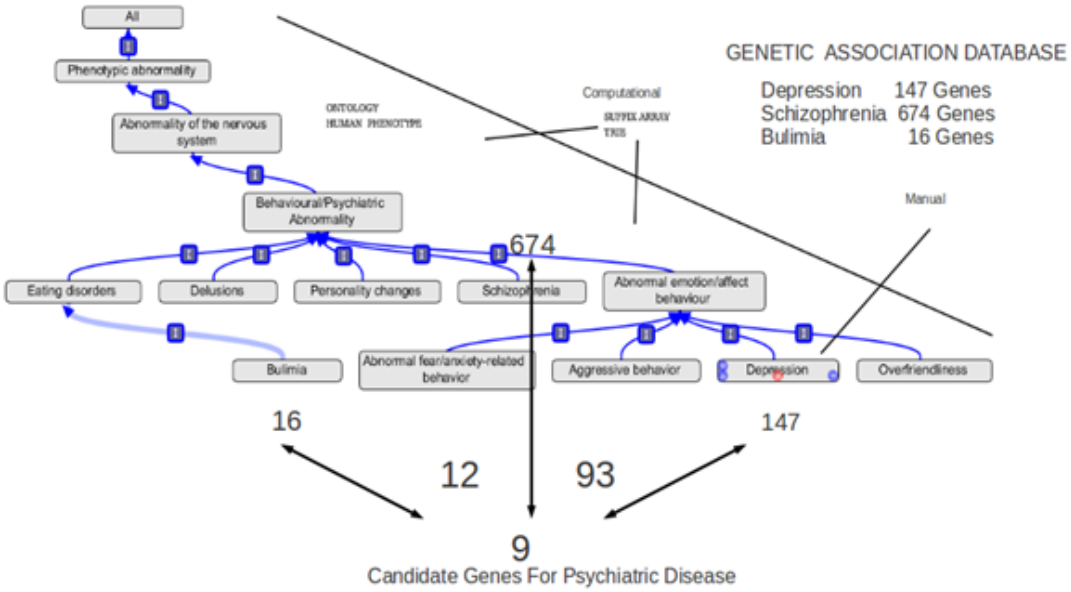

Figure 11 OBO-edit representation of the HOP focusing on the Psychiatric abnormality path

The results shown above indicate that it is indeed possible that phenoclusters are controlled by a set of genes. However, the results shown so far were not obtained systematically, and it is very hard to place any measure of certainty on these observations. In order to provide a systematic way to test the hypothesis posted above and to provide a way in which a measure of certainty can be placed on the observations, we first performed the aforementioned pattern matching approach. We obtained 23,303 unique matches of GAP associations with HOP concepts. There are 84,558 entries on GAP, which is a placement success of 27.5%. The percentage of placement could be

significantly improved by measures such as the use of synonyms for the HOP terms, a set of appropriate stopping words and manual curation. Although, the percentage of matching was not high, we decided to test the phenocluster related to Behavioural/Psychiatric abnormality to determine whether there were any sets of genes related to several phenotypes beyond what is expected by chance. Table 4 shows the sets of genes obtained by counting the number of phenotypes associated with each gene among the Psychiatric abnormality concept.

Table 4 Phenotype counts per each gene found associated with psychiatric abnormality.

Gene	Phenotype Count
SLC6A4	20
NOS1	16
HLA-A	13
APOE	11
HLA-DRB1	10
NOS2A	10
TOR1A	10
TOR1B	10
BCHE	9
CCL2	9
SERPINI1	9
VLDLR	9
HTR2A	8

MAOA	8
BDNF	7
CACNA1A	7
DRD3	6
DRD4	6
Intergenic	6
NOS3	6
TBX22	6

To assign a *p* value to the phenotype counts per gene, we obtained random permutations of the matches between GAP and HOP. For each permutation, we counted the number of phenotypes in the cluster associated with each gene. We had to perform 178 simulations in order to find a permutation that randomly found 20 or more phenotypes associated with a particular gene. This provided an empirical *p* value = 0.0056, indicating that the matching of 20 phenotypes for the gene SLC6A4 (Solute carrier family 6 (neurotransmitter transporter, serotonin), member 4) is not random. It is important to highlight that SLC6A4 is among those genes found in the first non-systematic trial that was performed with only three phenotypes. That shows that these genes may have a central role in the physio-pathology of psychiatric phenotypes. This is not hard to explain given the SLC6A4 is a carrier for the neurotransmitter serotonin (Ho et al., 2012; Kang et al., 2013; Lohoff, Narasimhan, & Rickels, 2012). Serotonin is a neurotransmitter known to participate in the regulation of mood, appetite, and sleep.

Serotonin has also been shown to regulate the development of areas of the brain related to emotional processing (Nordquist, 2010). SLC6A4 is one of the main molecules involved in the regulation of serotonin and many polymorphisms on this gene have been associated with mental disorders. Additionally, the other main regulator of serotonin is monoamine oxidase A (MAOA), the key enzyme responsible for the degradation of serotonin. MAOA was also among the list of genes found, but this gene was not associated with enough phenotypes to be considered statistically significant.

In addition, it took 41 permutations to find 16 phenotypes matched to a gene, such as NOS1 (nitric oxide synthase 1 (neuronal)). This corresponds to a p value = 0.02, which also indicates that the relationship of NOS1 to this phenocluster is statistically significant. All other counts resulted in p values > 0.05, i.e. are expected to simply be a result of chance.

NOS1 is a reactive free radical that displays the characteristics of a neurotransmitter in the brain and the peripheral nervous system. The presence of this gene in this list along with other nitric oxidase genes and with the MAOA gene points to oxidation as the common point among these genes and may indicate a potential point of interest to nominate as a pharmaceutical target.

It would be very interesting to examine the results of an association test that includes all these genes: SLC6A4, MAOA, NOS1, NOS2A and NOS3. It is possible that epistatic relationships exist among these genes.

In addition, when some of the information available for these genes is examined more closely, it is easy to establish a network based on prior information.

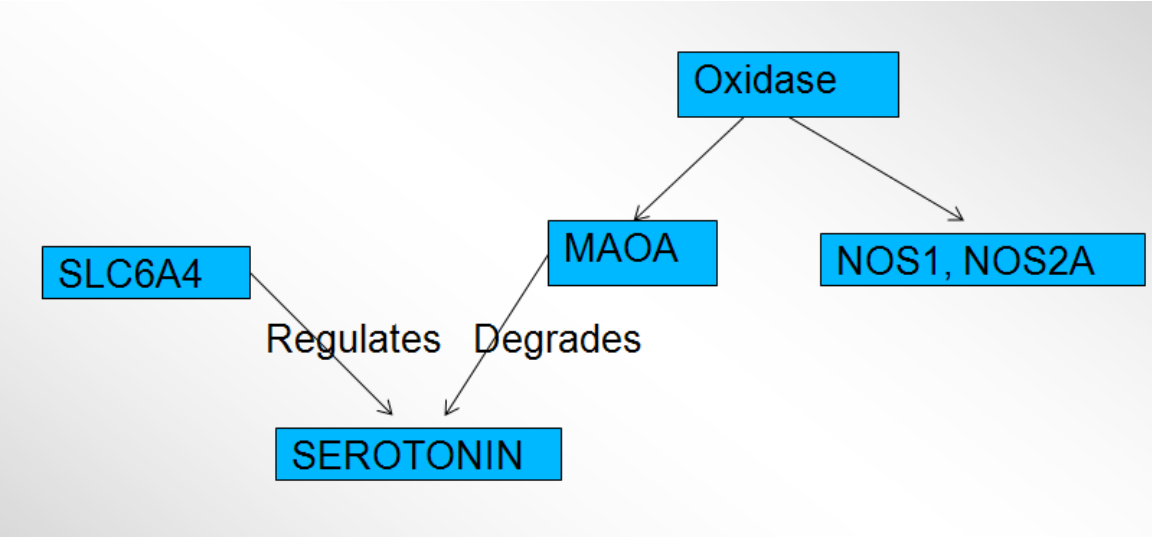

Figure 12 Information network for genes related to mental disorders

Figure 11 shows how the SLC6A4 regulates serotonin, but another gene in the list, MAOA degrades serotonin. NOS1 and NOS2A could participate in the same pathway by the oxidation of MAOA.

CHAPTER FIVE: RESULTS

Objectives accomplished

6.1 The improvement of the pattern matching was accomplished. A stopping word list was compiled and a PERL script was written in order to remove the stopping words from the HOP concepts (code addendum 1: stop_word_remover.pl). Both texts were converted to lowercase and all non alphanumeric characters were removed. In addition, a pattern matching algorithm was implemented specifically to improve the matching. The binary search algorithm (code addendum 2: binarysearch.h, binarysearch.cc, hop_gad_matcher.cc) was implemented to facilitate the search of a suffix array. Therefore, rather than the overlapping result sets approach used during the preliminary results phase, the suffix array allowed for exact match of each GAD term with the HOP concepts. First, matching allowing only one word in common was attempted; this resulted in a percentage of assignment quite higher than the one obtained during the preliminary phase (41.1% vs. 27.5%). The error rate was calculated next. The error rate was calculated by taking 1,000 matches randomly (a sample of about 1.1%) and assessing their relationship one at the time; the observed error rate in five samplings had an average of 30% error rate. The observed error rate was clearly too high, therefore an attempt was conducted on the other side of the stringency spectrum; this time the matching was attempted ensuring that an exact match was found including the complete

text of the GAD entry. In other words, all words in the GAD entry had to match the HOP concept exactly. Utilizing this approach the percentage of assignment was much lower only 19%. However, the error rate was calculated for this approach as well. The error rate was calculated in a similar manner as before; five random samples of size 1,000 were drawn. The observed error rate for this more stringent approach was 2% on average. It is quite difficult to establish how many links really exist between HOP and GAD. It is possible that the high percentage of assignment found at first was due simply to spurious matches as indicated by the high error rate. The 2% error rate in the high stringency approach seems acceptable. Therefore, that set was used to populate the database.

6.2 The proposal called for the implementation of a web application to browse the ontology. However, in order to avoid all the problems associated with hosting, a standalone application was implemented in JAVA-SWING. Everything required to run the application can be downloaded in a single archive file. The application was named OGA (Ontology of Genetic Associations) (code addendum 3: OGA, Analyzer.java, BinarySearch.java, DataManager.java, Intermediator.java, OgaBrowser.java, OgaMainFrame.java, SearchDialog.java).

6.3 The calculation of the empirical *p* values: OGA calculates the empirical *p* values in a manner that is significantly more efficient that the one described during the preliminary work. OGA takes all the concepts that have been selected for analysis and draws random samples of gene sets with sizes equal to the number of genes associated to

each concept. Thereafter, the intersection of the gene sets is calculated, therefore returning only those genes that are common to all the concepts selected. To generate the empirical *p* value, this process is repeated until a gene count is found that is equal or greater to the number genes found without random sampling.

6.4 A paper with the title "OGA Ontology of Genetic associations" has been prepared and is now in the submission process for the journal Bioinformatics.

Table 5 List of stopping words used for OGA

a, about, above, across, after, again, against, all, almost, alone, along, already, also, although, always, among, an, and, another, any, anybody, anyone, anything, anywhere, are, area, areas, around, as, ask, asked, asking, asks, at, away, back, backed, backing, backs, be, became, because, become, becomes, been, beta, before, began, behind, being, beings, best, better, between, big, both, but, by, came, can, cannot, case, cases, certain, certainly, clear, clearly, come, could, did, differ, different, differently, disease, do, does, done, down, down, downed, downing, downs, during, each, early, either, end, ended, ending, ends, enough, even, evenly, ever, every, everybody, everyone, everything, everywhere, face, faces, fact, facts, far, felt, few, find, finds, first, for, four, from, full, fully, further, furthered, furthering, furthers, gave, general, generally, get, gets, give, given, gives, go, going, good, goods, got, great, greater, greatest, group, grouped, grouping, groups, had, has, have, having, he, her, here, herself, high, high, high, higher, highest, him, himself, his, how, however, i, ii, if, important, in, interest, interested, interesting, interests, into, is, it, its, itself, just, keep, keeps, kind, knew, know, known, knows, large, largely, last, later, latest, least, level, less, let, lets, like,likely, long, longer, longest, made, make, making, man, many, may, me, member, members, men, might, more, most, mostly, mr, mrs, much, must, my, myself, necessary, need, needed, needing, needs, never, new, new, newer, newest, next, no, nobody, non, noone, not, nothing, now, nowhere, number, numbers, of, off, often, old, older, oldest, on, once, one, only, open, opened, opening, opens, or, order, ordered, ordering, orders, other, others, our, out, over, part, parted, parting, parts, per, perhaps, place, places, point, pointed, pointing, points, possible, present, presented, presenting, presents, problem, problems, put, puts, quite, rather, really, right, right, room, rooms, said, same, saw, say, says, second, seconds, see, seem, seemed, seeming, seems, sees, several, shall, she, should, show, showed, showing, shows, side, sides, since, small, smaller, smallest, so, some, somebody, someone, something, somewhere, state, states, still, still, such, sure, take, taken, than, that, the, their, them,

```
then, there, therefore, these, they, thing, things, think, thinks,
this, those, though, thought, thoughts, three, through, thus, to,
today, together, too, took, toward, turn, turned, turning, turns, two,
under, until, up, upon, us, use, used, uses, very, want, wanted,
wanting, wants, was, way, ways, we, well, wells, went, were, what,
when, where, whether, which, while, who, whole, whose, why, will, with,
within, without, work, worked, working, works, would, year, years, yet,
you, young, younger, youngest, your, yours, syndrome, syndromes,
diseases, type
```

Table 6 Evaluation of pattern matching by the error rate calculation

HOP Id	Hop name	Gad id	Gad name	Error rate
HP:0001723	restrictive cardiomyopathy	153219	Cardiomyopathy	
HP:0008584	progressive high-frequency hearing loss	562292	Hearing Loss	
HP:0005700	increased bone density with cystic changes	145209	bone density	
HP:0008619	bilateral sensorineural hearing impairment	577559	Hearing Loss	
HP:0003233	decreased circulating high-density lipoprotein cholesterol	129735	cholesterol; coronary heart disease; lipoproteins	
HP:0002642	arteriovenous fistulas of celiac and mesenteric vessels	129360	celiac disease; Wegener's granulomatosis; cervical cancer	
HP:0002511	alzheimer disease	114520	Alzheimer's Disease	
HP:0004780	hypertrichosis limited to elbows	136120	Hypertension	1
HP:0005145	coronary artery stenosis	557620	Coronary Artery Disease|	
HP:0002640	hypertension associated with pheochromocytoma	133671	Hypertension	
HP:0005152	oncocytic	120170	Cardiomyopathy	

	cardiomyopathy			
HP:0000857	neonatal insulin-dependent diabetes mellitus	134973	insulin; smoking behavior; obesity	
HP:0001956	truncal obesity	579048	Obesity	
HP:0002511	alzheimer disease	569822	Alzheimer's disease	
HP:0001513	obesity	562180	Obesity	
HP:0000365	hearing impairment	572720	Hearing Loss	
HP:0001717	coronary artery calcification	126307	coronary artery disease	
HP:0001409	portal hypertension	574147	Hypertension	
HP:0001626	abnormality of the cardiovascular system	116698	Cardiovascular	
HP:0002625	deep venous thrombosis	575321	Thrombosis	
HP:0006510	chronic obstructive pulmonary disease	558301	chronic obstructive pulmonary disease	
HP:0001875	neutropenia	576837	Neutropenia	
HP:0011131	perianal rash	151892	perianal disease	
HP:0004348	abnormality of bone mineral density	122409	bone mineral density	
HP:0100019	cortical cataract	121334	Cataract	
HP:0004941	extrahepatic portal hypertension	116211	hypertension; blood pressure	
HP:0100753	schizophrenia	153627	Schizophrenia	
HP:0004349	reduced bone mineral density	582587	Bone Mineral Density	
HP:0004956	systolic hypertension			
HP:0002278	staring episodes during seizures	123733	seizures	
HP:0006869	myoclonic epilepsy			
HP:0001535	poor weight gain	566192	Weight Gain	
HP:0000875	episodic hypertension	120203	hypertension	
HP:0000717	autism	590215	Autism	
HP:0005686	patchy osteosclerosis	593497	Bone Mineral Density	1
HP:0003362	increased circulating very-low-density lipoprotein cholesterol	596890	LDL cholesterol	
HP:0007213	late-onset form of familial alzheimer disease	138070	Alzheimer's Disease	
HP:0007213	late-onset form of	117234	Alzheimer's Disease	

	familial alzheimer disease		
HP:0006625	breast cancer		
HP:0000608	macular degeneration	579388	Macular Degeneration
HP:0007213	late-onset form of familial alzheimer disease	118444	Alzheimer's disease; Parkinson's disease
HP:0008014	central fundal arteriolar microaneurysms	151221	aneurysm
HP:0004956	systolic hypertension		
HP:0000083	renal failure	576783	Renal Insufficiency
HP:0008332	mild hypertriglyceridemia	138445	triglycerides
HP:0008283	hyperinsulinemia		
HP:0003512	adult female height 130-157 cm	598716	height
HP:0003002	breast carcinoma	149526	breast cancer
HP:0008603	congenital severe sensorineural hearing impairment	570300	Hearing Loss
HP:0001824	weight loss	132672	weight loss
HP:0000098	tall stature	599291	height
HP:0004761	post-angioplasty coronary artery restenosis	560066	Coronary Disease
HP:0005138	coronary artery disease presenting after age 30 years in heterozygotes		
HP:0008028	cystoid macular degeneration	145420	macular degeneration
HP:0005764	polyarticular arthritis	136969	arthritis
HP:0008071	maternal hypertension	145965	hypertension
HP:0008619	bilateral sensorineural hearing impairment	577567	Hearing Loss
HP:0008617	progressive bilateral sensorineural hearing loss	570800	Hearing Loss
HP:0001409	portal hypertension	124052	hypertension
HP:0006755	cutaneous leiomyosarcoma	131624	EO
HP:0007213	late-onset form of	114478	Alzheimer's Disease

	familial alzheimer disease		
HP:0008258	congenital adrenal hyperplasia	120688	congenital adrenal hyperplasia
HP:0010569	elevated 7-dehydrocholesterol	132671	cholesterol; cholesterol
HP:0004323	abnormality of body weight	573752	Body Weight
HP:0002511	alzheimer disease	572053	Alzheimer's disease
HP:0008741	hypertension due to renal artery hyperplasia	134236	hypertension; cerebral infarction
HP:0000510	retinitis pigmentosa	133792	retinitis pigmentosa
HP:0002626	venous varicosities of celiac and mesenteric vessels	587617	Celiac Disease|
HP:0001297	stroke	129540	stroke; sickle cell anemia
HP:0006510	chronic obstructive pulmonary disease	574491	chronic obstructive pulmonary disease
HP:0002642	arteriovenous fistulas of celiac and mesenteric vessels	567594	Celiac Disease|
HP:0004563	increased spinal bone density	145617	bone density
HP:0002037	inflammatory bowel disease	582026	inflammatory bowel disease
HP:0001654	abnormality of the heart valves	595214	heart disease
HP:0005539	t-cell chronic lymphocytic lymphoma/leukemia	136685	lymphoma
HP:0007716	malignant intraocular melanoma	144852	melanoma
HP:0003233	decreased circulating high-density lipoprotein cholesterol	117215	lipoprotein; lipids
HP:0003077	hyperlipidemia	117391	hyperlipidemia
HP:0001952	abnormal glucose tolerance	128902	glucose tolerance; insulin; polycystic ovary syndrome; androgen levels; anthropometric measuments
HP:0000822	hypertension	572495	Hypertension

HP:0100661	trigeminal neuralgia	127920	IgE	1
HP:0004848	ph-positive acute lymphoblastic leukemia	576283	leukemia	
HP:0005686	patchy osteosclerosis	122419	bone mineral density	
HP:0001977	thrombosis	585000	Thrombosis	
HP:0002354	memory impairment	141391	memory impairment	
HP:0002636	arterial aneurysm of celiac and mesenteric vessels	598699	Celiac disease	
HP:0001658	myocardial infarction	153292	myocardial infarct	
HP:0001677	coronary artery disease	123343	coronary disease	
HP:0200047	chrondritis of pinna	115119	age	1
HP:0001977	thrombosis	153379	thrombosis	
HP:0007716	malignant intraocular melanoma	587218	melanoma	
HP:0008071	maternal hypertension	595726	Hypertension	
HP:0004324	increased body weight	569396	Body Weight	
HP:0004941	extrahepatic portal hypertension	123841	hypertension; renal sodium handling	
HP:0004581	increased anterior vertebral height	600454	height	
HP:0005506	chronic myelogenous leukemia	129366	leukemia	
HP:0008071	maternal hypertension	131792	hypertension	
HP:0001717	coronary artery calcification	558931	Coronary Disease	
HP:0100753	schizophrenia	581747	schizophrenia	
HP:0000939	osteoporosis	137838	osteoporosis	
HP:0011025	abnormality of cardiovascular system physiology	567204	Cardiovascular Diseases	
HP:0003362	increased circulating very-low-density lipoprotein cholesterol	116995	lipoprotein	

HP:0004956	systolic hypertension			1
HP:0005978	noninsulin-dependent diabetes mellitus	593931	Diabetes Mellitus	
HP:0007213	late-onset form of familial alzheimer disease	585833	Alzheimer's disease	
HP:0004936	venous thrombosis	569418	Thrombosis	
HP:0000047	hypospadias	147309	hypospadias	
HP:0000822	hypertension	116170	hypertension	
HP:0007028	choreoathetosis			
HP:0001638	cardiomyopathy	131250	cardiomyopathy	
HP:0002099	asthma	118834	asthma	
HP:0100021	cerebral paralysis	591626	Cerebral Palsy	
HP:0004941	extrahepatic portal hypertension	570430	Hypertension	
HP:0002725	systemic lupus erythematosus	140672	lupus erythematosus; rheumatoid arthritis	
HP:0008513	bilateral conductive deafness	123976	deafness	
HP:0000027	azoospermia	586880	Azoospermia	
HP:0007302	bipolar affective disorder	144752	bipolar disorder	
HP:0002650	scoliosis	577384	Scoliosis	
HP:0005700	increased bone density with cystic changes	132454	bone density	
HP:0010659	patchy increased and decreased bone mineral density	596044	Bone Mineral Density	
HP:0011001	increased bone mineral density	587828	Bone Mineral Density	
HP:0005101	high-frequency hearing impairment	557523	Hearing Loss	
HP:0007673	cystic macular degeneration	579365	macular degeneration	
HP:0005101	high-frequency hearing impairment	562345	Hearing Loss	
HP:0003002	breast carcinoma	117535	breast cancer	
HP:0003002	breast carcinoma	146110	breast cancer	
HP:0001033	facial flushing after alcohol intake	593404	alcohol	
HP:0000938	osteopenia	584764	Bone Mineral Density	

HP:0007123	subcortical dementia			1
HP:0100817	renovascular hypertension	115664	hypertension	
HP:0003940	osteoarthritis of the elbow	135759	arthritis	
HP:0009726	renal neoplasm	150612	cancer	
HP:0001668	heart block	593195	heart disease	
HP:0000753	autism with high cognitive abilities	596216	Autism	
HP:0006702	spontaneous coronary artery dissection	129028	coronary disease	
HP:0001729	congenital hearing loss	577567	Hearing Loss	
HP:0002725	systemic lupus erythematosus	130530	lupus erythematosus	
HP:0007573	late onset atopic dermatitis	128448	Atopic dermatitis	
HP:0000875	episodic hypertension	566926	Hypertension	
HP:0006726	increased risk of leukemia	114591	leukemia	
HP:0002608	celiac disease	569881	Celiac Disease|	
HP:0003512	adult female height 130-157 cm	573892	height	
HP:0001909	leukemia	125361	leukemia	
HP:0005686	patchy osteosclerosis	561209	Bone Mineral Density	
HP:0002641	peripheral thrombosis	154281	thrombosis	
HP:0007201	cerebral artery atherosclerosis	119363	atherosclerosis	
HP:0008587	mild neurosensory hearing impairment	562222	Hearing Loss	
HP:0100817	renovascular hypertension	116362	hypertension; renal dysfunction	
HP:0001251	ataxia	117647	spinocerebellar ataxia	
HP:0003765	psoriasis	125498	psoriasis	
HP:0004956	systolic hypertension			
HP:0002140	ischemic stroke	595922	Stroke	
HP:0008603	congenital severe sensorineural	577557	Hearing Loss	

	hearing impairment		
HP:0000875	episodic hypertension	116909	hypertension
HP:0000717	autism	570396	Autism
HP:0001047	atopic dermatitis	583082	atopic dermatitis
HP:0010657	patchy reduction of bone mineral density	119427	bone mineral density
HP:0000822	hypertension	568086	hypertension
HP:0010658	patchy changes of bone mineral density	569281	Bone Mineral Density
HP:0006726	increased risk of leukemia	139643	leukemia
HP:0001370	rheumatoid arthritis	580163	rheumatoid arthritis
HP:0001677	coronary artery disease	563408	Coronary Disease
HP:0002092	pulmonary hypertension	116344	hypertension
HP:0002083	migraine without aura	575435	migraine
HP:0000708	behavioural/psychiatric abnormality	134898	psychiatric disorders
HP:0005686	patchy osteosclerosis	557311	Bone Mineral Density
HP:0002961	dysgammaglobulinemia	142793	globulin
HP:0100817	renovascular hypertension	574147	hypertension
HP:0002099	asthma	592113	asthma
HP:0001513	obesity	586012	Obesity
HP:0004754	paroxysmal or chronic atrial fibrillation	586463	Atrial Fibrillation
HP:0002861	malignant melanoma	148737	melanoma
HP:0004843	familial acute myelogenous leukemia	572721	leukemia
HP:0006510	chronic obstructive pulmonary disease	574529	chronic obstructive pulmonary disease
HP:0006625	breast cancer		
HP:0011001	increased bone mineral density	556301	Bone Mineral Density

HP:0005138	coronary artery disease presenting after age 30 years in heterozygotes		
HP:0010659	patchy increased and decreased bone mineral density	581147	Bone Mineral Density
HP:0000825	hyperinsulinemic hypoglycemia	127001	insulin
HP:0005145	coronary artery stenosis	585389	Coronary Disease
HP:0004836	acute promyelocytic leukemia	125161	leukemia
HP:0001409	portal hypertension	152191	hypertension
HP:0005161	premature sudden cardiac death	565764	sudden cardiac death
HP:0008603	congenital severe sensorineural hearing impairment	581091	Hearing Loss
HP:0008041	late onset congenital glaucoma	571453	Glaucoma
HP:0001936	idiopathic thrombocytopenia	124153	thrombocytopenia
HP:0003518	adult male height 142-169 cm	597603	height
HP:0007302	bipolar affective disorder	145533	bipolar disorder
HP:0000528	anophthalmia	134478	anophthalmia; coloboma; microphthalmia
HP:0004812	human pre-b-cell acute lymphoblastic leukemia	597861	leukemia
HP:0010972	anemia of inadequate production	119818	anemia; thrombocytopenic purpura
HP:0002642	arteriovenous fistulas of celiac and mesenteric vessels	596672	Celiac Disease|
HP:0011059	localized periodontitis	119508	periodontitis
HP:0011058	generalized periodontitis	115208	periodontitis
HP:0000822	hypertension	561863	Hypertension
HP:0001730	progressive hearing impairment	558922	hearing loss
HP:0005539	t-cell chronic lymphocytic lymphoma/leukemi	564781	lymphoma

	a			
HP:0001665	abnormality of cardiac conduction	134299	cardiac disease	
HP:0010675	abnormality of the mineralisation and ossification of bones of the feet	569703	Bone Mineral Density	
HP:0010927	abnormality of divalent inorganic cation homeostasis	131637	EO	1
HP:0002565	complex cardiac malformations	134127	cardiovascular disease; periodontal disease	
HP:0004348	abnormality of bone mineral density	592824	Bone Mineral Density	
HP:0002896	neoplasm of the liver	131414	liver cancer	
HP:0005117	elevated diastolic blood pressure	599295	Diastolic blood pressure	
HP:0002640	hypertension associated with pheochromocytoma	116785	hypertension	
HP:0003141	hyperbetalipoprotei nemia	121220	lipoproteins	
HP:0006851	spinal nerve root neurofibromas			1
HP:0006625	breast cancer			1
HP:0005744	generalized osteoporosis with pathologic fractures	568710	Osteoporosis	
HP:0010675	abnormality of the mineralisation and ossification of bones of the feet	582248	Bone Mineral Density	
HP:0008874	truncal obesity developing in mid-childhood	116840	obesity	
HP:0001626	abnormality of the cardiovascular system	115479	Cardiovascular Disease	
HP:0001297	stroke	122820	stroke	
HP:0001045	vitiligo	585077	Vitiligo	
HP:0002725	systemic lupus erythematosus	137023	Lupus	
HP:0004956	systolic			1

	hypertension			
HP:0006679	granulomatous coronary arteritis	563816	Coronary Disease	
HP:0001025	urticaria	145150	HIV	1
HP:0000825	hyperinsulinemic hypoglycemia	133463	insulin; body fat; metabolic syndrome	
HP:0005506	chronic myelogenous leukemia	147179	leukemia	
HP:0002099	asthma	589664	asthma	
HP:0007673	cystic macular degeneration	579430	Macular Degeneration	
HP:0007302	bipolar affective disorder	145546	bipolar disorder	
HP:0001915	aplastic anemia	116541	anemia	
HP:0008071	maternal hypertension	129298	hypertension	
HP:0002615	hypotension	120988	blood pressure	
HP:0007123	subcortical dementia			1
HP:0004780	hypertrichosis limited to elbows	152017	hypertension	
HP:0005550	chronic lymphatic leukemia	132228	leukemia	
HP:0008071	maternal hypertension	133686	hypertension	
HP:0100279	ulcerative colitis	600092	ulcerative colitis	
HP:0001409	portal hypertension	145965	hypertension	
HP:0008849	low birth weight in males	124739	birth weight	
HP:0008871	height less than 3rd percentile	568641	height	
HP:0003141	hyperbetalipoproteinemia	596854	LDL cholesterol	
HP:0007150	dementia			
HP:0002641	peripheral thrombosis	575303	Thrombosis	
HP:0001904	autoimmune neutropenia	116415	autoimmune; polyendocrinopathy-candidiasis-ectodermal dystrophy; polyendocrinopathy-candidiasis-ectodermal; dystrophy	
HP:0100753	schizophrenia	121256	schizophrenia	
HP:0100808	gastric diverticulum	124499	gastric disease	
HP:0011001	increased bone	147317	bone density	

	mineral density			
HP:0005700	increased bone density with cystic changes	145639	bone density	
HP:0005700	increased bone density with cystic changes	150142	bone density	
HP:0006733	acute megakaryocytic leukemia	138256	leukemia; lung cancer; laryngeal cancer; bladder cancer; oral-pharyngeal cancer	
HP:0000875	episodic hypertension	124042	Hypertension	
HP:0005686	patchy osteosclerosis	589494	Bone Mineral Density	
HP:0007213	late-onset form of familial alzheimer disease	588465	Alzheimer's disease	
HP:0004956	systolic hypertension			1
HP:0004956	systolic hypertension			1
HP:0008538	sensorineural deafness	123969	deafness	
HP:0000877	insulin-resistant diabetes mellitus at puberty	128932	insulin; obesity	
HP:0003002	breast carcinoma	135305	breast cancer	
HP:0008915	truncal obesity apparent in childhood	574307	Obesity	
HP:0004421	elevated systolic blood pressure	565892	blood pressure	
HP:0007694	pigmented macular degeneration	563520	Macular Degeneration	
HP:0001677	coronary artery disease	597217	coronary disease	
HP:0007694	pigmented macular degeneration	562058	Macular Degeneration	
HP:0008332	mild hypertriglyceridemia	566107	Hypertriglyceridemia	
HP:0008625	severe sensorineural hearing impairment	576311	Hearing Loss	
HP:0007947	pericentral retinitis pigmentosa	132409	retinitis pigmentosa	

HP:0003940	osteoarthritis of the elbow	140554	arthritis
HP:0002511	alzheimer disease	587651	Alzheimer's disease
HP:0004563	increased spinal bone density	150580	bone density
HP:0004843	familial acute myelogenous leukemia	126122	leukemia
HP:0008071	maternal hypertension	572273	Hypertension
HP:0002511	alzheimer disease	587953	Alzheimer's disease
HP:0001729	congenital hearing loss	572720	Hearing Loss
HP:0008332	mild hypertriglyceridemia	598633	triglycerides
HP:0003002	breast carcinoma	124802	breast cancer
HP:0001513	obesity	130323	obesity; energy expenditure
HP:0000831	insulin-resistant diabetes mellitus	137639	insulin; obesity
HP:0002511	alzheimer disease	123819	Alzheimer's disease; chronic obstructive pulmonary disease/COPD
HP:0008915	truncal obesity apparent in childhood	587758	obesity
HP:0003084	fractures of the long bones	119473	fractures
HP:0001889	megaloblastic anemia	147843	anemia
HP:0006510	chronic obstructive pulmonary disease	586421	chronic obstructive pulmonary disease
HP:0005145	coronary artery stenosis	560566	Coronary Artery Disease|
HP:0005686	patchy osteosclerosis	556575	Bone Mineral Density
HP:0008610	infantile sensorineural hearing impairment	556243	hearing loss
HP:0007716	malignant intraocular melanoma	132088	melanoma
HP:0006754	posterior fossa and upper cervical meningiomas	118585	meningioma

HP:0004322	short stature	587668	height
HP:0006718	increased risk of colorectal cancer	135633	colorectal cancer
HP:0004348	abnormality of bone mineral density	563598	Bone Mineral Density
HP:0008741	hypertension due to renal artery hyperplasia	124098	hypertension
HP:0001127	progressive retinitis pigmentosa	128666	retinitis pigmentosa
HP:0002894	neoplasm of the pancreas	140991	pancreatic cancer
HP:0005700	increased bone density with cystic changes	129379	bone density; fractures; osteocalcin
HP:0002755	osteomyelitis due to immunodeficiency	584299	Osteomyelitis
HP:0006733	acute megakaryocytic leukemia	133473	leukemia; Noonan syndrome
HP:0001956	truncal obesity	123690	obesity
HP:0004808	acute myeloid leukemia	136047	leukemia
HP:0002608	celiac disease	140087	celiac disease
HP:0100292	amyloidosis of peripheral nerves	130512	amyloidosis; Familial Mediterranean Fever
HP:0003156	increased liver function tests	115984	liver disease
HP:0003512	adult female height 130-157 cm	597354	height
HP:0003518	adult male height 142-169 cm	561337	height
HP:0001677	coronary artery disease	583641	Coronary Disease
HP:0008245	tsh deficient hypothyroidism	594333	Thyroid Diseases
HP:0001672	symmetric		
HP:0002642	arteriovenous fistulas of celiac and mesenteric vessels	569275	Celiac Disease|
HP:0001727	thromboembolic stroke may occur	141305	stroke
HP:0004323	abnormality of body weight	578263	Body Weight

HP:0002665	lymphoma	592195	lymphoma	
HP:0008874	truncal obesity developing in mid-childhood	130311	obesity	
HP:0003193	allergic rhinitis	128208	rhinitis	
HP:0004929	coronary atherosclerosis	586535	Coronary Disease	
HP:0100753	schizophrenia	133620	schizophrenia; autism; alcoholism; bipolar disorder; attention deficit hyperactivity disorder	
HP:0001513	Obesity	140740	obesity	
HP:0000964	Eczema	591101	eczema	
HP:0008205	insulin-dependent but ketosis-resistant diabetes	133065	insulin; glucose	
HP:0005686	patchy osteosclerosis	566776	Bone Mineral Density	
HP:0002626	venous varicosities of celiac and mesenteric vessels	559049	Celiac Disease	
HP:0003002	breast carcinoma	599891	Breast cancer	
HP:0002861	malignant melanoma	580727	Melanoma	
HP:0006727	t-cell acute lymphoblastic leukemias	129366	leukemia	
HP:0001409	portal hypertension	133674	hypertension	
HP:0004581	increased anterior vertebral height	597169	height	
HP:0003124	hypercholesterolemia	116890	serum cholesterol	
HP:0000028	cryptorchidism	126360	cryptorchidism	
HP:0001045	Vitiligo	579643	Vitiligo	
HP:0006625	breast cancer			
HP:0007213	late-onset form of familial alzheimer disease	577974	Alzheimer's disease	
HP:0003131	cystinuria	141798	cystinuria	
HP:0005897	severe osteoporosis	131637	EO	1
HP:0007302	bipolar affective disorder	581545	Bipolar Disorder	
HP:0008874	truncal obesity developing in mid-childhood	140835	obesity	

HP:0000822	hypertension	569645	Hypertension	
HP:0000875	episodic hypertension	120201	hypertension; hyperaldosteronism	
HP:0001665	abnormality of cardiac conduction	591922	heart disease	
HP:0008071	maternal hypertension	141418	hypertension	
HP:0002511	alzheimer disease	126995	Alzheimer's Disease	
HP:0008741	hypertension due to renal artery hyperplasia	577834	Hypertension	
HP:0004581	increased anterior vertebral height	600484	height	
HP:0006775	multiple myeloma	146498	multiple myeloma	
HP:0007302	bipolar affective disorder	556360	Bipolar Disorder	
HP:0010658	patchy changes of bone mineral density	581121	Bone Mineral Density	
HP:0002156	homocystinuria	558748	Cystinuria	
HP:0002636	arterial aneurysm of celiac and mesenteric vessels	557504	Celiac Disease|	
HP:0005681	rheumatoid arthritis			
HP:0010675	abnormality of the mineralisation and ossification of bones of the feet	573251	Bone Mineral Density	
HP:0007213	late-onset form of familial alzheimer disease	572105	Alzheimer's disease	
HP:0004904	insulin-dependent maturity-onset diabetes of the young	135837	diabetes	
HP:0001658	myocardial infarction	119618	myocardial infarct; atherosclerosis	
HP:0100753	schizophrenia	146705	schizophrenia	
HP:0001286	low intelligence	129368	IgE	1
HP:0008619	bilateral sensorineural hearing impairment	594375	hearing loss	
HP:0000753	autism with high cognitive abilities	558335	Autism	

HP:0001087	congenital glaucoma	117099	glaucoma
HP:0003124	hypercholesterolemia	114718	cholesterol; cholesterol
HP:0008874	truncal obesity developing in mid-childhood	599169	obesity
HP:0002401	stroke-like episodes	129121	stroke
HP:0003518	adult male height 142-169 cm	598892	height
HP:0003146	hypocholesterolemia	121189	cholesterol
HP:0001370	rheumatoid arthritis	133502	rheumatoid arthritis
HP:0004563	increased spinal bone density	118237	bone density
HP:0007302	bipolar affective disorder	126869	bipolar disorder
HP:0001956	truncal obesity	134370	obesity; retinal vascular occlusion
HP:0008741	hypertension due to renal artery hyperplasia	141232	hypertension
HP:0003518	adult male height 142-169 cm	599319	height
HP:0001297	Stroke	141068	stroke
HP:0007018	attention deficit hyperactivity disorder	135008	attention deficit hyperactivity disorder
HP:0004820	acute myelomonocytic leukemia	138517	leukemia
HP:0008282	unconjugated hyperbilirubinemia	154951	bilirubin
HP:0006510	chronic obstructive pulmonary disease	575814	chronic obstructive pulmonary disease
HP:0000002	abnormality of body height	599698	height
HP:0004941	extrahepatic portal hypertension	588621	Hypertension
HP:0004956	systolic hypertension		
HP:0010658	patchy changes of bone mineral density	568632	Bone Mineral Density
HP:0004929	coronary atherosclerosis	574377	Coronary Disease

HP:0001518	low birth weight	136084	small for gestational age	
HP:0005681	rheumatoid arthritis			
HP:0010675	abnormality of the mineralisation and ossification of bones of the feet	593655	Bone Mineral Density	
HP:0009831	mononeuropathy	142824	neuropathy	
HP:0000716	depression	556501	depression	
HP:0006834	developmental stagnation at onset of seizures	123733	seizures	
HP:0005978	noninsulin-dependent diabetes mellitus	150454	insulin	
HP:0006753	neoplasm of the stomach	584840	gastric cancer	
HP:0002401	stroke-like episodes	560733	Stroke	
HP:0001717	coronary artery calcification	567957	Coronary Disease	
HP:0004780	hypertrichosis limited to elbows	565663	hypertension	
HP:0005907	broad metatarsals and phalanges	586292	ALS	1
HP:0005100	premature birth following premature rupture of fetal membranes	578537	Premature Birth	
HP:0100817	renovascular hypertension	135981	hypertension	
HP:0005517	t-cell lymphoma/leukemia	120341	lymphoma; non-Hodgkin's lymphoma	
HP:0004780	hypertrichosis limited to elbows	141681	hypertension; glucose tolerance; insulin; hematology indices	
HP:0001513	Obesity	577452	Obesity	
HP:0200025	mandibular pain	593863	Pain	
HP:0007813	nongranulomatous uveitis	136705	uveitis	
HP:0004956	systolic hypertension			
HP:0003464	abnormal cholesterol homeostasis	116808	cholesterol; triglycerides; cholesterol	
HP:0100817	renovascular	572543	Hypertension	

	hypertension		
HP:0005550	chronic lymphatic leukemia	139622	leukemia
HP:0100753	schizophrenia	126754	schizophrenia; suicide
HP:0000704	periodontitis	139497	periodontal disease
HP:0000704	periodontitis	130704	periodontitis
HP:0000510	retinitis pigmentosa	128666	retinitis pigmentosa
HP:0001409	portal hypertension	116522	hypertension
HP:0004941	extrahepatic portal hypertension	124041	hypertension
HP:0002626	venous varicosities of celiac and mesenteric vessels	129341	celiac disease
HP:0004563	increased spinal bone density	131846	bone density
HP:0005516	myelodysplasia and acute myelogenous leukemia	138266	leukemia; lung cancer; oropharyngolaryngeal cancers; laryngeal cancer; bladder cancer
HP:0004421	elevated systolic blood pressure	600004	Blood pressure
HP:0006718	increased risk of colorectal cancer	557792	colorectal cancer
HP:0005558	chronic leukemia	125251	leukemia
HP:0003765	psoriasis	590568	Psoriasis
HP:0002140	ischemic stroke	575278	Stroke
HP:0002608	celiac disease	581773	Celiac Disease
HP:0001513	Obesity	585751	Obesity
HP:0100753	schizophrenia	148141	schizophrenia
HP:0008071	maternal hypertension	556835	Hypertension
HP:0008598	mild conductive hearing impairment	577484	hearing loss
HP:0008871	height less than 3rd percentile	600485	height
HP:0003233	decreased circulating high-density lipoprotein cholesterol	117446	lipoprotein
HP:0100817	renovascular hypertension	124280	hypertension
HP:0006233	osteoarthritis of the distal interphalangeal joint	121061	arthritis; diabetes

HP:0000753	autism with high cognitive abilities	572716	Autism	
HP:0005517	t-cell lymphoma/leukemia	140144	leukemia	
HP:0000938	osteopenia	589642	Bone Mineral Density	
HP:0008849	low birth weight in males	596981	weight	
HP:0008579	bilateral sensorineural deafness			1
HP:0000791	uric acid urolithiasis	117488	urolithiasis	
HP:0001727	thromboembolic stroke may occur	124139	stroke	
HP:0002641	peripheral thrombosis	122647	thrombosis	
HP:0005681	rheumatoid arthritis			
HP:0005516	myelodysplasia and acute myelogenous leukemia	155267	leukemia	
HP:0005700	increased bone density with cystic changes	144276	bone density	
HP:0003513	ratio of renal calcium clearance to creatinine clearance usually below 0.01	141274	renal disease	
HP:0001977	thrombosis	591742	Thrombosis	
HP:0006510	chronic obstructive pulmonary disease	594718	chronic obstructive pulmonary disease	
HP:0007417	discoid lupus erythematosus	118935	lupus erythematosus	
HP:0006718	increased risk of colorectal cancer	561895	colorectal cancer	
HP:0006625	breast cancer			
HP:0001727	thromboembolic stroke may occur	575578	Stroke	
HP:0010819	atonic seizures	123733	seizures	
HP:0003563	hypobetalipoproteinemia	150752	lipoprotein	
HP:0003991	osteosclerosis of the ulna	131624	EO	1
HP:0100021	cerebral paralysis	560720	Cerebral Palsy	
HP:0011058	generalized	130822	periodontitis	

	periodontitis		
HP:0004956	systolic hypertension		
HP:0010675	abnormality of the mineralisation and ossification of bones of the feet	599823	Bone Mineral Density
HP:0004868	severe hemolytic anemia	124374	anemia
HP:0004780	hypertrichosis limited to elbows	557827	Hypertension
HP:0006702	spontaneous coronary artery dissection	585161	Coronary Artery Disease
HP:0008598	mild conductive hearing impairment	577510	Hearing Loss
HP:0010837	decreased serum ceruloplasmin	126376	ceruloplasmin
HP:0007694	pigmented macular degeneration	579428	Macular Degeneration
HP:0006127	long		
HP:0000939	osteoporosis	567282	Osteoporosis
HP:0002092	pulmonary hypertension	591788	hypertension
HP:0010658	patchy changes of bone mineral density	559383	Bone Mineral Density
HP:0008741	hypertension due to renal artery hyperplasia	133674	hypertension
HP:0001658	myocardial infarction	129332	myocardial infarct; Crohn's disease; asthma; malaria; Malaria infection; kawasaki disease; psoriasis vulgaris;
HP:0002511	alzheimer disease	134171	Alzheimer's Disease
HP:0008598	mild conductive hearing impairment	566276	hearing loss
HP:0004780	hypertrichosis limited to elbows	116242	hypertension
HP:0004943	accelerated atherosclerosis	593801	Atherosclerosis
HP:0009726	renal neoplasm	131430	cancer; HIV infection; gastrointestinal bleeding; thiopurine methyltransferase activity
HP:0002621	atherosclerosis	577474	Atherosclerosis

HP:0001127	progressive retinitis pigmentosa	123530	retinitis pigmentosa; macular dystrophy	
HP:0008071	maternal hypertension	581110	hypertension	
HP:0001956	truncal obesity	115828	obesity	
HP:0007673	cystic macular degeneration	149213	macular degeneration	
HP:0006718	increased risk of colorectal cancer	152528	colorectal cancer	
HP:0000717	autism	581261	Autism	
HP:0004956	systolic hypertension			
HP:0002725	systemic lupus erythematosus	119567	systemic lupus erythematosus	
HP:0002666	pheochromocytoma	137989	Pheochromocytoma	
HP:0001275	epilepsy	577069	Epilepsy	
HP:0001520	macrosomia	585882	Birth Weight	
HP:0004570	increased vertebral height	599322	height	
HP:0000047	hypospadias	147374	hypospadias	
HP:0005681	rheumatoid arthritis			
HP:0007213	late-onset form of familial alzheimer disease	129891	Alzheimer's Disease	
HP:0100647	graves disease	142399	Graves disease	
HP:0006718	increased risk of colorectal cancer	138298	cancer	
HP:0006727	t-cell acute lymphoblastic leukemias	580539	leukemia	
HP:0005117	elevated diastolic blood pressure	120153	blood pressure	
HP:0003512	adult female height 130-157 cm	599513	height	
HP:0003464	abnormal cholesterol homeostasis	138459	cholesterol; cholesterol	
HP:0100753	schizophrenia	129312	schizophrenia	
HP:0008671	rapid loss of renal function	116203	renal function	
HP:0000855	insulin resistance	134973	insulin; smoking behavior; obesity	
HP:0003131	cystinuria	141792	cystinuria	
HP:0010049	hypoplastic/short metacarpal bones	586292	ALS	1

HP:0008071	maternal hypertension	131801	hypertension; obesity	
HP:0006702	spontaneous coronary artery dissection	571574	Coronary Artery Disease|	
HP:0002592	gastric ulcer	151268	gastric ulcer	
HP:0004941	extrahepatic portal hypertension	588967	Hypertension	
HP:0008915	truncal obesity apparent in childhood	562127	obesity	
HP:0001513	obesity	589077	obesity	
HP:0008617	progressive bilateral sensorineural hearing loss	577500	hearing loss	
HP:0000751	personality changes	568987	personality	
HP:0011059	localized periodontitis	125320	periodontitis	
HP:0005130	restrictive heart failure	591379	heart disease	
HP:0004780	hypertrichosis limited to elbows	144271	hypertension	
HP:0003002	breast carcinoma	124675	breast cancer	
HP:0100817	renovascular hypertension	572017	hypertension	
HP:0001668	heart block	559155	heart disease	
HP:0005101	high-frequency hearing impairment	567880	Hearing Loss	
HP:0002608	celiac disease	150513	celiac disease	
HP:0006726	increased risk of leukemia	138256	leukemia; lung cancer; laryngeal cancer; bladder cancer; oral-pharyngeal cancer	
HP:0006625	breast cancer			
HP:0001297	stroke	151513	stroke	
HP:0004929	coronary atherosclerosis	141817	atherosclerosis	
HP:0000843	hyperparathyroidism	130436	Hyperparathyroidism	
HP:0009132	abnormality of bone mineral density involving tarsal bones	570826	Bone Mineral Density	
HP:0011042	abnormality of potassium homeostasis	131637	EO	1

HP:0100279	ulcerative colitis	600094	ulcerative colitis	
HP:0004780	hypertrichosis limited to elbows	119149	hypertension	
HP:0003002	breast carcinoma	137687	cancer	
HP:0007302	bipolar affective disorder	594806	Bipolar Disorder	
HP:0003124	hypercholesterolemia	119124	Hypercholesterolemia	
HP:0003141	hyperbetalipoproteinemia	596875	LDL cholesterol	
HP:0002592	gastric ulcer	134340	gastric ulcer	
HP:0000822	hypertension	567882	hypertension	
HP:0100753	schizophrenia	569325	schizophrenia	
HP:0007133	progressive peripheral neuropathy	132712	neuropathy; lactic acidosis	
HP:0007673	cystic macular degeneration	144933	macular degeneration	
HP:0000717	autism	568799	Autism	
HP:0100848	neoplasia of the male external genitalia	131624	EO	1
HP:0002794	apnea during seizure spells	123657	seizures	
HP:0002665	lymphoma	120462	lymphoma; Hodgkin's disease	
HP:0000405	conductive hearing impairment	594379	Hearing Loss	
HP:0001710	conotruncal defect	575469	heart disease	
HP:0009733	glioma	119017	Glioma	
HP:0004581	increased anterior vertebral height	598623	height	
HP:0000076	vesicoureteral reflux	136037	vesicoureteral reflux	
HP:0002640	hypertension associated with pheochromocytoma	566892	Hypertension	
HP:0005539	t-cell chronic lymphocytic lymphoma/leukemia	138201	leukemia; bladder cancer; radiotherapy	
HP:0006718	increased risk of colorectal cancer	152970	colorectal cancer	
HP:0005086	knee osteoarthritis	125809	arthritis	
HP:0004761	post-angioplasty coronary artery	566291	Coronary Artery Disease	

	restenosis			
HP:0100279	ulcerative colitis	127592	ulcerative colitis	
HP:0004780	hypertrichosis limited to elbows	572197	Hypertension	
HP:0008619	bilateral sensorineural hearing impairment	599086	hearing impairment	
HP:0006520	progressive pulmonary function impairment	136391	pulmonary function	
HP:0001909	leukemia	120447	leukemia	
HP:0005506	chronic myelogenous leukemia	580303	leukemia	
HP:0010675	abnormality of the mineralisation and ossification of bones of the feet	136556	bone mineral density	
HP:0001956	truncal obesity	114836	obesity	
HP:0000855	insulin resistance	140749	insulin	
HP:0002412	dystonia			
HP:0006625	breast cancer			
HP:0003156	increased liver function tests	135299	liver disease; oxidative stress	
HP:0007213	late-onset form of familial alzheimer disease	585600	Alzheimer's disease	
HP:0004780	hypertrichosis limited to elbows	136819	hypertension	
HP:0100817	renovascular hypertension	139329	hypertension	
HP:0001370	rheumatoid arthritis	139402	arthritis	
HP:0001657	prolonged qt interval on ekg	144311	long QT syndrome	
HP:0008625	severe sensorineural hearing impairment	588268	Hearing Loss	
HP:0005243	partial abdominal muscle agenesis	115119	age	1
HP:0008727	idiopathic nephrotic syndrome	131976	nephrotic syndrome	
HP:0009726	renal neoplasm	131767	kidney disease	
HP:0002641	peripheral thrombosis	575531	Thrombosis	
HP:0001997	gout	136101	arthritis; diabetes	

HP:0002642	arteriovenous fistulas of celiac and mesenteric vessels	597992	Celiac disease	
HP:0007533	severe atopic dermatitis	583175	atopic dermatitis	
HP:0010675	abnormality of the mineralisation and ossification of bones of the feet	573790	Bone Mineral Density	
HP:0008741	hypertension due to renal artery hyperplasia	595475	Hypertension	
HP:0007694	pigmented macular degeneration	144472	macular degeneration	
HP:0004421	elevated systolic blood pressure	570460	blood pressure	
HP:0000704	periodontitis	133596	periodontitis	
HP:0100817	renovascular hypertension	593668	Hypertension	
HP:0006679	granulomatous coronary arteritis	564068	Coronary Disease	
HP:0005558	chronic leukemia	139975	leukemia	
HP:0010659	patchy increased and decreased bone mineral density	569686	Bone Mineral Density	
HP:0006679	granulomatous coronary arteritis	597229	coronary disease	
HP:0002120	cerebral cortical atrophy	584961	aging	1
HP:0004581	increased anterior vertebral height	599931	height	
HP:0002511	alzheimer disease	557115	Alzheimer's disease	
HP:0000831	insulin-resistant diabetes mellitus	141643	insulin; diabetes	
HP:0000939	osteoporosis	584028	Osteoporosis	
HP:0002636	arterial aneurysm of celiac and mesenteric vessels	140077	celiac disease	
HP:0001513	obesity	137822	obesity	
HP:0008071	maternal hypertension	572017	hypertension	
HP:0005086	knee osteoarthritis	130673	arthritis	
HP:0006562	viral hepatitis	131305	Hepatitis	
HP:0007213	late-onset form of	566601	Alzheimer's disease	

	familial alzheimer disease		
HP:0100753	schizophrenia	121503	schizophrenia
HP:0002934	distal limb muscle atrophy due to peripheral neuropathy	132712	neuropathy; lactic acidosis
HP:0005539	t-cell chronic lymphocytic lymphoma/leukemia	152162	leukemia
HP:0002608	celiac disease	142584	celiac disease
HP:0008245	tsh deficient hypothyroidism	154834	hypothyroidism
HP:0008527	congenital sensorineural hearing impairment	577510	Hearing Loss
HP:0008619	bilateral sensorineural hearing impairment	562345	Hearing Loss
HP:0007284	myoclonic seizures may occur	129262	seizures
HP:0002092	pulmonary hypertension	145172	hypertension
HP:0008741	hypertension due to renal artery hyperplasia	124280	hypertension
HP:0008620	congenital sensorineural deafness	577490	Hearing Loss
HP:0002650	scoliosis	564152	Scoliosis
HP:0002636	arterial aneurysm of celiac and mesenteric vessels	151119	celiac disease
HP:0008907	dwarfism		
HP:0001730	progressive hearing impairment	600703	hearing impairment
HP:0004941	extrahepatic portal hypertension	120112	hypertension
HP:0009020	exercise-induced muscle fatigue	558465	Fatigue
HP:0008581	early conductive hearing loss	556262	Hearing Loss
HP:0002511	alzheimer disease	138001	Alzheimer's Disease

HP:0005602	progressive vitiligo	148544	vitiligo	
HP:0004929	coronary atherosclerosis	572172	Coronary Disease	
HP:0005531	myeloid/lymphoid leukemia	117698	leukemia	
HP:0001626	abnormality of the cardiovascular system	132932	cardiovascular	
HP:0000002	abnormality of body height	600476	Height	
HP:0007108	demyelinating peripheral neuropathy	135357	Neuropathy	
HP:0010675	abnormality of the mineralisation and ossification of bones of the feet	560396	Bone Mineral Density	
HP:0001081	cholelithiasis	577310	Gallstones	
HP:0005681	rheumatoid arthritis			
HP:0002511	alzheimer disease	595895	Alzheimer's disease	
HP:0002511	alzheimer disease	117443	Alzheimer's Disease	
HP:0002092	pulmonary hypertension	131787	Hypertension	
HP:0005110	atrial fibrillation	598367	Atrial fibrillation	
HP:0002027	abdominal pain	593843	Pain	
HP:0008871	height less than 3rd percentile	600000	Height	
HP:0004780	hypertrichosis limited to elbows	120091	Hypertension	1
HP:0100027	recurrent pancreatitis	133262	Pancreatitis	
HP:0002511	alzheimer disease	570332	Alzheimer's disease	
HP:0001275	epilepsy	124221	Epilepsy	
HP:0000510	retinitis pigmentosa	123530	retinitis pigmentosa; macular dystrophy	
HP:0005526	lymphoid leukemia	131905	leukemia	
HP:0007213	late-onset form of familial alzheimer disease	128122	Alzheimer's Disease	
HP:0002642	arteriovenous fistulas of celiac and mesenteric vessels	565100	Celiac Disease|	
HP:0000822	hypertension	145374	hypertension	
HP:0004812	human pre-b-cell acute lymphoblastic	125349	leukemia	

	leukemia			
HP:0008615	late onset sensorineural deafness	594375	hearing loss	
HP:0004941	extrahepatic portal hypertension	565518	hypertension	
HP:0007782	peripheral retinal cone degeneration	124891	cone degeneration	
HP:0001677	coronary artery disease	141320	coronary artery disease; nitric oxide	
HP:0008071	maternal hypertension	570883	hypertension	
HP:0007354	amyotrophic lateral sclerosis	565925	Amyotrophic Lateral Sclerosis	
HP:0005040	distal ulceration and osteomyelitis leading to autoamputation	131637	EO	1
HP:0002640	hypertension associated with pheochromocytoma	593094	Hypertension	
HP:0005558	chronic leukemia	131781	leukemia	
HP:0001409	portal hypertension	569056	hypertension	
HP:0000365	hearing impairment	563644	Hearing Loss	
HP:0004929	coronary atherosclerosis	560695	Coronary Disease	
HP:0004780	hypertrichosis limited to elbows	141213	Hypertension	
HP:0002626	venous varicosities of celiac and mesenteric vessels	584824	Celiac Disease|	
HP:0010657	patchy reduction of bone mineral density	593528	Bone Mineral Density	
HP:0002608	celiac disease	139828	celiac disease	
HP:0002608	celiac disease	557521	Celiac Disease|	
HP:0004780	hypertrichosis limited to elbows	115381	hypertension	
HP:0001730	progressive hearing impairment	570300	Hearing Loss	
HP:0006725	pancreatic adenocarcinoma	596518	pancreatic adenocarcinoma	
HP:0003141	hyperbetalipoprotei nemia	599407	LDL cholesterol	

HP:0002092	pulmonary hypertension	116157	hypertension	
HP:0003296	hyperthreoninuria	131637	EO	1
HP:0006679	granulomatous coronary arteritis	574249	Coronary Disease|	
HP:0004941	extrahepatic portal hypertension	115880	Hypertension	
HP:0000404	deafness	123970	deafness	
HP:0001402	hepatocellular carcinoma	589952	hepatocellular carcinoma	
HP:0003349	low cholesterol esterification rates	139303	cholesterol; cholesterol	
HP:0002401	stroke-like episodes	141594	stroke	
HP:0002642	arteriovenous fistulas of celiac and mesenteric vessels	567567	Celiac Disease|	
HP:0002092	pulmonary hypertension	114957	hypertension	
HP:0005572	decreased renal tubular phosphate excretion	115160	renal disease	
HP:0006958	abnormal auditory evoked potentials	134810	auditory evoked potential	
HP:0003002	breast carcinoma	138228	breast cancer	
HP:0007868	senile macular degeneration	562064	Macular Degeneration	
HP:0100753	schizophrenia	594840	schizophrenia	
HP:0007673	cystic macular degeneration	571822	Macular Degeneration	
HP:0000101	chronic renal failure	141245	chronic renal failure	
HP:0007417	discoid lupus erythematosus	123224	lupus erythematosus	
HP:0000875	episodic hypertension	592314	hypertension	
HP:0000842	hyperinsulinemia	128950	insulin; blood pressure	
HP:0008843	hip osteoarthritis	125585	osteoarthritis	
HP:0001677	coronary artery disease	562177	Coronary Disease	
HP:0007213	late-onset form of familial alzheimer disease	148506	Alzheimer's disease	
HP:0003002	breast carcinoma	124823	breast cancer	
HP:0004820	acute myelomonocytic	151509	leukemia	

	leukemia		
HP:0000722	obsessive-compulsive disorder	153709	obsessive compulsive disorder
HP:0002511	alzheimer disease	130603	Alzheimer's Disease
HP:0004956	systolic hypertension		
HP:0002626	venous varicosities of celiac and mesenteric vessels	151697	celiac disease
HP:0007889	iridescent posterior subcapsular cataract	569610	Cataract
HP:0004761	post-angioplasty coronary artery restenosis	581047	Coronary Artery Disease|
HP:0002640	hypertension associated with pheochromocytoma	581066	Hypertension
HP:0005130	restrictive heart failure	572502	Heart Failure
HP:0005686	patchy osteosclerosis	562980	Bone Mineral Density
HP:0010628	facial nerve palsy	588511	ALS
HP:0006775	multiple myeloma	131592	multiple myeloma
HP:0006226	osteoarthritis of the first carpometacarpal joint	135756	arthritis
HP:0002292	frontal balding	599252	male-pattern baldness
HP:0005393	recurrent major bacterial infections	119270	bacterial infection
HP:0004780	hypertrichosis limited to elbows	559548	hypertension
HP:0001409	portal hypertension	569950	Hypertension
HP:0004956	systolic hypertension		
HP:0004421	elevated systolic blood pressure	120153	blood pressure
HP:0000408	progressive sensorineural hearing impairment	577546	Hearing Loss
HP:0001729	congenital hearing loss	577490	Hearing Loss
HP:0007716	malignant intraocular	138615	melanoma

	melanoma		
HP:0003518	adult male height 142-169 cm	599782	height
HP:0002621	atherosclerosis	120223	atherosclerosis
HP:0008915	truncal obesity apparent in childhood	585387	Obesity
HP:0003002	breast carcinoma	118263	breast cancer
HP:0007122	dementia		
HP:0007211	gradual onset of cognitive impairment	121545	cognitive impairment
HP:0002511	alzheimer disease	127006	Alzheimer's Disease
HP:0008651	uric acid urolithiasis independent of gout	588357	Gout
HP:0001370	rheumatoid arthritis	153553	rheumatoid arthritis
HP:0002642	arteriovenous fistulas of celiac and mesenteric vessels	600401	Celiac disease
HP:0001733	pancreatitis	136038	pancreatitis
HP:0001409	portal hypertension	114958	hypertension
HP:0001397	hepatic steatosis	569594	Fatty Liver
HP:0002640	hypertension associated with pheochromocytoma	115053	hypertension
HP:0002099	asthma	119283	asthma
HP:0007131	acute demyelinating polyneuropathy	145598	polyneuropathy
HP:0001409	portal hypertension	561009	Hypertension
HP:0100817	renovascular hypertension	116079	hypertension
HP:0007673	cystic macular degeneration	563460	Macular Degeneration
HP:0011001	increased bone mineral density	557313	Bone Mineral Density
HP:0000563	keratoconus	558857	keratoconus
HP:0006834	developmental stagnation at onset of seizures	123738	seizures
HP:0002613	biliary cirrhosis	135930	cirrhosis; hepatocellular carcinoma
HP:0004940	arterial calcification		
HP:0004950	peripheral arterial disease	141582	arterial disease

HP:0002092	pulmonary hypertension	569532	hypertension
HP:0010926	aculeiform cataract	569610	Cataract
HP:0002511	alzheimer disease	143555	Alzheimer's disease
HP:0008741	hypertension due to renal artery hyperplasia	557052	Hypertension
HP:0005653	moderate generalized osteoporosis	567564	Osteoporosis
HP:0004416	precocious atherosclerosis	119375	atherosclerosis
HP:0001409	portal hypertension	575466	Hypertension
HP:0008574	severe early sensorineural hearing loss	594374	Hearing Loss
HP:0010659	patchy increased and decreased bone mineral density	557393	Bone Mineral Density
HP:0007716	malignant intraocular melanoma	592067	melanoma
HP:0006702	spontaneous coronary artery dissection	129100	coronary artery disease
HP:0008615	late onset sensorineural deafness	577524	Deafness
HP:0008185	precocious puberty in males	128911	precocious puberty
HP:0006525	lung segmentation defects	572379	Lung Diseases
HP:0008874	truncal obesity developing in mid-childhood	115977	obesity
HP:0100753	schizophrenia	121614	schizophrenia
HP:0006753	neoplasm of the stomach	124842	gastric disease
HP:0004850	recurrent deep vein thrombosis	122881	thrombosis
HP:0008874	truncal obesity developing in mid-childhood	595004	Obesity
HP:0005978	noninsulin-	140904	diabetes mellitus

	dependent diabetes mellitus		
HP:0100817	renovascular hypertension	132000	hypertension
HP:0007204	brain imaging shows diffuse white matter abnormalities	600648	Brain imaging
HP:0008071	maternal hypertension	126326	hypertension
HP:0100817	renovascular hypertension	593958	Hypertension
HP:0006625	breast cancer		
HP:0006677	short pr interval and prolonged qrs		
HP:0008871	height less than 3rd percentile	573891	Height
HP:0003518	adult male height 142-169 cm	600459	Height
HP:0100753	schizophrenia	124252	schizophrenia
HP:0002511	alzheimer disease	578042	Alzheimer's disease
HP:0005138	coronary artery disease presenting after age 30 years in heterozygotes		
HP:0008843	hip osteoarthritis	119522	osteoarthritis
HP:0003940	osteoarthritis of the elbow	115389	arthritis
HP:0008071	maternal hypertension	141490	hypertension
HP:0000405	conductive hearing impairment	577539	Hearing Loss
HP:0002037	inflammatory bowel disease	568533	inflammatory bowel disease
HP:0004433	selective iga deficiency	566873	IgA Deficiency
HP:0004845	acute monocytic leukemia	565379	leukemia
HP:0001370	rheumatoid arthritis	118806	arthritis; spondyloarthropathies
HP:0008741	hypertension due to renal artery hyperplasia	576561	hypertension
HP:0008071	maternal hypertension	559548	hypertension
HP:0008513	bilateral conductive deafness	124015	deafness
HP:0004904	insulin-dependent	128965	insulin; glucose; polycystic ovary syndrome

	maturity-onset diabetes of the young		
HP:0000875	episodic hypertension	568930	Hypertension
HP:0000822	Hypertension	134541	hypertension
HP:0004780	hypertrichosis limited to elbows	131825	hypertension; cardiovascular disease
HP:0100279	ulcerative colitis	597949	ulcerative colitis
HP:0004956	systolic hypertension		
HP:0003512	adult female height 130-157 cm	597094	Height
HP:0010659	patchy increased and decreased bone mineral density	591272	Bone Mineral Density
HP:0008573	Progressive		
HP:0008540	bilateral congenital sensorineural deafness	577565	Hearing Loss
HP:0009900	unilateral deafness	577495	Deafness
HP:0001658	myocardial infarction	117273	myocardial infarct
HP:0011058	generalized periodontitis	139461	periodontitis
HP:0001956	truncal obesity	138835	Obesity
HP:0004956	systolic hypertension		
HP:0006625	breast cancer		
HP:0000704	Periodontitis	139524	periodontitis
HP:0000855	insulin resistance	124182	Insulin
HP:0008741	hypertension due to renal artery hyperplasia	115664	hypertension
HP:0005145	coronary artery stenosis	587299	Coronary Artery Disease
HP:0002728	chronic mucocutaneous candidiasis	120113	candidiasis
HP:0000708	behavioural/psychiatric abnormality	126733	behavioral disorder
HP:0005159	dilated cardiomyopathy may be present	139842	cardiomyopathy

HP:0007213	late-onset form of familial alzheimer disease	557838	Alzheimer's disease
HP:0002972	reduced delayed hypersensitivity	565278	Hypersensitivity
HP:0003558	rhabdomyolysis may follow severe exercise in hot conditions		
HP:0001730	progressive hearing impairment	568142	Hearing Loss
HP:0000875	episodic hypertension	116324	hypertension
HP:0002626	venous varicosities of celiac and mesenteric vessels	125595	celiac disease
HP:0001897	normocytic anemia	149113	anemia
HP:0006733	acute megakaryocytic leukemia	150574	leukemia
HP:0001392	abnormality of the liver	120884	liver disease
HP:0002943	thoracic scoliosis	147314	scoliosis
HP:0001717	coronary artery calcification	129488	Coronary Artery Disease
HP:0001399	hepatic failure	120884	liver disease
HP:0002187	intellectual disability		
HP:0002640	hypertension associated with pheochromocytoma	117442	hypertension
HP:0001118	juvenile cataract	569610	Cataract
HP:0005145	coronary artery stenosis	567954	Coronary Artery Disease
HP:0001677	coronary artery disease	598437	coronary disease
HP:0004348	abnormality of bone mineral density	593965	Bone Mineral Density
HP:0001956	truncal obesity	583048	obesity
HP:0005516	myelodysplasia and acute myelogenous leukemia	130582	leukemia
HP:0001727	thromboembolic stroke may occur	122091	Stroke
HP:0000098	tall stature	147990	Height
HP:0002274	Dementia		
HP:0003940	osteoarthritis of the elbow	128120	arthritis

HP:0004941	extrahepatic portal hypertension	564197	Hypertension
HP:0002099	Asthma	136516	Asthma
HP:0008538	sensorineural deafness	562362	Hearing Loss
HP:0000857	neonatal insulin-dependent diabetes mellitus	138364	Insulin
HP:0100739	Bulimia	558585	Bulimia
HP:0005181	premature coronary artery disease	136648	coronary artery disease
HP:0006704	abnormality of the coronary arteries	591964	Coronary Disease
HP:0002636	arterial aneurysm of celiac and mesenteric vessels	600434	Celiac disease
HP:0004780	hypertrichosis limited to elbows	151836	hypertension
HP:0001824	weight loss	596763	weight
HP:0000112	Nephropathy	123144	nephropathy
HP:0100753	Schizophrenia	566322	schizophrenia
HP:0007213	late-onset form of familial alzheimer disease	596887	Alzheimer's disease
HP:0000875	episodic hypertension	588879	hypertension
HP:0000124	renal tubular dysfunction	141631	renal disease
HP:0002668	Paragangliomas	559774	Glioma
HP:0000938	Osteopenia	587686	Bone Mineral Density
HP:0001649	Tachycardia	559010	heart disease
HP:0001735	acute pancreatitis	124578	pancreatitis
HP:0007018	attention deficit hyperactivity disorder	135046	attention deficit hyperactivity disorder
HP:0000704	Periodontitis	123257	periodontitis
HP:0000875	episodic hypertension	131801	hypertension; obesity
HP:0100577	urinary bladder inflammation	569441	Inflammation
HP:0004936	venous thrombosis	591525	Thrombosis
HP:0000729	pervasive developmental disorder	130375	mental disorder

HP:0006625	breast cancer		
HP:0010658	patchy changes of bone mineral density	127253	bone mineral density
HP:0004875	neonatal inspiratory stridor	131624	EO
HP:0005681	rheumatoid arthritis		
HP:0005181	premature coronary artery disease	576957	Coronary Disease
HP:0004941	extrahepatic portal hypertension	139585	hypertension; autonomic nervous system dysfunction
HP:0001520	Macrosomia	125869	birth weight
HP:0007213	late-onset form of familial alzheimer disease	593928	Alzheimer's disease
HP:0007533	severe atopic dermatitis	128867	atopic dermatitis
HP:0006753	neoplasm of the stomach	127974	gastric cancer
HP:0008874	truncal obesity developing in mid-childhood	569176	obesity
HP:0006523	pulmonary fibrosis due to recurrent infections	134419	pulmonary fibrosis
HP:0005978	noninsulin-dependent diabetes mellitus	133463	insulin; body fat; metabolic syndrome
HP:0002861	malignant melanoma	122390	melanoma
HP:0007302	bipolar affective disorder	121439	bipolar disorder
HP:0007716	malignant intraocular melanoma	150998	melanoma
HP:0006718	increased risk of colorectal cancer	562788	colorectal cancer
HP:0007150	Dementia		
HP:0008741	hypertension due to renal artery hyperplasia	564051	hypertension
HP:0000753	autism with high cognitive abilities	586186	Autism
HP:0100753	Schizophrenia	135123	schizophrenia

HP:0011001	increased bone mineral density	582132	Bone Mineral Density
HP:0004780	hypertrichosis limited to elbows	121201	hypertension
HP:0008014	central fundal arteriolar microaneurysms	151303	aneurysm
HP:0004941	extrahepatic portal hypertension	116229	hypertension
HP:0100817	renovascular hypertension	572241	Hypertension
HP:0100753	Schizophrenia	583636	schizophrenia
HP:0008024	congenital nuclear cataract	119632	congenital nuclear cataract
HP:0004761	post-angioplasty coronary artery restenosis	560815	Coronary Disease|
HP:0005299	premature peripheral vascular disease	122765	Vascular Disease
HP:0002592	gastric ulcer	120869	gastric disease
HP:0002636	arterial aneurysm of celiac and mesenteric vessels	563785	Celiac Disease|
HP:0000851	congenital hypothyroidism	561894	Congenital Hypothyroidism
HP:0004820	acute myelomonocytic leukemia	589917	leukemia
HP:0008915	truncal obesity apparent in childhood	577805	obesity
HP:0006725	pancreatic adenocarcinoma	590239	pancreatic adenocarcinoma
HP:0005686	patchy osteosclerosis	561016	Bone Mineral Density
HP:0006704	abnormality of the coronary arteries	593000	Coronary Disease
HP:0005519	myeloid cell leukemia 1	148421	leukemia
HP:0001727	thromboembolic stroke may occur	131788	Stroke
HP:0010775	vascular ring	128969	Vascular Disease
HP:0003126	low-molecular-weight proteinuria	579133	Proteinuria

HP:0001899	increased hematocrit	597853	hematocrit
HP:0001657	prolonged qt interval on ekg	597747	QT interval
HP:0001513	Obesity	133018	obesity
HP:0001658	myocardial infarction	152702	myocardial infarct
HP:0000938	Osteopenia	557171	Bone Mineral Density
HP:0005625	osteoporosis of vertebrae	568118	Osteoporosis
HP:0001658	myocardial infarction	129051	myocardial infarction; stroke
HP:0000842	Hyperinsulinemia	140297	insulin; obesity
HP:0002511	alzheimer disease	577688	Alzheimer's disease
HP:0005145	coronary artery stenosis	587947	Coronary Disease
HP:0006679	granulomatous coronary arteritis	579559	Coronary Disease
HP:0007213	late-onset form of familial alzheimer disease	569250	Alzheimer's disease
HP:0001956	truncal obesity	150252	obesity
HP:0005700	increased bone density with cystic changes	150802	bone density
HP:0001952	abnormal glucose tolerance	135006	glucose
HP:0001657	prolonged qt interval on ekg	598935	QT interval
HP:0002092	pulmonary hypertension	570343	hypertension
HP:0002725	systemic lupus erythematosus	139463	lupus erythematosus
HP:0100027	recurrent pancreatitis	583720	Pancreatitis
HP:0004820	acute myelomonocytic leukemia	587062	leukemia
HP:0010885	aseptic necrosis	122765	Vascular Disease
HP:0002636	arterial aneurysm of celiac and mesenteric vessels	148839	celiac disease
HP:0003141	hyperbetalipoprotei nemia	132884	lipoprotein; lipids

HP:0005558	chronic leukemia	150519	leukemia
HP:0002626	venous varicosities of celiac and mesenteric vessels	563785	Celiac Disease|
HP:0002099	Asthma	135310	asthma
HP:0004941	extrahepatic portal hypertension	561665	hypertension
HP:0000857	neonatal insulin-dependent diabetes mellitus	581175	Diabetes Mellitus
HP:0006718	increased risk of colorectal cancer	557745	colorectal cancer
HP:0004780	hypertrichosis limited to elbows	591094	hypertension
HP:0006941	Polyneuropathy		
HP:0001710	conotruncal defect	569857	heart disease
HP:0006718	increased risk of colorectal cancer	141892	colorectal cancer
HP:0004853	lethal congenital nonspherocytic		
HP:0002511	alzheimer disease	566001	Alzheimer's disease
HP:0002155	Hypertriglyceridemia	117414	triglycerides; insulin; lipoproteins; apoB; apoC-III
HP:0001727	thromboembolic stroke may occur	141104	stroke
HP:0006733	acute megakaryocytic leukemia	125361	leukemia
HP:0004941	extrahepatic portal hypertension	588884	hypertension
HP:0100753	Schizophrenia	134977	schizophrenia
HP:0003462	elevated 8-dehydrocholesterol	122995	cholesterol; cholesterol
HP:0002626	venous varicosities of celiac and mesenteric vessels	142540	celiac disease
HP:0004348	abnormality of bone mineral density	565497	Bone Mineral Density
HP:0008071	maternal hypertension	128926	hypertension; glucose tolerance; insulin; hematology indices
HP:0002565	complex cardiac malformations	139485	cardiovascular disease; periodontal disease
HP:0004780	hypertrichosis limited to elbows	120230	hypertension

HP:0000002	abnormality of body height	600477	height	
HP:0010675	abnormality of the mineralisation and ossification of bones of the feet	576470	Bone Mineral Density	
HP:0008538	sensorineural deafness	124014	deafness	
HP:0008741	hypertension due to renal artery hyperplasia	136819	hypertension	
HP:0006721	acute lymphatic leukemia	148554	leukemia	
HP:0007280	acute infantile spinal muscular atrophy	558784	spinal muscular atrophy	
HP:0001511	intrauterine growth restriction	151310	intrauterine growth	
HP:0003124	Hypercholesterolemia	141662	cholesterol; cholesterol	
HP:0007983	reduced visual acuity by age 3 years	115119	age	1
HP:0010979	abnormality of the level of lipoprotein cholesterol	129806	cholesterol; triglycerides; lipoprotein	
HP:0007154	nonprogressive intellectual disability	153070	mental retardation	
HP:0008071	maternal hypertension	563904	Hypertension	
HP:0002401	stroke-like episodes	152706	stroke	
HP:0007213	late-onset form of familial alzheimer disease	129902	Alzheimer's Disease	
HP:0001087	congenital glaucoma	132136	glaucoma	
HP:0004904	insulin-dependent maturity-onset diabetes of the young	130316	insulin; obesity; leptin	
HP:0001776	bilateral club feet	581161	Clubfoot	
HP:0007694	pigmented macular degeneration	566395	Macular Degeneration	
HP:0001730	progressive hearing impairment	577485	Hearing Loss	

HP:0008071	maternal hypertension	588621	Hypertension
HP:0008874	truncal obesity developing in mid-childhood	590180	Obesity
HP:0003002	breast carcinoma	117881	breast cancer; ovarian cancer
HP:0008620	congenital sensorineural deafness	588278	Hearing Loss
HP:0002608	celiac disease	583303	Celiac Disease|
HP:0004956	systolic hypertension		
HP:0008711	benign prostatic hyperplasia	155036	prostatic hyperplasia
HP:0002092	pulmonary hypertension	572246	Hypertension
HP:0002092	pulmonary hypertension	153268	hypertension
HP:0002626	venous varicosities of celiac and mesenteric vessels	588701	Celiac Disease|
HP:0000716	Depression	576764	depression
HP:0001033	facial flushing after alcohol intake	568349	alcohol
HP:0008071	maternal hypertension	116154	hypertension
HP:0000361	pulsatile tinnitus (tympanic paraganglioma)	564494	Glioma
HP:0001370	rheumatoid arthritis	580672	rheumatoid arthritis
HP:0006718	increased risk of colorectal cancer	119253	cancer
HP:0001409	portal hypertension	116317	hypertension; left ventricular hypertrophy
HP:0008617	progressive bilateral sensorineural hearing loss	582624	Hearing Loss
HP:0000002	abnormality of body height	599509	height
HP:0003002	breast carcinoma	598450	Breast cancer
HP:0010657	patchy reduction of bone mineral density	585392	Bone Mineral Density
HP:0000822	Hypertension	573863	Hypertension
HP:0006510	chronic obstructive	571370	chronic obstructive pulmonary disease

	pulmonary disease			
HP:0004780	hypertrichosis limited to elbows	121608	hypertension	
HP:0002565	complex cardiac malformations	140592	cardiovascular	
HP:0011001	increased bone mineral density	594193	Bone Mineral Density	
HP:0000822	Hypertension	561833	hypertension	
HP:0006226	osteoarthritis of the first carpometacarpal joint	140086	osteoarthritis	
HP:0002725	systemic lupus erythematosus	140102	lupus erythematosus	
HP:0002642	arteriovenous fistulas of celiac and mesenteric vessels	126191	celiac disease; Wegener's granulomatosis; cervical cancer	
HP:0005686	patchy osteosclerosis	131624	EO	1
HP:0006702	spontaneous coronary artery dissection	589253	Coronary Disease	
HP:0001638	Cardiomyopathy	586454	cardiomyopathy	
HP:0008615	late onset sensorineural deafness	559429	Hearing Loss	
HP:0010900	abnormality of threonine metabolism	131637	EO	1
HP:0000875	episodic hypertension	126469	hypertension	
HP:0002077	migraine with aura	131658	migraine; migraine with aura	
HP:0011025	abnormality of cardiovascular system physiology	115496	cardiovascular disease	
HP:0008741	hypertension due to renal artery hyperplasia	574154	Hypertension	
HP:0003768	periodic paralysis	129234	periodic paralysis	
HP:0008915	truncal obesity apparent in childhood	568584	obesity	
HP:0010659	patchy increased and decreased bone	566455	Bone Mineral Density	

	mineral density			
HP:0100647	graves disease	586728	Graves disease	
HP:0008915	truncal obesity apparent in childhood	126834	obesity	
HP:0004956	systolic hypertension			1
HP:0100817	renovascular hypertension	151979	hypertension	
HP:0010658	patchy changes of bone mineral density	582486	Bone Mineral Density	
HP:0001370	rheumatoid arthritis	127932	rheumatoid arthritis	
HP:0000399	prelingual sensorineural deafness	123965	deafness	
HP:0008590	progressive childhood hearing loss	577492	Hearing Loss	
HP:0007284	myoclonic seizures may occur	114716	seizures	
HP:0008071	maternal hypertension	141231	hypertension	
HP:0006730	myelodysplastic syndrome	154717	myelodysplastic syndrome	
HP:0005517	t-cell lymphoma/leukemia	137270	lymphoma	
HP:0002613	biliary cirrhosis	135343	cirrhosis	
HP:0008915	truncal obesity apparent in childhood	115863	obesity	
HP:0005544	familial chronic myelocytic leukemia-like syndrome	123429	leukemia	
HP:0002511	alzheimer disease	114483	Alzheimer's Disease	
HP:0100753	schizophrenia	121827	schizophrenia	
HP:0008071	maternal hypertension	564196	Hypertension	
HP:0004956	systolic hypertension			1
HP:0008071	maternal hypertension	123440	hypertension	

HP:0001658	myocardial infarction	134199	myocardial infarction
HP:0010675	abnormality of the mineralisation and ossification of bones of the feet	581103	Bone Mineral Density

Table 5 shows the evaluation of a 1,000 links between HOP and GAD using the suffix array. A "1" was entered on the "error rate" column every time a mismatch was found. As the table shows, most matchings were correct.

Table six shows a comparison between the basis statistics found for OGA and the corresponding statistics reported for Neurocarta.

Table 7 Comparison between OGA and Neurocarta

	OGA	**Neurocarta**
Number of links	98,698	30,000
Number of concepts	2,708	2,000
Number of genes	4,666	7,000
Backbone	HOP	HOP, DO, MPO
Curated	No	Yes
Statistical analysis	Yes	No
Interface	Standalone	Website

Table six shows top ten genes with the most associations in OGA.

Table 8 Ten five hundred genes with the most associations in OGA

Gene	Number of HOP Concepts
ACE	1923
NOS3	1659
APOE	1573
GJB2	1042
HLA-DRB1	1008
AGT	971
MTHFR	960
NOS1	866
TNF	770
HLA-DQB1	724
GSTM1	689
HLA-A	655
CYP11B2	632
AGTR1	630
IL6	618
IL1B	577
VDR	568
SLC6A4	532
ESR1	496
ADD1	477
ADRB2	476
IL10	463
SERPINE1	458
GSTT1	457
ABCB1	452
HLA-B	431
PPARG	411
TGFB1	403
PON1	401
GJB6	385
GNB3	384
IL1A	380
CYP2C9	372
HLA-DQA1	372
CFH	369
SOD2	346
F5	337
APOC3	327

CTLA4	326
UGT1A1	323
CYP2D6	319
IL1RN	317
F2	308
COMT	305
LPL	303
ITGB3	294
CYP1A1	277
HTR2A	266
APOB	264
BDNF	264
Intergenic	263
IGF1	262
CD14	258
RNR1	258
TP53	257
XRCC1	254
CYP2C19	251
IL4	250
FLT3	248
FCGR3A	247
NQO1	247
FCGR2A	245
LEPR	244
SLC26A4	244
intergenic	243
MPO	242
TPMT	238
APOA1	236
HLA-C	233
CYP3A4	232
CYP3A5	230
CETP	228
ADRB1	227
GSTP1	227
LRP5	225
CYBA	224
IFNG	224
TYMS	223

ADRB3	220
ESR2	219
IL4R	218
TLR4	213
REN	212
FGB	206
CYP1B1	205
MMP3	204
CYP17A1	202
PPARA	202
LDLR	200
TNFRSF1B	199
ERCC2	197
COL1A1	193
MC4R	193
VEGFA	193
XRCC3	189
APOA5	186
HFE	183
OPRM1	183
ADIPOQ	181
ITGA2	174
MMP9	174
AGTR2	173
LEP	173
LTA	169
CCR2	167
CDKN2A	167
IRS1	167
LIPC	163
EDN1	161
CYP2E1	160
MAPT	158
TNFRSF11B	157
AR	154
CAT	150
EPHX1	149
APOA4	148
MMP1	148
NOS2A	147

CFTR	145
ATM	143
DRD4	143
ABCA1	141
UCP2	141
FTO	140
PTGS2	139
CCR5	137
CRP	137
DCN	137
MAOA	136
G6PD	135
TNFRSF1A	135
IL2	134
BRCA1	133
MLH1	133
TRNS1	133
GJB3	132
PSEN1	132
NAT2	131
CASR	130
BDKRB2	129
SLC6A3	129
GABRG2	126
HLA	126
CCL2	125
DRD3	124
MBL2	122
P2RX7	122
CYP19A1	121
MTR	121
PKD1	121
PTPN22	121
ALDH2	120
ERCC1	119
HMOX1	119
MC1R	118
NPPA	118
NR3C1	118
DRD2	117

FABP2	116
SELE	116
SPINK1	116
APOC1	115
ITPR2	115
PTGIS	115
A2M	114
CDKN2B	114
FAS	114
FCGR3B	113
ARMS2	112
CAPN10	112
FGGY	112
GP1BA	112
ICAM1	109
EGR2	108
ENPP1	108
IL1R1	107
MMP2	107
MTRR	107
INSR	106
PON2	106
PPARGC1A	105
IL8	104
MSH2	103
GHR	102
CCND1	101
F13A1	101
VKORC1	101
LPA	99
DMPK	98
GNAS	98
ADH1B	95
CXCL12	94
F7	94
GHRL	92
BRCA2	91
HTRA1	91
CYP1A2	90
NPY	90

SULT1A1	90
TAP2	90
POMC	89
NA	88
CBS	88
MYH7	88
WNK4	88
ZBTB38	88
TLR2	87
ALOX5	86
IL13	86
INS	85
ADH1C	84
NBN	84
NOD2	84
NR3C2	84
SERPINA3	84
HHIP	83
HP	83
HSD11B2	83
IGF2	82
MICA	82
MYOC	82
APEX1	81
IL18	80
RAD51	80
CYP4A11	79
IL12B	79
GCH1	78
HLA-DPB1	78
MYBPC3	78
ITGA2B	77
JAK2	77
PTH	77
SCNN1G	77
TH	77
TPH1	77
IL12A	76
KRAS	76
SLC45A2	76

ASIP	74
CYP11B1	74
DHFR	74
IRS2	74
LRP1	74
CHEK2	73
KIR2DL3	73
HBB	72
HNF1A	72
GCK	71
GRIN1	71
HSPA1B	71
KIR2DL1	71
TCF1	71
C3	70
GSTM3	70
HTR2C	70
SCNN1B	70
SLCO1B1	70
TNNT2	70
CFB	69
SCN1A	69
UCP3	69
COL2A1	68
EGF	68
KCNJ11	68
CMA1	67
FLG	67
GABRA1	67
HMGA2	67
KLK1	67
SELP	67
ABCG2	66
ASPN	66
BCHE	66
FMO3	66
NPHS2	66
RHO	66
CALCR	65
IGFBP3	65

RUNX1	65
ADRA1A	64
ADRA2A	64
MTTP	64
PCSK9	64
HSD11B1	63
CDK6	62
CYP46A1	62
DIO2	62
EFEMP1	62
IDE	62
IL6R	62
KIR3DL1	62
KL	62
CALCA	61
CCL5	61
CX3CR1	61
KIR2DS2	61
NPR1	61
SERPINA1	61
ACE2	60
APP	60
KIT	60
TNFRSF11A	60
ADA	59
DBH	59
INSIG2	59
KIR2DL5A	59
KIR2DS5	59
NRG1	59
TSHR	59
ABCG5	58
ERCC5	58
GDF5	58
KIR2DS4	58
MEFV	58
POLG	58
ALOX5AP	57
APC	57
BCR	57

HMGCR	57
KCNJ10	57
KIR3DS1	57
UCP1	57
ADAMTSL3	56
APOH	56
CDKN1A	56
HSPA1L	56
NOTCH1	56
PTPN11	56
SLC12A3	56
SLC26A5	56
ACSM3	55
ADD2	55
DMD	55
FCGR2B	55
FGFR2	55
KIR3DL2	55
PLAT	55
UGT1A7	55
MT-ND1	54
PLAU	54
SRD5A2	54
C2	53
FRZB	53
GPR126	53
KCNE1	53
PRSS1	53
TIMP1	53
GPX1	52
HSPA1A	52
PDCD1	52
TNFSF11	52
ADRA2B	51
ALDOA	51
BRAF	51
CST3	51
CYP2J2	51
FOLH1	51
IL15	51

KCNQ1	51
KIR2DL4	51
KIR3DL3	51
TAP1	51
ABCG8	50
ABO	50
CTSD	50
CYP4B1	50
GSTM4	50
MC3R	50
MSH3	50
MT-ND2	50
MT-ND4	50
XRCC2	50
AHSG	49
BCL2	49
BMPR2	49
FGFR3	49
NPR3	49
SAA1	49
UGT1A6	49
DTNBP1	48
FMR1	48
HMGA1	48
KIR2DP1	48
MT-ND6	48
P2RY12	48
PLAG1	48
PSEN2	48
PSRC1	48
TOR1A	48
VCAM1	48
ERBB2	47
GJA8	47
KIR2DL2	47
MLL	47
MTHFD1	47
MYH9	47
PPARD	47
BAX	46

CRYGA	46
CRYGB	46
NAT1	46
NFKBIA	46
RGS2	46
ACAN	45
BMP2	45
DPYD	45
IL21R	45
MDM2	45
MGMT	45
RUNX2	45
SPP1	45
ABCA4	44
GRK4	44
HLA-DRB4	44
OGG1	44
PLIN	44
SERPINC1	44
SLC12A1	44
SLC19A1	44
TWIST1	44
AMPD1	43
ERCC4	43
GJA4	43
HTR1B	43
IHH	43
LCT	43
LIPE	43
MT-TL1	43
PLN	43
TCF7L2	43
UGT2B15	43
CDH13	42
EDNRA	42
PGR	42
PLA2G7	42
RP1	42
RPGR	42
SOD1	42

TERT	42
TGFB3	42
ADH7	41
ADIPOR1	41
CRH	41
EDNRB	41
GH1	41
MECP2	41
PSMB8	41
SOST	41
SREBF1	41
TGFBR1	41
UGT1A9	41
ADRA2C	40
APOC2	40
CDKN1B	40
COL1A2	40
DEFB1	40
DLEU7	40
FASLG	40
FGA	40
HIST1H1D	40
IL10RA	40
KCNH2	40
LCORL	40
PARP1	40
PRNP	40
UQCC	40
CNR1	39
DRD1	39
HSD3B1	39
HTR1A	39
IRF4	39
KIR2DS1	39
PTCH1	39
PTPN1	39
RET	39
RGS4	39
SCN5A	39
TRNK	39

TRNL1	39
ARL11	38
BACE1	38
CYP11A1	38
EGFR	38
HNMT	38
IGF1R	38
IL6ST	38
MS4A2	38
VWF	38
ABCB11	37
AHR	37
CASP8	37
CLU	37
GC	37
GHRHR	37
SCARB1	37
THBD	37
CCDC26	36
F8	36
FBN1	36

Table nine shows the first five hundred phenotypes with the most gene associations in OGA.

Table 9 Top five hundred phenotypes with the most gene associations in OGA

HOP Concept	Phenotype name	Gene count
HP:0002511	alzheimer disease	2433
HP:0100753	Schizophrenia	1816
HP:0006718	increased risk of colorectal cancer	1581
HP:0100817	renovascular hypertension	1251
HP:0002092	pulmonary hypertension	1239
HP:0002640	hypertension associated with pheochromocytoma	1237
HP:0008741	hypertension due to renal artery hyperplasia	1233
HP:0000822	Hypertension	1228

HP:0003002	breast carcinoma	1211
HP:0002099	Asthma	911
HP:0011001	increased bone mineral density	798
HP:0001370	rheumatoid arthritis	687
HP:0002636	arterial aneurysm of celiac and mesenteric vessels	648
HP:0001658	myocardial infarction	643
HP:0001513	Obesity	641
HP:0002642	arteriovenous fistulas of celiac and mesenteric vessels	637
HP:0002608	celiac disease	636
HP:0002626	venous varicosities of celiac and mesenteric vessels	636
HP:0004761	post-angioplasty coronary artery restenosis	631
HP:0001717	coronary artery calcification	547
HP:0005181	premature coronary artery disease	543
HP:0005145	coronary artery stenosis	542
HP:0001677	coronary artery disease	541
HP:0007302	bipolar affective disorder	538
HP:0000938	Osteopenia	503
HP:0005686	patchy osteosclerosis	503
HP:0004348	abnormality of bone mineral density	501
HP:0000717	Autism	495
HP:0000753	autism with high cognitive abilities	495
HP:0006510	chronic obstructive pulmonary disease	420
HP:0005517	t-cell lymphoma/leukemia	406
HP:0005539	t-cell chronic lymphocytic lymphoma/leukemia	406
HP:0004322	short stature	392
HP:0000002	abnormality of body height	389
HP:0000098	tall stature	389
HP:0003518	adult male height 142-169 cm	389
HP:0003512	adult female height 130-157 cm	388
HP:0004570	increased vertebral height	388
HP:0004581	increased anterior vertebral height	388
HP:0008871	height less than 3rd percentile	388
HP:0000716	Depression	376
HP:0001025	Urticaria	336
HP:0004808	acute myeloid leukemia	332
HP:0001626	abnormality of the cardiovascular system	314
HP:0011025	abnormality of cardiovascular system physiology	314
HP:0005550	chronic lymphatic leukemia	298
HP:0004563	increased spinal bone density	295
HP:0005700	increased bone density with cystic changes	295

HP:0007417	discoid lupus erythematosus	290
HP:0004812	human pre-b-cell acute lymphoblastic leukemia	281
HP:0004848	ph-positive acute lymphoblastic leukemia	281
HP:0006721	acute lymphatic leukemia	281
HP:0006727	t-cell acute lymphoblastic leukemias	281
HP:0000704	Periodontitis	279
HP:0002488	acute leukemia	279
HP:0004836	acute promyelocytic leukemia	279
HP:0005506	chronic myelogenous leukemia	279
HP:0001909	Leukemia	278
HP:0005086	knee osteoarthritis	275
HP:0004929	coronary atherosclerosis	271
HP:0000608	macular degeneration	269
HP:0003765	Psoriasis	269
HP:0007673	cystic macular degeneration	269
HP:0007694	pigmented macular degeneration	269
HP:0007868	senile macular degeneration	269
HP:0008028	cystoid macular degeneration	269
HP:0002758	Osteoarthritis	262
HP:0003141	hyperbetalipoproteinemia	248
HP:0000166	severe periodontitis	239
HP:0011058	generalized periodontitis	239
HP:0011059	localized periodontitis	239
HP:0003362	increased circulating very-low-density lipoprotein cholesterol	230
HP:0002565	complex cardiac malformations	221
HP:0002155	Hypertriglyceridemia	220
HP:0008332	mild hypertriglyceridemia	220
HP:0002037	inflammatory bowel disease	218
HP:0005978	noninsulin-dependent diabetes mellitus	217
HP:0003233	decreased circulating high-density lipoprotein cholesterol	214
HP:0003124	Hypercholesterolemia	203
HP:0002861	malignant melanoma	201
HP:0007018	attention deficit hyperactivity disorder	201
HP:0007716	malignant intraocular melanoma	196
HP:0100279	ulcerative colitis	186
HP:0002140	ischemic stroke	182
HP:0001997	Gout	180
HP:0008538	sensorineural deafness	177
HP:0008615	late onset sensorineural deafness	177
HP:0008620	congenital sensorineural deafness	177

HP:0001727	thromboembolic stroke may occur	176
HP:0008540	bilateral congenital sensorineural deafness	175
HP:0001297	Stroke	172
HP:0001535	poor weight gain	172
HP:0002401	stroke-like episodes	172
HP:0005292	intimal thickening in the coronary arteries	169
HP:0006679	granulomatous coronary arteritis	169
HP:0006704	abnormality of the coronary arteries	169
HP:0002894	neoplasm of the pancreas	166
HP:0006753	neoplasm of the stomach	165
HP:0002896	neoplasm of the liver	164
HP:0005744	generalized osteoporosis with pathologic fractures	156
HP:0000855	insulin resistance	155
HP:0000939	osteoporosis	154
HP:0005625	osteoporosis of vertebrae	154
HP:0005653	moderate generalized osteoporosis	154
HP:0005897	severe osteoporosis	154
HP:0000877	insulin-resistant diabetes mellitus at puberty	151
HP:0001369	Arthritis	150
HP:0003095	septic arthritis	150
HP:0005059	arthralgia/arthritis	150
HP:0005764	polyarticular arthritis	150
HP:0000857	neonatal insulin-dependent diabetes mellitus	144
HP:0000819	diabetes mellitus	143
HP:0000831	insulin-resistant diabetes mellitus	142
HP:0010979	abnormality of the level of lipoprotein cholesterol	142
HP:0001657	prolonged qt interval on ekg	139
HP:0000408	progressive sensorineural hearing impairment	137
HP:0008527	congenital sensorineural hearing impairment	137
HP:0008603	congenital severe sensorineural hearing impairment	137
HP:0008610	infantile sensorineural hearing impairment	137
HP:0008619	bilateral sensorineural hearing impairment	137
HP:0008625	severe sensorineural hearing impairment	137
HP:0000365	hearing impairment	135
HP:0000405	conductive hearing impairment	135
HP:0001730	progressive hearing impairment	135
HP:0005101	high-frequency hearing impairment	135
HP:0008587	mild neurosensory hearing impairment	135
HP:0008598	mild conductive hearing impairment	135
HP:0008574	severe early sensorineural hearing loss	132

HP:0008617	progressive bilateral sensorineural hearing loss	132
HP:0009726	renal neoplasm	132
HP:0000410	mixed hearing loss	130
HP:0001643	patent ductus arteriosus	130
HP:0001729	congenital hearing loss	130
HP:0008536	bilateral conductive hearing loss	130
HP:0008542	low-frequency hearing loss	130
HP:0008581	early conductive hearing loss	130
HP:0008584	progressive high-frequency hearing loss	130
HP:0008590	progressive childhood hearing loss	130
HP:0008591	congenital conductive hearing loss	130
HP:0001033	facial flushing after alcohol intake	129
HP:0002665	lymphoma	128
HP:0005305	cerebral venous thrombosis	119
HP:0001639	hypertrophic cardiomyopathy	117
HP:0005157	concentric hypertrophic cardiomyopathy	117
HP:0008205	insulin-dependent but ketosis-resistant diabetes	116
HP:0004904	insulin-dependent maturity-onset diabetes of the young	115
HP:0003464	abnormal cholesterol homeostasis	114
HP:0005117	elevated diastolic blood pressure	114
HP:0002625	deep venous thrombosis	112
HP:0003107	abnormality of cholesterol metabolism	112
HP:0003146	hypocholesterolemia	112
HP:0003349	low cholesterol esterification rates	112
HP:0003462	elevated 8-dehydrocholesterol	112
HP:0004850	recurrent deep vein thrombosis	112
HP:0010569	elevated 7-dehydrocholesterol	112
HP:0000825	hyperinsulinemic hypoglycemia	111
HP:0004936	venous thrombosis	111
HP:0002077	migraine with aura	109
HP:0000842	hyperinsulinemia	108
HP:0001644	dilated cardiomyopathy	107
HP:0005159	dilated cardiomyopathy may be present	107
HP:0008189	insulin insensitivity	107
HP:0004421	elevated systolic blood pressure	105
HP:0001635	congestive heart failure	103
HP:0001665	abnormality of cardiac conduction	103
HP:0002083	migraine without aura	103
HP:0001722	high-output congestive heart failure	101
HP:0009805	low-output congestive heart failure	101

HP:0001638	cardiomyopathy	99
HP:0001723	restrictive cardiomyopathy	99
HP:0005152	oncocytic cardiomyopathy	99
HP:0000751	personality changes	95
HP:0002076	Migraine	94
HP:0005130	restrictive heart failure	94
HP:0002615	hypotension	91
HP:0001275	Epilepsy	90
HP:0002621	atherosclerosis	90
HP:0004416	precocious atherosclerosis	90
HP:0004943	accelerated atherosclerosis	90
HP:0007201	cerebral artery atherosclerosis	90
HP:0003003	colon cancer	88
HP:0100806	Sepsis	88
HP:0004420	arterial thrombosis	86
HP:0001735	acute pancreatitis	84
HP:0002632	low-to-normal blood pressure	84
HP:0001518	low birth weight	81
HP:0006280	chronic pancreatitis	81
HP:0001733	pancreatitis	79
HP:0001977	thrombosis	79
HP:0002641	peripheral thrombosis	79
HP:0004419	recurrent thrombophlebitis	79
HP:0005236	chronic calcifying pancreatitis	79
HP:0007854	glaucomatous visual field defects	79
HP:0008007	primary congenital glaucoma	79
HP:0100027	recurrent pancreatitis	79
HP:0000501	Glaucoma	78
HP:0001087	congenital glaucoma	78
HP:0006775	multiple myeloma	78
HP:0008041	late onset congenital glaucoma	78
HP:0007354	amyotrophic lateral sclerosis	77
HP:0001622	premature birth	75
HP:0001520	macrosomia	73
HP:0002452	cerebrovascular accident	70
HP:0008849	low birth weight in males	70
HP:0000722	obsessive-compulsive disorder	66
HP:0004325	decreased body weight	65
HP:0100647	graves disease	65
HP:0001045	Vitiligo	61

HP:0005602	progressive vitiligo	61
HP:0004749	atrial fibrillation or flutter	59
HP:0005179	atrial fibrillation may occur	59
HP:0001712	left ventricular hypertrophy	58
HP:0000794	iga nephropathy	56
HP:0002145	frontotemporal dementia	56
HP:0004953	abdominal aortic aneurysm	56
HP:0000787	kidney stones	52
HP:0001402	hepatocellular carcinoma	52
HP:0001824	weight loss	52
HP:0002613	biliary cirrhosis	51
HP:0010301	spinal dysraphism	51
HP:0010919	abnormality of homocysteine metabolism	51
HP:0008504	moderate sensorineural hearing impairment	50
HP:0000726	Dementia	49
HP:0000727	frontal lobe dementia	49
HP:0002439	frontolimbic dementia	49
HP:0003193	allergic rhinitis	49
HP:0000981	discoid lupus in carriers or adults with mild disease	47
HP:0007204	brain imaging shows diffuse white matter abnormalities	47
HP:0007645	retinitis pigmentosa type i	47
HP:0008035	retinitis pigmentosa inversa	47
HP:0000367	conductive deafness	45
HP:0000399	prelingual sensorineural deafness	45
HP:0000404	Deafness	45
HP:0000510	retinitis pigmentosa	45
HP:0000833	glucose intolerance	45
HP:0001127	progressive retinitis pigmentosa	45
HP:0001728	congenital deafness	45
HP:0001731	prelingual deafness	45
HP:0001757	high-tone sensorineural deafness	45
HP:0007826	atypical retinitis pigmentosa	45
HP:0007947	pericentral retinitis pigmentosa	45
HP:0008506	central retinitis pigmentosa	45
HP:0008513	bilateral conductive deafness	45
HP:0008522	high-frequency deafness	45
HP:0008525	congenital conductive deafness	45
HP:0008530	bilateral sensorineural deafness	45
HP:0008596	progressive postlingual sensorineural	45
HP:0008607	progressive conductive deafness	45

HP:0009900	unilateral deafness	45
HP:0100651	diabetes mellitus type i	45
HP:0100652	diabetes mellitus type ii	45
HP:0003265	neonatal hyperbilirubinemia	44
HP:0004924	abnormal oral glucose tolerance	44
HP:0001952	abnormal glucose tolerance	43
HP:0004754	paroxysmal or chronic atrial fibrillation	43
HP:0004757	paroxysmal atrial fibrillation	43
HP:0005110	atrial fibrillation	43
HP:0008176	neonatal unconjugated hyperbilirubinemia	43
HP:0008651	uric acid urolithiasis independent of gout	43
HP:0002251	congenital megacolon	42
HP:0000100	nephrotic syndrome	41
HP:0000113	polycystic kidney dysplasia	41
HP:0001047	atopic dermatitis	41
HP:0001511	intrauterine growth restriction	41
HP:0002564	cardiac malformation	41
HP:0002904	hyperbilirubinemia	41
HP:0002908	conjugated hyperbilirubinemia	41
HP:0007533	severe atopic dermatitis	41
HP:0007573	late onset atopic dermatitis	41
HP:0008282	unconjugated hyperbilirubinemia	41
HP:0100021	cerebral paralysis	41
HP:0000112	nephropathy	40
HP:0008654	tubulointerstitial nephropathy	40
HP:0001649	tachycardia	39
HP:0003140	t-wave inversion in the right precordial leads and late potentials in signal-averaging ecg	39
HP:0008638	terminal nephrotic syndrome	39
HP:0008695	transient nephrotic syndrome	39
HP:0008727	idiopathic nephrotic syndrome	39
HP:0010885	aseptic necrosis	39
HP:0002120	cerebral cortical atrophy	38
HP:0006523	pulmonary fibrosis due to recurrent infections	38
HP:0006672	congenital heart block	38
HP:0001654	abnormality of the heart valves	37
HP:0001668	heart block	37
HP:0001709	complete heart block	37
HP:0001710	conotruncal defect	37
HP:0001721	irregular heart beat	37

HP:0002206	pulmonary fibrosis	37
HP:0002419	molar tooth sign on mri	37
HP:0005170	complete heart block with broad rs complexes	37
HP:0005178	complete heart block with narrow qrs complexes	37
HP:0007007	cavitation of the basal ganglia	37
HP:0007257	imaging shows signal abnormalities in basal ganglia	37
HP:0000027	azoospermia	36
HP:0001081	cholelithiasis	36
HP:0001394	Cirrhosis	36
HP:0001413	micronodular cirrhosis	36
HP:0002377	cranial nerve palsies can arise with head and neck paragangliomas	36
HP:0006577	macronodular cirrhosis	36
HP:0008209	premature ovarian failure	36
HP:0008255	transient neonatal diabetes mellitus	36
HP:0011005	mixed cirrhosis	36
HP:0000545	Myopia	35
HP:0001110	progressive myopia	35
HP:0003126	low-molecular-weight proteinuria	35
HP:0008711	benign prostatic hyperplasia	35
HP:0011003	severe myopia	35
HP:0000862	addison disease	34
HP:0002229	alopecia areata	34
HP:0002668	paragangliomas	34
HP:0002864	paraganglioma of head and neck	34
HP:0004908	mild diabetes mellitus	34
HP:0001397	hepatic steatosis	33
HP:0002963	abnormal delayed hypersensitivity skin test	33
HP:0002972	reduced delayed hypersensitivity	33
HP:0005299	premature peripheral vascular disease	33
HP:0005309	peripheral vascular insufficiency	33
HP:0005427	lack of delayed skin hypersensitivity reaction	33
HP:0005530	high molecular weight kininogen deficiency	33
HP:0006516	hypersensitivity pneumonitis	33
HP:0100326	immunologic hypersensitivity	33
HP:0001251	Ataxia	32
HP:0004323	abnormality of body weight	32
HP:0004324	increased body weight	32
HP:0008245	tsh deficient hypothyroidism	32
HP:0000361	pulsatile tinnitus (tympanic paraganglioma)	31
HP:0001263	developmental delay	31

HP:0001917	renal amyloidosis	31
HP:0002886	vagal paraganglioma	31
HP:0003563	hypobetalipoproteinemia	31
HP:0004944	cerebral aneurysm	31
HP:0005381	recurrent meningococcal disease	31
HP:0006715	glomus tympanicum paraganglioma	31
HP:0100635	carotid paraganglioma	31
HP:0001854	gout (feet)	30
HP:0008158	hyperapobetalipoproteinemia	30
HP:0008181	abetalipoproteinemia	30
HP:0010980	hyperlipoproteinemia	30
HP:0010981	hypolipoproteinemia	30
HP:0000122	unilateral renal agenesis	29
HP:0001004	lymphedema	29
HP:0001762	talipes equinovarus	29
HP:0002138	subarachnoid hemorrhage	29
HP:0003293	elevated liver enzymes	29
HP:0003605	onset of lymphedema around puberty	29
HP:0004788	intestinal lymphedema	29
HP:0006530	interstitial pulmonary disease	29
HP:0010522	Dyslexia	29
HP:0100636	pulmonary paraglioma	29
HP:0009733	Glioma	28
HP:0009734	optic glioma	28
HP:0010795	cerebellar glioma	28
HP:0010796	brainstem glioma	28
HP:0000003	multicystic kidney dysplasia	27
HP:0000047	Hypospadias	27
HP:0000051	perineal hypospadias	27
HP:0000792	kidney malformation	27
HP:0000807	glandular hypospadias	27
HP:0000808	penoscrotal hypospadias	27
HP:0003244	penile hypospadias	27
HP:0008743	coronal hypospadias	27
HP:0002592	gastric ulcer	26
HP:0002650	Scoliosis	26
HP:0002751	Kyphoscoliosis	26
HP:0002943	thoracic scoliosis	26
HP:0002944	thoracolumbar scoliosis	26
HP:0003423	thoracolumbar kyphoscoliosis	26

HP:0004615	mild thoracic scoliosis	26
HP:0004619	lumbar kyphoscoliosis	26
HP:0004626	lumbar scoliosis	26
HP:0005659	thoracic kyphoscoliosis	26
HP:0008453	congenital kyphoscoliosis	26
HP:0001249	intellectual disability	25
HP:0002111	restrictive respiratory insufficiency	25
HP:0005743	avascular necrosis of the capital femoral epiphysis	25
HP:0005952	decreased pulmonary function	25
HP:0006520	progressive pulmonary function impairment	25
HP:0007154	nonprogressive intellectual disability	25
HP:0008222	female infertility	25
HP:0010637	conjunctival amyloidosis	25
HP:0011034	Amyloidosis	25
HP:0100292	amyloidosis of peripheral nerves	25
HP:0100742	vascular neoplasia	25
HP:0001410	decreased liver function	24
HP:0001411	abnormal liver function tests	24
HP:0001776	bilateral club feet	24
HP:0003156	increased liver function tests	24
HP:0005317	increased pulmonary vascular resistance	24
HP:0005520	chronic disseminated intravascular coagulation	24
HP:0006576	hepatic vascular malformations	24
HP:0006578	subclinical abnormal liver function tests	24
HP:0007797	retinal vascular malformations	24
HP:0008846	severe intrauterine growth retardation	24
HP:0008883	mild intrauterine growth retardation	24
HP:0100805	precocious menopause	24
HP:0002156	Homocystinuria	23
HP:0002597	abnormality of the vasculature	23
HP:0003562	abnormal metaphyseal vascular invasion	23
HP:0004746	membranoproliferative glomerulonephritis type ii	23
HP:0004934	vascular calcifications	23
HP:0004948	vascular tortuosity	23
HP:0005296	occlusive vascular disease	23
HP:0005297	premature occlusive vascular disease	23
HP:0005521	disseminated intravascular coagulation	23
HP:0007768	central retinal vessel vascular tortuosity	23
HP:0007850	retinal vascular proliferation	23
HP:0010775	vascular ring	23

HP:0004736	ectopic kidney with fusion	22
HP:0005100	premature birth following premature rupture of fetal membranes	22
HP:0005507	alpha-thalassemia with microcytosis	22
HP:0005537	decreased mean platelet volume	22
HP:0005582	tubulointerstitial medullary cystic kidney disease	22
HP:0011014	abnormal glucose homeostasis	22
HP:0100611	hypoplastic glomerulocystic kidney disease	22
HP:0000851	congenital hypothyroidism	21
HP:0001904	autoimmune neutropenia	21
HP:0003006	Neuroblastoma	21
HP:0003077	Hyperlipidemia	21
HP:0003540	abnormal platelet aggregation	21
HP:0005560	Thalassemia	21
HP:0006742	congenital neuroblastoma	21
HP:0006747	Ganglioneuroblastoma	21
HP:0006768	localized neuroblastoma	21
HP:0008159	variable hyperlipidemia	21
HP:0008279	transient hyperlipidemia	21
HP:0008356	combined hyperlipidemia	21
HP:0001645	sudden cardiac death	20
HP:0002373	febrile seizures	20
HP:0005161	premature sudden cardiac death	20
HP:0005549	congenital neutropenia	20
HP:0008177	abnormal platelet aggregation in response to various agents	20
HP:0008320	reduced platelet aggregation response to collagen and thrombin	20
HP:0001875	Neutropenia	19
HP:0001973	immune thrombocytopenia	19
HP:0002661	painless fractures due to injury	19
HP:0003229	prolonged whole-blood clotting time in severe hemophilia	19
HP:0004295	abnormality of the gastric mucosa	19
HP:0004809	neonatal alloimmune thrombocytopenia	19
HP:0005518	erythrocyte macrocytosis	19
HP:0006566	neonatal cholestatic liver disease	19
HP:0007596	painful trunk and limb subcutaneous lipomas	19
HP:0010833	spontaneous pain sensation	19
HP:0100839	hepatic agenesis	19
HP:0000821	Hypothyroidism	18
HP:0000832	primary hypothyroidism	18
HP:0001286	low intelligence	18
HP:0003131	Cystinuria	18

HP:0003251	male infertility	18
HP:0003419	low back pain	18
HP:0003573	increased total bilirubin	18
HP:0006557	polycystic liver disease	18
HP:0007141	sensorimotor neuropathy	18
HP:0007180	low normal intelligence	18
HP:0007423	skin inflammation	18
HP:0008148	absent platelet aggregation response to epinephrine	18
HP:0008168	increased total bilirubin may occur	18
HP:0008191	athyroidal hypothyroidism	18
HP:0008223	compensated hypothyroidism	18
HP:0008237	hypothalamic hypothyroidism	18
HP:0008254	neonatal primary hyperparathyroidism	18
HP:0100662	Chondritis	18
HP:0200047	chrondritis of pinna	18
HP:0000131	uterine leiomyoma	17
HP:0000554	Uveitis	17
HP:0001271	Polyneuropathy	17
HP:0001301	chronic sensorineural polyneuropathy	17

Motivating examples

In order to showcase the OGA functionality we have identified four motivating examples.

Colon cancer and *Helicobacter pylori* infection susceptibility

The colon cancer (HP:0003003) and *Helicobacter pylori* infection susceptibility (HP:0005202) concepts were placed in the selected list during an OGA session. The OGA analysis was conducted and one gene was found to be common to both concepts, CYP2C19. CYP2C19 is the gene symbol for the Cytochrome P450, family 2, subfamily C, polypeptide 19 gene. The simulation to find the empirical *p* value was repeated five times and the following *p* values were found: 0.3333, 0.5, 0.5, 0.25, 0.333. This is not

surprising since the chances of finding a single gene common for two concepts are expected to be high. However, it is still intriguing that these two seemly unrelated concepts share this association. The relationship between *H. pylori* and colon cancer is very well documented (Strofilas et al., 2012). It seems clear that *H. pylori* infection is a risk factor for colon cancer. In addition, CYP2C19 has a gain-of-function polymorphism associated with peptic ulcer disease (Musumba et al., 2013). The association between CYP2C19 and *H. pylori* maybe related to a polymorphism that affects the proton pump inhibitors (PPI) and, that makes some patients respond poorly to treatment with omeprazole (O'Donoghue, 2011). Furthermore, the relationship between CYP2C19 and cancer is equally well documented (Yamamoto et al., 2013). This observation could be completely spurious, especially because the empirical *p* value was not significant. Nonetheless, the observation provokes some interesting questions: what is the relationship between CYP2C19 and colon cancer specifically? What is the relationship between this gene, bleeding and the sensibility to *H. pylori* infection? Should CYP2C19 polymorphisms be included in the genetic risk calculations for colon cancer? A pubmed search for the terms “colon cancer CYP2C19 and H. pylori” returned no results. Therefore, there seem not to be evidence in either direction of the relationship between colon cancer and CYP2C19, this presents a prime opportunity for the exploration of GWAS databases for colon cancer and CYP2C19 genotypes.

Lipid metabolism, diabetes, obesity, and hypertension

Lipid metabolism abnormalities, diabetes, obesity and hypertension are four clearly related afflictions. They have been the recipients of special attention in the last

decades, because of their role in public health in developed countries, especially the United States. Therefore, we selected these concepts in OGA: Abnormality of glycolipid metabolism (HP:0010969), Abnormal glucose homeostasis (HP:0011014), Systemic hypertension (HP:0000822) and Obesity (HP:0011014) in order to find the common associations to these three phenotypes. Nine genes were found and the simulations for these finding resulted in an empirical p value < 0.001.

Table 10 Genes related to obesity, diabetes, hypertension and lipid metabolism

Gene Symbol	Gene Name	Comment (associations according to OMIM)
APOE	Apolipoprotein E	Alzheimer disease-2, Hyperlipoproteinemia, type III, Myocardial infarction susceptibility
ACE	Angiotensin I-converting enzyme	Myocardial infarction susceptibility, Alzheimer disease, Stroke
CETP	Cholesteryl Ester Transfer protein	Hyperalphalipoproteinemia
AGT	Angiotensinogen	Hypertension
IL6	Interleukin 6	Diabetes
FGB	Fibrinogen B, Beta polypeptide	None
PON1	Paraoxonase 1	Coronary artery disease, Microvascular complications of diabetes
LPL	Lipoprotein lipase	Combined hyperlipidemia, familial
MTHFR	5,10-Methylenetetrahydrofolate	Vascular disease, Schizophrenia

reductase

It is possible that the common association of these nine genes to the four phenotype lineages is just the result of the same genes been preferentially selected for association studies based on previous reports, a self feeding loop. This is certainly the case for the candidate gene approach, when hypotheses were generated while compiling a list of genes to be included in a study; however this is no longer true for GWAS when all the associations are hypothesis free and they are corrected and adjusted for multiple testing. As more hypothesis free results become available that effect should dissipate. Furthermore, it would still be interesting to investigate more carefully if there is actually a biological phenomenon that makes these particular genes constantly reappear in the analysis. It is particularly interesting to observe two of these genes associated to Alzheimer ’s disease (AD), because there have been some recent reports relating obesity and diabetes to an increased risk of AD (Vignini et al., 2013). These genes would be prime candidates for meta-analysis of GWAS studies that include these phenotypes and risk models combining these nine genes. In addition to the relationship found by OGA, it is also possible to try to reconstruct other types of relationships that these genes may have using a program such as Cytoscape (Shannon et al., 2003).

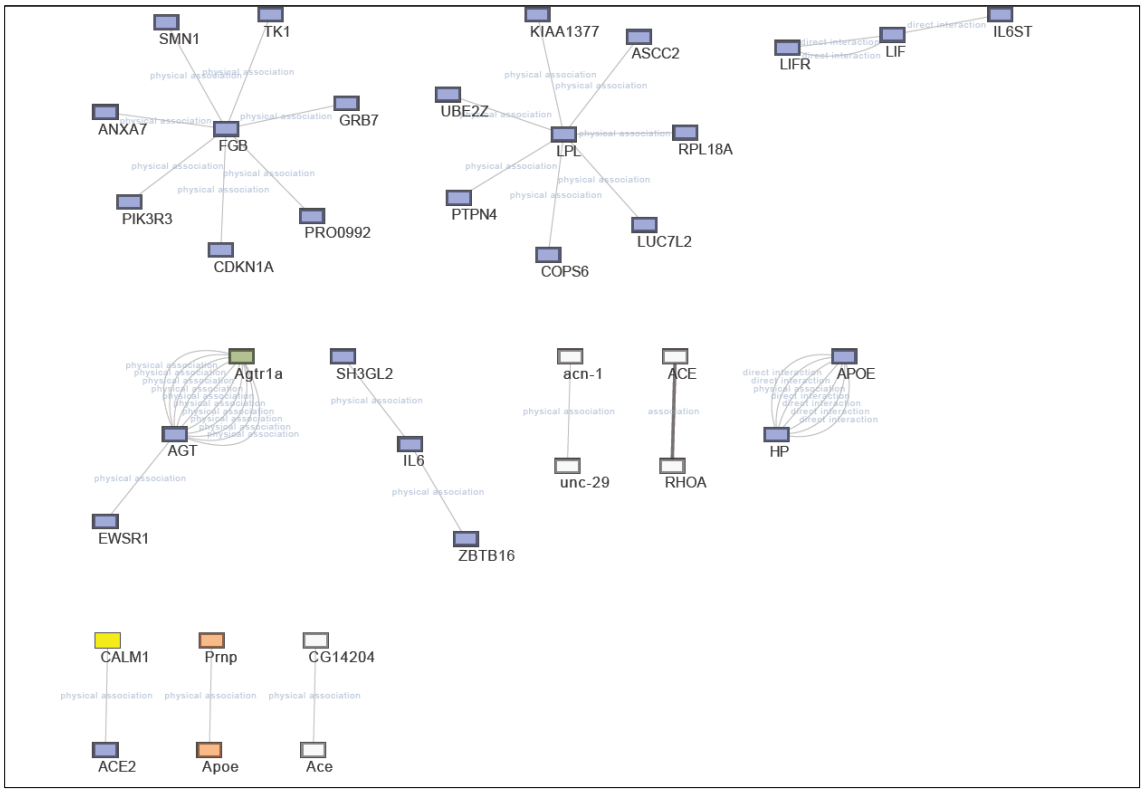

Figure 13 Cytoscape network for genes related to diabetes, obesity, lipid metabolism and hypertension

Schizophrenia, bulimia, depression and psychosis

The same example used during the preliminary phase was repeated with the current implementation of OGA. This time the concepts Psychosis (HP:0000709), Bulimia (HP:0100739), Depression (HP:0000716), and Schizophrenia (HP:0100753) were used in the analysis. The analysis in the preliminary phase took several weeks to complete. The analysis process included downloading, parsing and formatting the databases and writing scripts to perform the queries. OGA allowed the same type of analysis in just minutes. Furthermore, three genes were found that are common to all four phenotypes with an empirical *p* value < 0.001.

Table 11 Genes related to four neurological disorders

Gene Symbol	Gene Name	Comment (OMIM)
HTR2A	HTR2A 5-hydroxytryptamine (serotonin) receptor 2A, G protein-coupled	A neurotransmitter associated with depression, schizophrenia, anorexia
SLC6A3	Solute carrier family 6 (neurotransmitter transporter, dopamine), member 3	Eating disorders, attention deficit-hyperactivity disorder, Major affective disorder 1
SLC6A4	Solute carrier family 6 (neurotransmitter transporter, dopamine), member 3	Anxiety, Obsessive-compulsive disorder

The same gene HTR2A was found again as common to all these phenotypes. The importance of this neurotransmitter in relation neurological disorders seems clear (Gu et al., 2013; Jin, Xu, Yuan, Wang, & Cheng, 2013; Lohoff, et al., 2012). This gene seems to be a prime target for pharmacological studies, something that maybe well known to experts in field, but that OGA has made very easy to ascertain for everyone. SLC6A3 and SLC6A4 are both neurotransmitters as well. These genes have been associated with eating disorders and several neurological abnormalities (Ho, et al., 2012; Kang, et al., 2013; Lohoff, et al., 2012; Papenberg et al., 2013; Spencer et al., 2012).

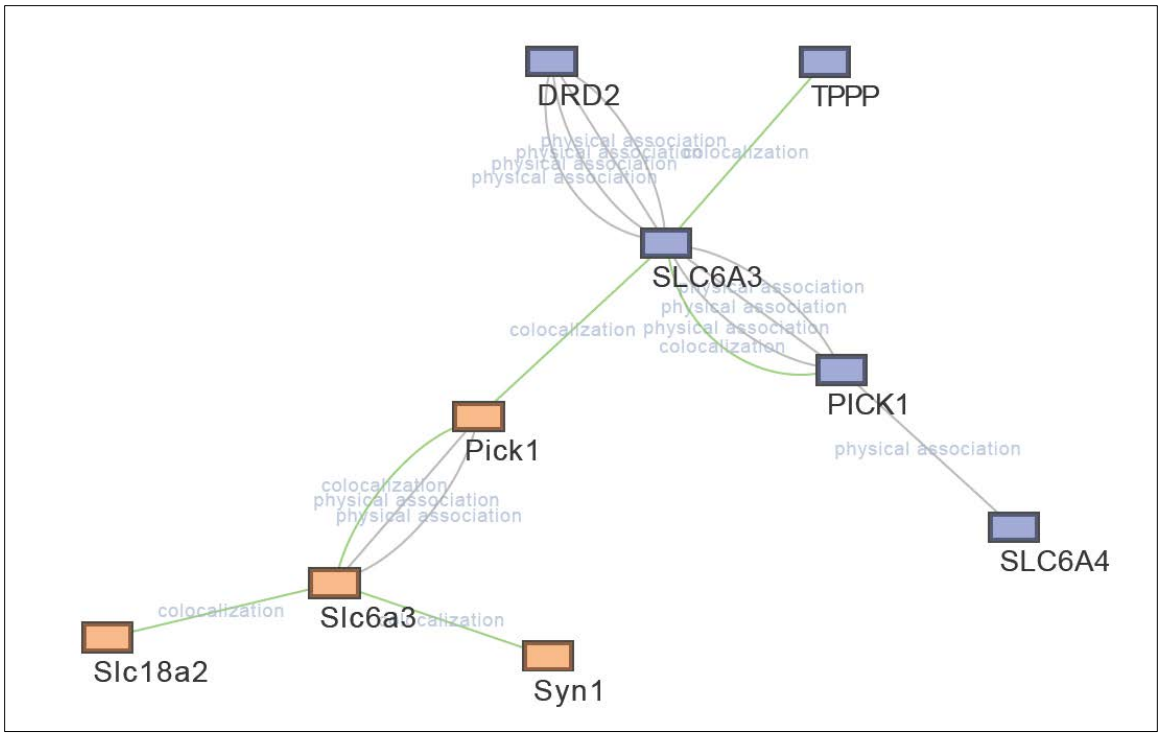

Figure 14 Cytoscape network for genes related to neurological disorders

Autism and Cerebral palsy

The last motivating example is a replication of the use case introduced by the Neurocarta publication. For this purpose we selected the concepts Autism (HP:0000717) and Cerebral palsy (HP:0100021). We could not include the third concept in the Neurocarta use case, Fetal Alcohol Spectrum Disorder (FASD), because this concept is not included in HOP. The concept FASD is not expected to be included in HOP, because it is an affliction caused by environmental conditions and not a phenotype expected to have a significantly heritable component. OGA found seven genes linked between these two concepts: PTGS2, APOE, SERPINE1, TFPI, ITGB3, TNF, and MTHFR. The empirical *p* value for this intersection was *p* value < 0.5. The non significant *p* value may

be a strong indication that these genes were found simply by chance. However, the ability to identify the genes common to these afflictions can still be important for potential meta-analysis or for the design of future studies.

Table 12 Genes associated with autism and cerebral palsy

Gene Symbol	Gene Name	Comment (OMIM)
PTGS2	prostaglandin-endoperoxide synthase 2 (prostaglandin G/H synthase and cyclooxygenase)	Prostaglandin synthesis
APOE	Apolipoprotein E	Alzheimer disease-2, Hyperlipoproteinemia, type III, Myocardial infarction susceptibility
SERPINE1	Serpin peptidase inhibitor, clade E (nexin, plasminogen activator inhibitor type 1), member 1	Plasminogen activator inhibitor-1 deficiency 613329 Modulator of transcription of plasminogen activator inhibitor
TFPI	Tissue factor pathway inhibitor (lipoprotein-associated coagulation inhibitor)	Also known as lipoprotein-associated coagulation inhibitor
ITGB3	Integrin, beta 3 (platelet glycoprotein IIIa, antigen CD61)	Glanzmann thrombasthenia, purpura posttransfusion, thrombocytopenia,

		neonatal alloimmune , susceptibility to Myocardial infarction.
TNF	Tumor necrosis factor	Asthma, cardiovascular disease
MTHFR	5,10-Methylenetetrahydrofolate reductase	Vascular disease, Schizophrenia

CHAPTER SIX: DISCUSSION

OGA not only allows for the easy navigation of the phenotype to gene association ontology, but it also allows for the easy generation of statistics that can be helpful in distinguishing spurious associations from associations that are more likely to be real. Tools and applications such as OMIM and Neurocarta aim at a similar target: the scrutiny of the existing phenotype to gene associations. OMIM is not an ontology and ignores the potential relationship among phenotypes. On the other hand, Neurocarta is an ontology and allows for the navigation of concepts in a way resembling OGA. However, there are some marked differences. Neurocarta does not provide any functionality to produce the intersection between concepts, i.e. the genes that are common to two or more phenotypes. In this sense, Neurocarta seems to miss the main advantage of the ontology, the opportunity to exploit the relationships between phenotypes. Neurocarta is manually curated and OGA is generated completely automatically. Since OGA is generated automatically it can constantly monitor updates to the sources and automatically recreate previous defined analysis; this is another marked difference from Neurocarta which is manually curated. The most commonly associated phenotype in OGA is Alzheimer's disease (2,433 genes) and the same phenotype is the 10^{th} most associated in Neurocarta (259 genes). The concept hypertension is the fourth most associated in OGA (1,251) and it is the third more associated in Neurocarta (439 genes). Overall five of the top most

associated phenotypes in OGA are the in the top ten most associated concepts in Neurocarta. The higher count of genes for OGA could be explained in part because OGA takes association in the same sense as described for phenopedia (Yu, et al., 2010); the link between phenotype and gene is established because there is an entry for them in GAD. That indicates that the interaction has been studied, but not necessarily that there is a positive association. The gene with the most associations in OGA is ACE (1,923 phenotypes) and ACE is not among the top ten in Neurocarta. However, HLA-DRB1, MTHFR, TNF are in the top ten list for both. APOE, ACE, AGT, and MTHFR are all genes among the top most frequently associated for OGA, but they are also in the list of genes associated with obesity, diabetes, and hypertension and lipid metabolism. This observation lends weight to the possibility that these common genes may be present in this type of analysis simply by chance, and that it is simply that these phenotypes are repeatedly studied for the same genes. That may certainly be the case, but it perhaps also lends some support to genes like PON1 (29^{th} in the list of most frequent) or CEPT (73th in the list). In the case of the neurological diseases, only one gene SLC6A3 is among the top 20 most associated genes, HTR2A is 48^{th} and SLCA3 does not even make the top 100.

The phenotypes with the most gene associations in OGA were Alzheimer's disease with 2,433 genes, followed by Schizophrenia with 1,816 and in third place Colorectal Cancer with 1,581. This observation is not surprising as neurological disease and cancer are among the most heavily studied traits and the ones with the most support.

CHAPTER EIGHT: TUTORIAL

OGA is fully implemented in JAVA. It can be downloaded as a jar file in a standalone archive that contains everything necessary to run the application.

The OGA file structure has five main files.

Table 13 OGA file structure

File	Description
oga.jar	The java jar file that contains all the classes necessary to run the application
merge.db	The SQLite database that implements the database design (see methods)
Concepts.data	The names of the HOP concepts
Concepts.data.bis	The index to support the suffix array based pattern matching
Libraries	sqlite-jdbc-3.7.2.jar a dependency to connect to the SQLite database

The application can be run from the command line with command “java –jar oga.jar” or in the windows environment double clicking on the jar file should suffice to start the application.

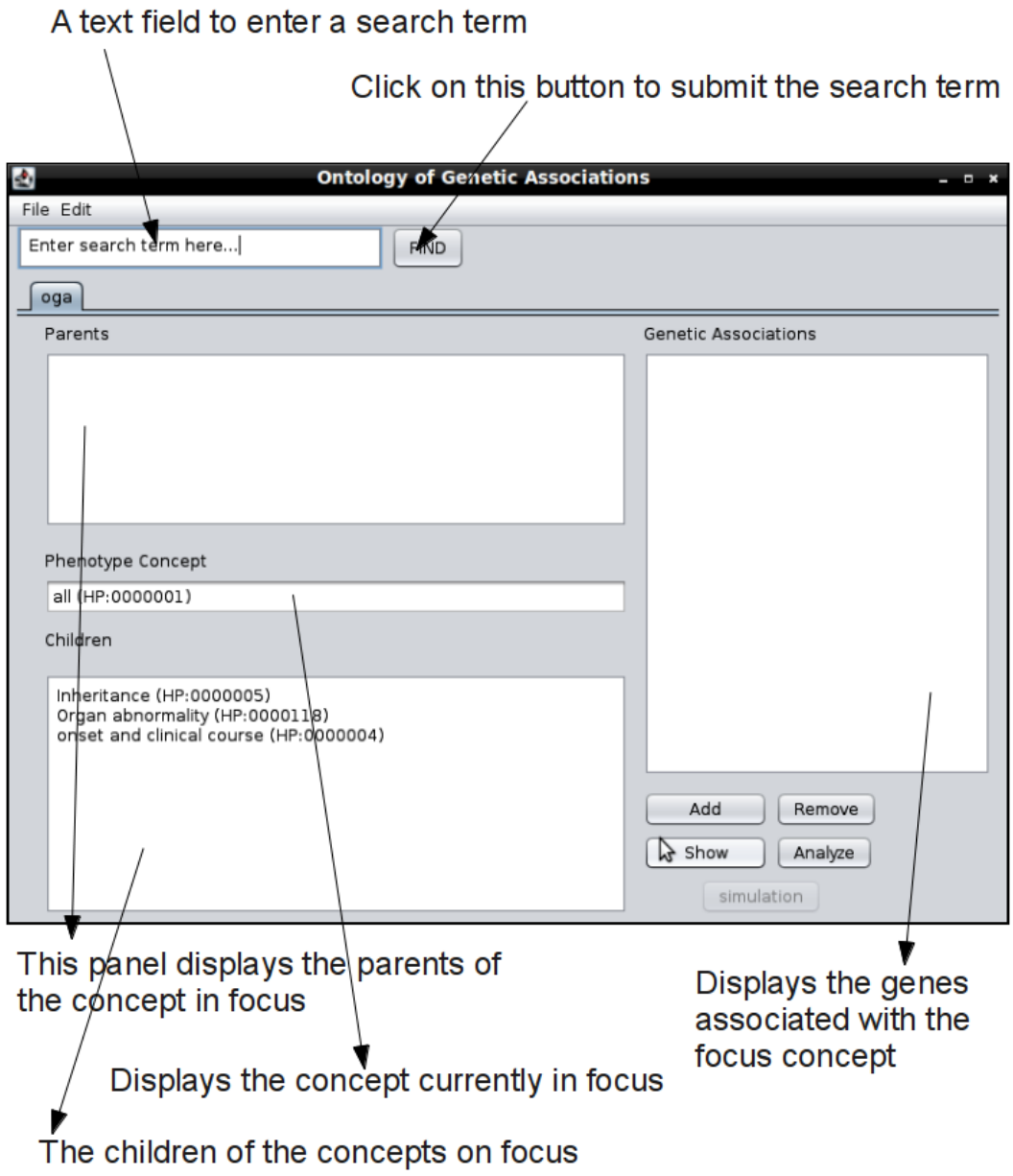

Figure 15 Description of navigation fields for OGA.

The application is divided in two main functional groups: the group related to navigation and the group related to analysis.

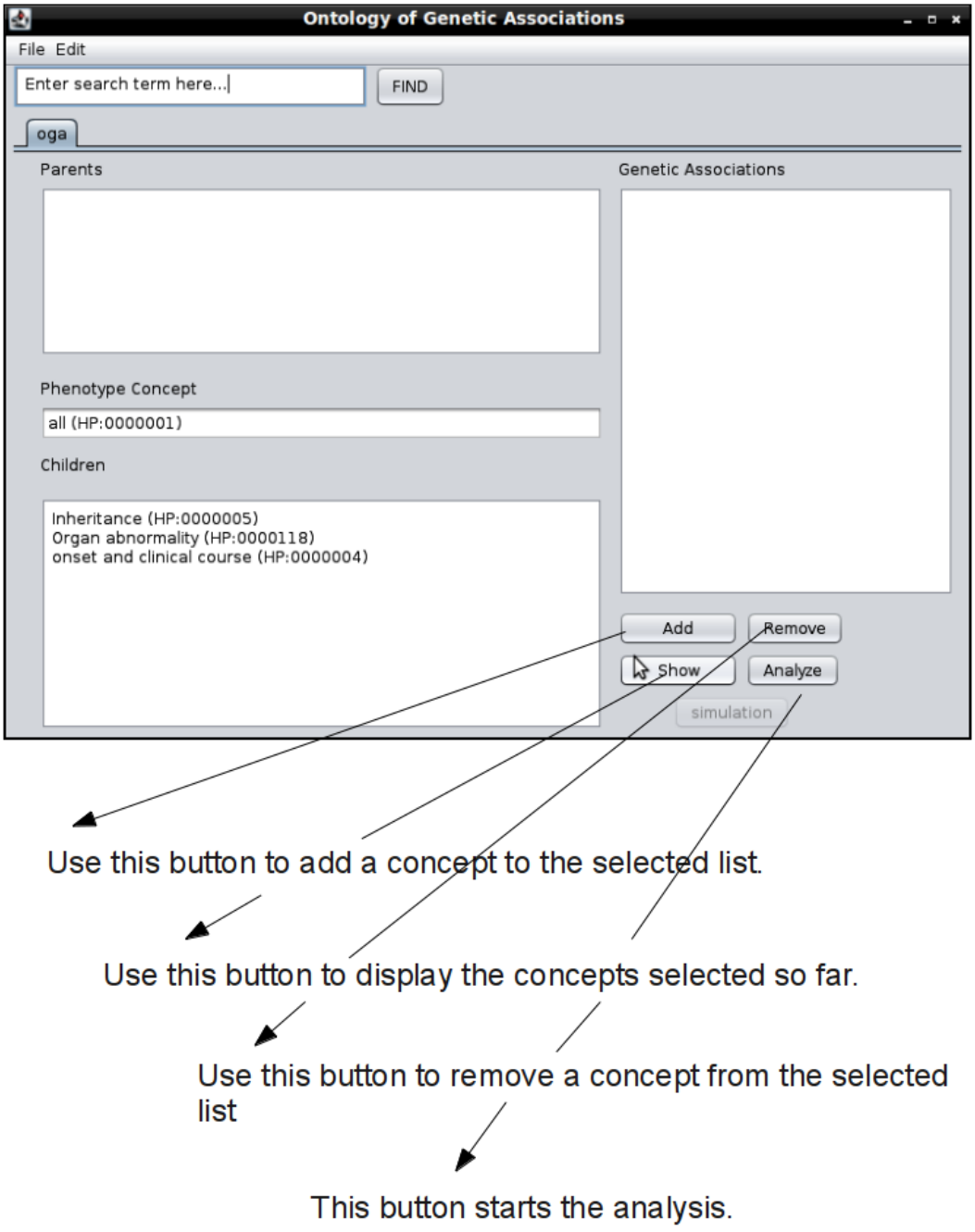

Figure 16 Description of analysis fields for OGA

The best way to start using OGA is to utilize the search field to enter a search term. The search term can be any part, prefix or suffix, of the term of interest. The application utilizes the suffix array based pattern matching to find all the HOP concepts that match the query. After the "FIND" button had been pressed a dialog window appears that contains the matching results.

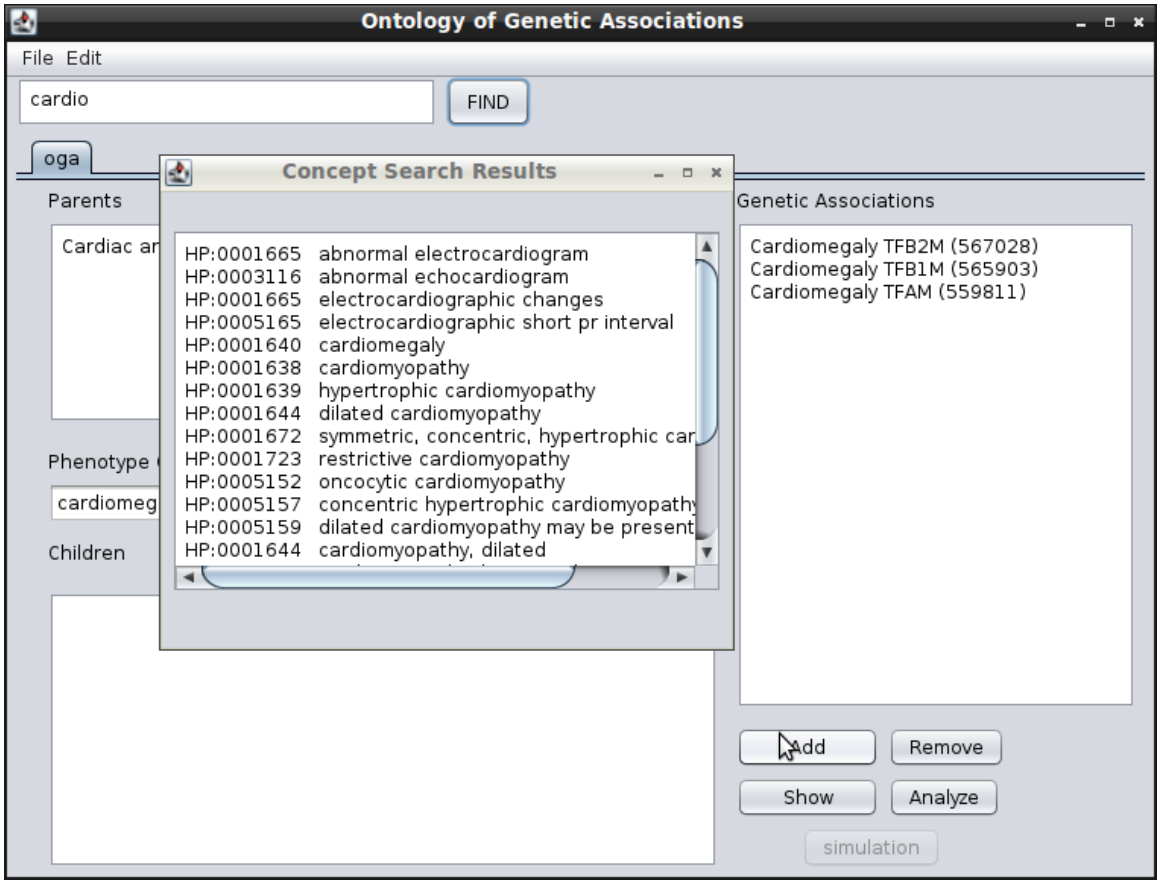

Figure 17 Results of the OGA pattern matching

In this example, OGA returns all the HOP concepts that contain the pattern "cardio". The user can utilize the dialog to click on any of the concepts of interest.

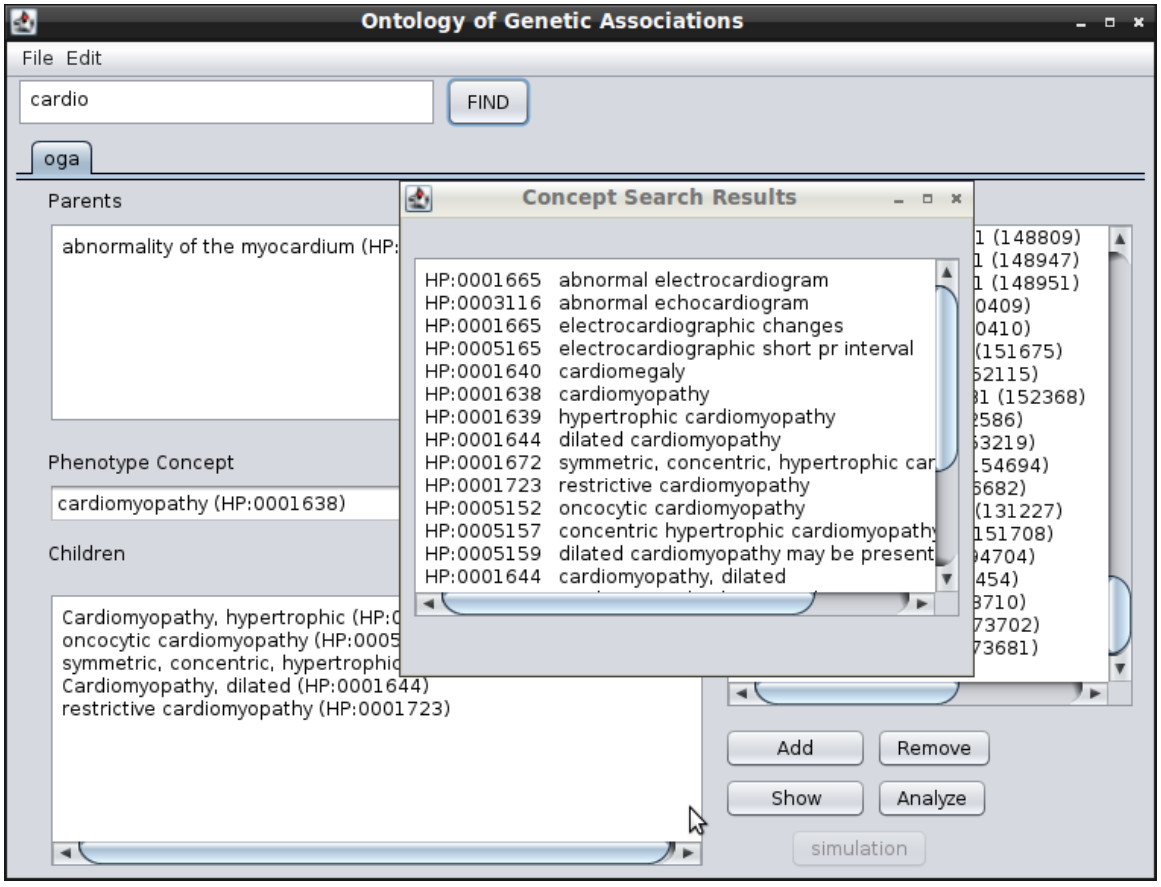

Figure 18 OGA navigation

For example, after the user has clicked the concept "cardiomyopathy" in the dialog, that concept becomes the focus phenotype as shown above.

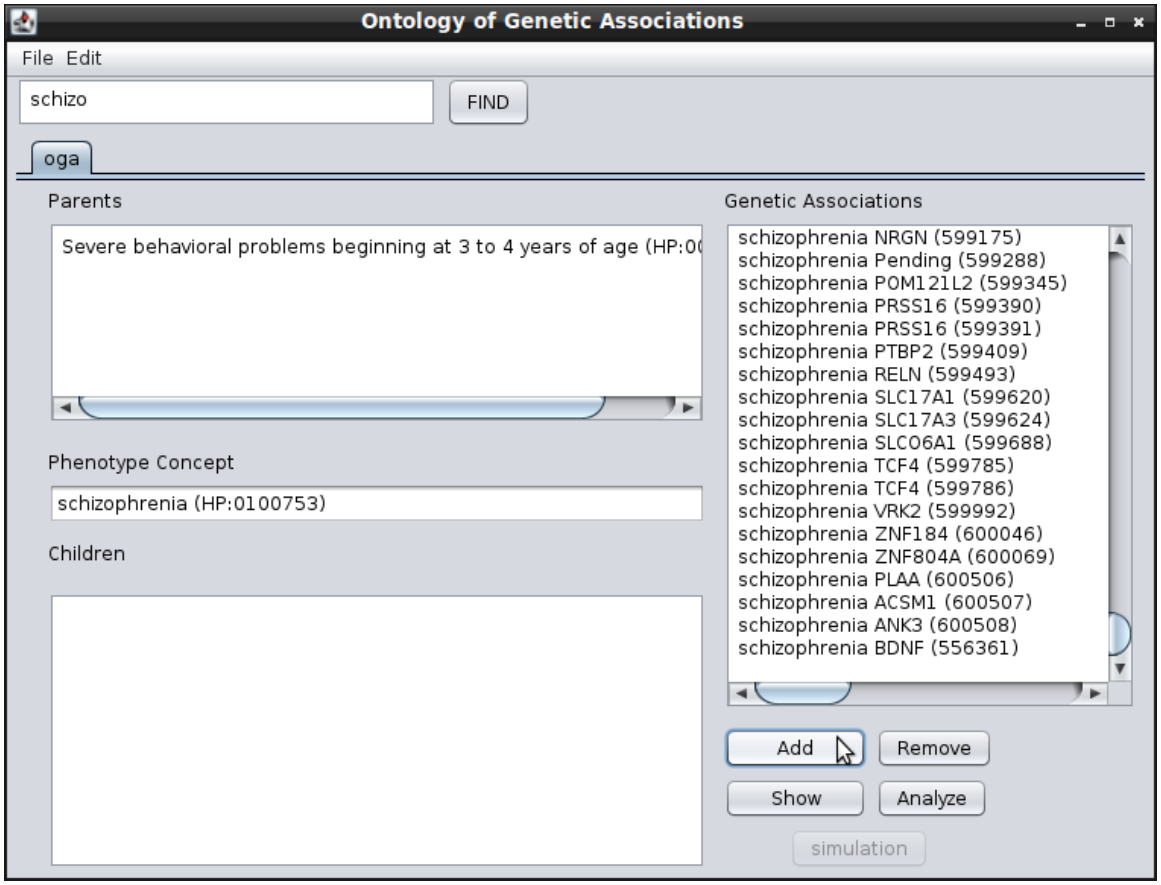

Figure 19 How to add a concept to the selected list

In order to add the concept currently in focus to the selected list, the user must click on the “add” button. In the example above, clicking on the “add” button will result in the concept “schizophrenia” being added to the selected list.

Given that the bulk of the analysis is based on finding the intersection of associated genes between one or more concepts, the selected list must have at least two concepts in order to be able to run the analysis.

After the user has selected all the concepts of interest, the selected concept list can be displayed by clicking on the “Show” button.

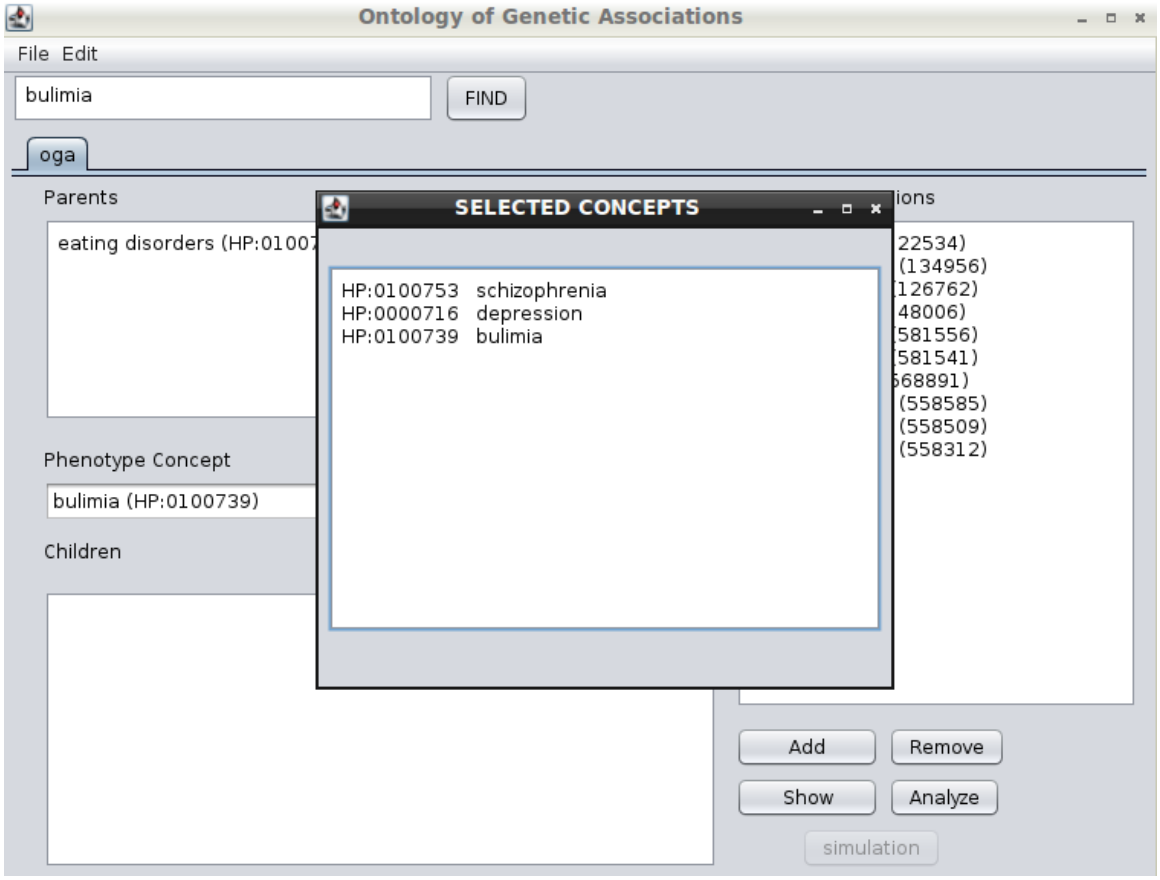

Figure 20 Selected concepts list displayed

Once the user has selected all the concepts of interest the “Analyze” button can be pressed to obtain the gene intersection.

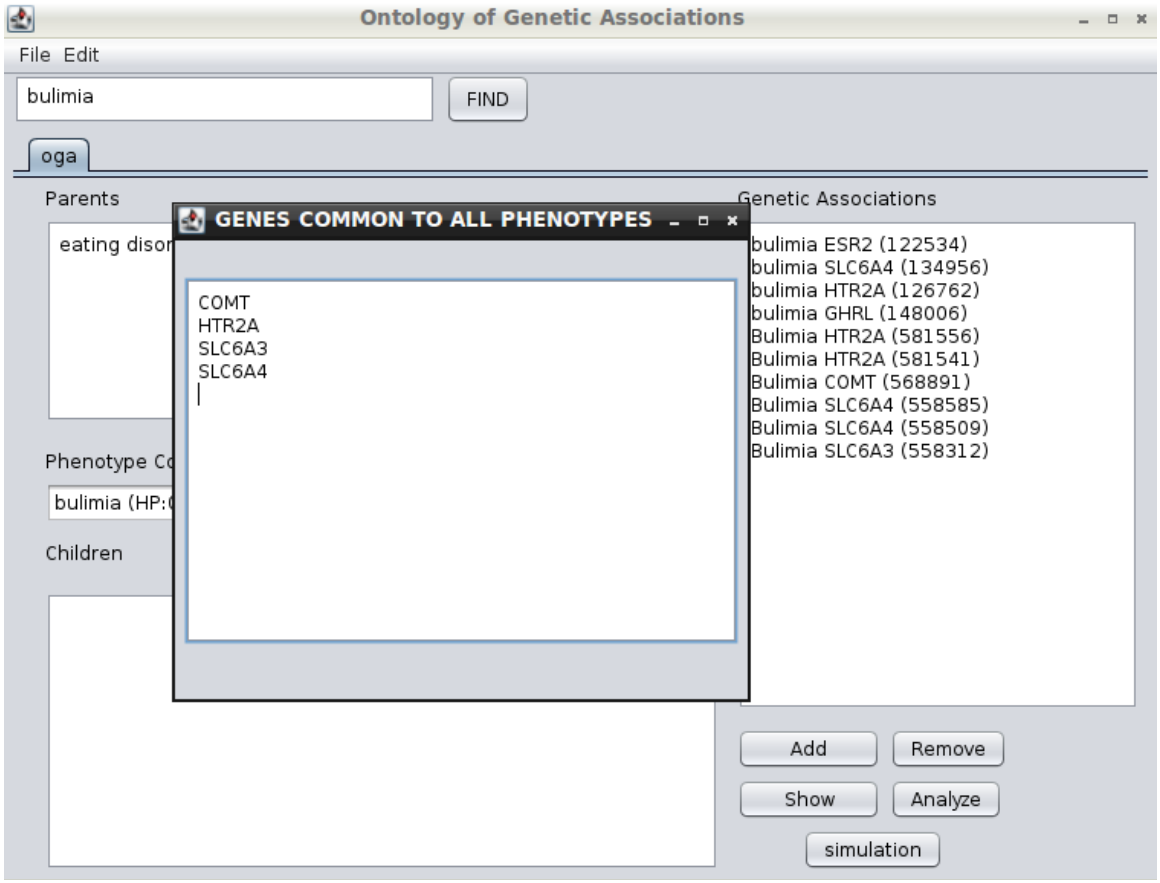

Figure 21 Results of OGA analysis

In the example, running the analysis on three selected concepts (Schizophrenia, bulimia and depression) resulted in a gene intersection of four genes (COMT, HTR2A, SLC6A3, SLC6A4). Notice that to this point the "simulation" button has been disabled. However, once at least one analysis has been ran, that button becomes enabled and an empirical *p* value can be calculated.

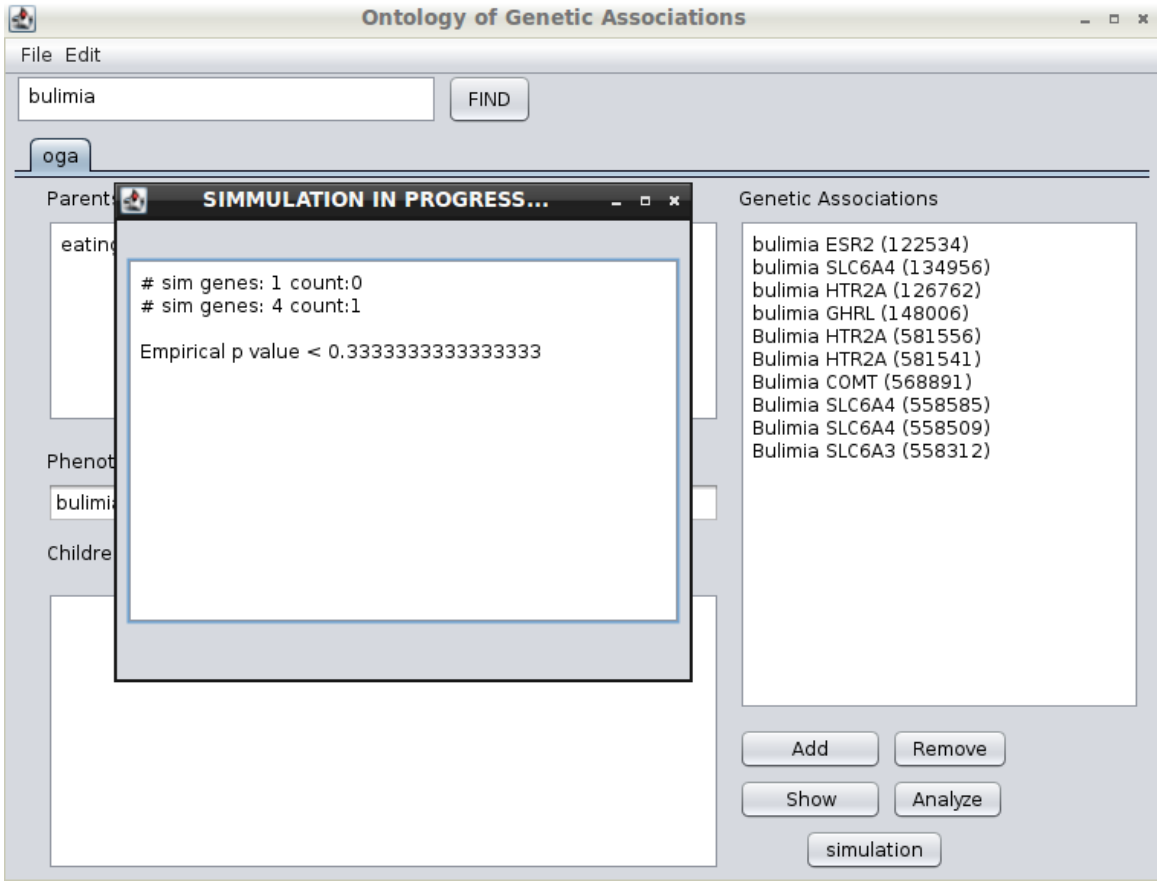

Figure 22 Empirical p value calculation

CHAPTER NINE: FUTURE DIRECTIONS

OGA currently uses HOP as the backbone; however, the extension of the backbone to use the DO and other ontologies may make the ontology more comprehensive and also may increase the potential for the discovery of more relationships and links. Other sources of genetic association could also be added, for example the addition of GWASdb (Li et al., 2012) may also increase the number of relationships within OGA, but more importantly most of the associations generated by GWAS studies are hypothesis free, so this may help solve the issue of the same genes being repeatedly studied for related phenotypes.

An important extension to OGA could be the ability to download and update the backbone and the association sources. The automation of OGA updating is quite feasible as the integration between the backbone and genetic associations is done automatically using the suffix array.

Another potential direction for OGA could be integration of data from dbGAP (Mailman, et al., 2007). There is already an application that could facilitate the integration with OGA (Wooten & Huggins, 2011). This type of functionality could allow users to perform tests of association within OGA for the gene intersections identified.

CODE ADDENDUMS

Code addendum 1: stop_word_remover.pl

```
#!/usr/bin/perl -w


MAIN:{
      open(FILE,"$ARGV[0]") || die "No input\n";
      my @words = <FILE>;
      close(FILE);
      my %stop_words;
      foreach my $word(@words)
         {
               $word=~s/\n//g;
               $stop_words{$word}++;
             }
      open(FILE,"$ARGV[1]") || die "No input\n";
      my @rows = <FILE>;
      close(FILE);
      foreach my $row(@rows)
         {
               $row=~s/\n//g;
               my @tals = split('\t',$row);
               my $id = $tals[0];
               my $name = $tals[1];
               print "$id";
               if(@tals>1){print "\t";}
               my @vals = split(' ',$name);
               my $new_name='';
               for(my $i=0;$i<@vals;$i++)
                  {
                          if(!defined($stop_words{$vals[$i]}) )
                             {
                             $new_name=$new_name.$vals[$i];
                             if($i<$#vals)
                                {
                                     $new_name=$new_name." ";
                                }
                             }
                     }
                  print "$new_name\n";
            }

         }
```

Code addendum 2

Binarysearch.h

```
#ifndef BINARYSEARCH_H
#define BINARYSEARCH_H

#include <set>
#include <map>
#include <queue>

using namespace std;

class binarysearch
   {
   private:
     std::string fixed_input;
     vector<int> ordering;

   public:
     binarysearch();
     void set_input(std::string ain);
     void set_ordering(vector<int> passed);
     vector<int> find_pattern(std::string pattern);
     double find_left_boundary(char to_test,double ls,double le,string input);
     double find_right_boundary(char to_test,double ls,double le,string input);
     std::string convert_string(double start, double end,double displ);
   };


#endif
```

Binarysearch.cc

```
#include <iostream>
#include <set>
#include <queue>
#include <map>
#include <cmath>
#include "binarysearch.h"
#include "bsort.h"


using namespace std;

/*
 * binary search to support a suffix array
 */

binarysearch::binarysearch()
   {
   }


double  binarysearch::find_right_boundary(char to_test,double ls,double le, string input)
   {

   double t1 =  floor( ls + ((le-ls)/2) ) ;
   double next =  t1+1;
   double answer  = -1;
   if (    (next==le)  &&  (input[t1]==to_test)
      )
      {
          return t1;
          }
   if (    (next==le)  &&  (input[t1-1]==to_test)
      )
      {
          return t1-1;
          }

   else if ( (next<le) &&(input[next]<to_test) && (input[t1]<to_test)
       )
      {
           ls=t1;
           answer = binarysearch::find_right_boundary(to_test,ls,le,input);
      }
 else if (   (input[t1]==to_test) && (input[next]==to_test)
       )
      {
           ls=t1;

           answer = binarysearch::find_right_boundary(to_test,ls,le,input);
      }
  else if (   (input[t1]<to_test) && (input[next]==to_test)
       )
      {
           ls=t1;

           answer = binarysearch::find_right_boundary(to_test,ls,le,input);
      }
   else if( (input[t1]==to_test) && (input[next]>to_test)
          )
      {
      answer = t1;
      }
   else if ( (next<le) && (input[t1]>to_test)  && (input[next]>to_test)
       )
      {
           le=t1;
           answer = binarysearch::find_right_boundary(to_test,ls,le,input);
```

```
      }


    return answer;
   }

double  binarysearch::find_left_boundary(char to_test,double ls,double le,string input)
   {

   double t1 =  floor( ls + ((le-ls)/2) );
   double prior =  t1-1;
   double answer  = -1;

   if( (le-ls)<1 )
      {
           return answer;
         }
   else if   (  (t1==ls) && ( input[t1]==to_test ) )
      {
       return t1;
         }
   else if   (  (t1==ls) && ( input[t1]!=to_test ) )
      {
       return answer;
         }
   else if ( (input[prior]<to_test) && (input[t1]<to_test)
       )
      {
           ls=t1;
           answer = binarysearch::find_left_boundary(to_test,ls,le,input);
      }
   else if(  (input[prior]<to_test) && (input[t1]==to_test)
          )
      {
      answer = t1;
      }
   else if ( (input[prior]>to_test) && (input[t1]>to_test)
       )
      {
           le=t1;
           answer = binarysearch::find_left_boundary(to_test,ls,le,input);
      }
    else if ( (input[prior]==to_test) && (input[t1]==to_test)
       )
      {
           le=t1;
           answer = binarysearch::find_left_boundary(to_test,ls,le,input);
      }
     else if ( (input[prior]==to_test) && (input[t1]>to_test)
       )
      {
           le=t1;
           answer = binarysearch::find_left_boundary(to_test,ls,le,input);
      }
    return answer;
   }

void binarysearch::set_input(string ain)
   {
    fixed_input = ain;
   }

void binarysearch::set_ordering(vector<int> passed)
   {
    ordering=passed;
   }
```

z

```
string binarysearch::convert_string(double start, double end,double displ)
    {
     string ordered;
    for(double i =start;i<end;i++)
      {
       ordered+=fixed_input[ordering[i]+displ];
      }
    return ordered;
    }


vector<int> binarysearch::find_pattern(string pattern)
   {
   double start =0;
   double end   = fixed_input.length();

   for(double i =0;i<pattern.length();i++)
      {

      string ordered = convert_string(start,end,i);

      double  left  =  find_left_boundary(pattern[i],0,ordered.length(),ordered);
      double  right  = find_right_boundary(pattern[i],0,ordered.length(),ordered);

          end   = start+right+1;
      start = start+left;
      if( (left<0) || (right<0)
        )
         {
             i=pattern.length();
             end=0;
             start=0;
             }
      }
    vector<int> matchings;

    for(double i =start;i<end;i++)
      {
          matchings.push_back(ordering[i]);
      }
   return matchings;
   }
```

Hop_gad_matcher.cc

```
#include <iostream>
#include <set>
#include <queue>
#include <map>
#include <fstream>
#include "binarysearch.h"
#include "Parser.h"


//Author: J. Enrique Herrera-Galeano
// 11/3/2012

using namespace std;

string remove_token(string txt, char token)
   {
   Parser po;
   vector<string> entries;
   po.ParseByToken(txt,token,entries);
   string out;
   for(unsigned int i=0;i<entries.size();i++)
      {
          out+=entries[i];
      }
    return out;
   }



void traverse_map(map<string,int> &hops,double cutoff,string gad_id )
  {
  map<string,int >::iterator it;
  for(it=hops.begin();it!=hops.end();it++)
    {
    int tmp = (*it).second;
    if((double)tmp>=cutoff)
       {
         //cout<<(*it).first<<"< -- >"<<tmp<<endl;
       cout<<(*it).first<<"\t"<<gad_id<<endl;
       }
    }
  }


void get_hop_ids(string &input, vector<int> &matchings,map<string,int> &hops)
  {
   for(unsigned int i=0;i<matchings.size();i++)
      {
           int start = matchings[i];
           //I know that this string is separated by \n
           int stx=start;
           unsigned int endx=start;
           while( (input[stx]!=':') && (stx>=0) )
              {
                  stx--;
              }
           stx=stx-2;
           if(stx<0){stx=0;}
           endx=stx;
           while( (input[endx]!='\t') && (endx< input.length() ))
              {
                   endx++;
              }
           endx=endx-stx;
           string matched = input.substr(stx,endx);
           if(hops.count(matched)==0)
              {
```

```
                hops[matched]=1;
                //cout<<matched<<endl;
            }
        else
            {
                hops[matched]=hops[matched]+1;
                //cout<<matched<<endl;
            }

    }

  }

void format_matches(string &input, vector<int> &matchings)
  {
   for(unsigned int i=0;i<matchings.size();i++)
      {
           int start = matchings[i];
           //I know that this string is separated by \n
           int stx=start;
           unsigned int endx=start;

           while( (input[endx]!='\n') && (endx< input.length() ))
              {
                   endx++;
              }
           while( (input[stx]!=':') && (stx>=0) )
              {
                  stx--;
              }
           stx=stx-2;
           if(stx<0){stx=0;}
           endx=endx-stx;

           string matched = input.substr(stx,endx);
           //cout<<matched<<" length:"<<matched.length()<<endl;
      }

  }

void getInput(fstream &infile,string &input)
    {
    if(infile.is_open())
     {
     string line;
     int index = 0;
     while(infile.good())
        {
        getline(infile,line);
        input+=line;
        input+='\n';
        index++;
        }
     }

  }


int main(int argc, char* argv[0])
  {
   //First argument has to be the orignal text in this case HOP   HOP is the text
   //the second argument is the index file
   //Third argument is the list of GAD entries GAD will be the patterns
   fstream infile;
   infile.open(argv[1],ios::in);
   string input;
```

```
getInput(infile,input);


//Takes the ordering file and
//places back on teh vector
fstream in_order;
in_order.open(argv[2],ios::in|ios::binary);
vector<int> ordering;
while( in_order.good())
   {
       int tmp;
       in_order.read( (char*) &tmp,sizeof(int));
       ordering.push_back(tmp);
   }

//The third part is to load the GAD
fstream gfile;
gfile.open(argv[3],ios::in);
string gad_str;
getInput(gfile,gad_str);

//set the binary search with the suffix array
binarysearch bs;
bs.set_input(input);
bs.set_ordering(ordering);
vector<int> matchings;

Parser po;
vector<string> gad_entries;
po.ParseByToken(gad_str,'\n',gad_entries);


//matchings = bs.find_pattern("infarct;");
//format_matches(input,matchings);

for(unsigned int i =0;i<gad_entries.size();i++)
   {
        //cout<<gad_entries[i]<<endl;
        vector<string> entry_parts;
        po.ParseByToken(gad_entries[i],'\t',entry_parts);
        if(entry_parts.size()>1)
           {
                string gad_id   = entry_parts[0];
            string gad_name = entry_parts[1];
            vector<string> name_parts;
            po.ParseByToken(gad_name,' ',name_parts);
           // cout<<gad_id<<"|";
           map<string,int> hops;
           for(unsigned int j =0;j<1;j++)
              {
                  string tmp_part = remove_token(name_parts[j],',');
                  string name_part = remove_token(tmp_part,';');


                  if(name_part.length()>1)
                     {
                             //cout<<name_part<<" -- ";
                     matchings = bs.find_pattern(gad_name);
                     get_hop_ids(input,matchings,hops);
                     //cout<<"size:"<<matchings.size()<<endl;
                     }

          //format_matches(input,matchings);
                }
              double proportion_of_overlapping_words = 0.8;
              double cut_off = name_parts.size() * proportion_of_overlapping_words;
              traverse_map(hops,cut_off,gad_id);




              //cout<<endl;
       }
   }


}
```

Code addendum 3

Analyzer.java

```
/*
 * To change this template, choose Tools | Templates
 * and open the template in the editor.
 */
package oga;

import java.util.HashMap;
import java.util.Iterator;
import java.util.Vector;

/**
 *
 * @author jesuse
 */
public class Analyzer {
   Vector<String> selected;
   DataManager dm;
   Vector<Integer> genesSeen;
   Vector<String>  signGenes;

    public Vector<String> getSignGenes() {
        return signGenes;
    }

    public void setSignGenes(Vector<String> signGenes) {
        this.signGenes = signGenes;
    }


    public Vector<Integer> getGenesSeen() {
        return genesSeen;
    }

    public void setGenesSeen(Vector<Integer> genesSeen) {
        this.genesSeen = genesSeen;
    }


   public void runSimulation()
      {
      int counter = 0;
      int sim_genes = 0;
            Intermediator im=new Intermediator() {
               public void processChoice(int id) {

               }
            };
      SearchDialog  sd = new SearchDialog(new javax.swing.JFrame(),false,im);
      sd.setTitle("SIMMULATION IN PROGRESS...");
      sd.setVisible(true);
      String content="";
      sd.setText(content);
      while (  (counter<1000)
             && (sim_genes<this.signGenes.size()))
         {
         sim_genes = getEmpirical();
         content="# sim genes: "+sim_genes+" count:"+counter+"\n";
         System.out.println("# sim genes: "+sim_genes+" count:"+counter);
         sd.append(content);
         counter++;
         }
      double pvalue = 1.0/(counter+1);
      String message = "Empirical p value < "+pvalue;
      sd.append("\n");
      sd.append(message);
      }
```

```
public int getEmpirical()
   {
   //First how many sets do I need to retrieve
   //as many as concepts were selected
   System.out.println("sel: "+selected.size());
    HashMap hm = new HashMap();
   for (int i=0;i<selected.size();i++)
      {
      Vector<String> genes_rand = dm.getGenesRand(getGenesSeen().get(i));
      System.out.println("size: "+getGenesSeen().get(i));
      //MAKE GENE LIST UNIQUE

      Vector<String> genes = new Vector<String>();
      for(String gene:genes_rand)
         {
         if(!genes.contains(gene))
            {
            genes.add(gene);
            }

         }

      for(String gene:genes)
         {
        int test=1;
        if ( (!gene.equals("null"))  &&  (hm.get(gene)==null) )
           {
           hm.put(gene,new Integer(test));
           }
        else if( (hm.get(gene)!=null) )
          {
          test=(Integer)hm.get(gene)+1;
          hm.put(gene,new Integer(test));
          }
         }
      }

  Iterator it = hm.keySet().iterator();
  int number_of_sig_genes_rand = 0;
  while(it.hasNext())
     {
     String key    = (String) it.next();
     int val    = (Integer)hm.get(key);
     if(val>=selected.size())
        {
        //System.out.println("key: "+key+" val:"+val);
        number_of_sig_genes_rand++;
        }
     }

   return number_of_sig_genes_rand;
   }

//This sets the data for
public void getData()
  {
  HashMap hm = new HashMap();
  Vector<Integer> sizes = new Vector<Integer>();
  for(String s : selected)
     {//TO-DO test getGadsLinage!
     String[] vals = s.split(":");
     String sel = vals[1];
     Vector<Integer> gads = dm.getGadsLineage(sel);
     Vector<String> genes = new Vector<String>();

     for(int i:gads)
```

```
            {
            String gene = dm.getGadGene(Integer.toString(i));
            if(!genes.contains(gene))
               {
               genes.add(gene);
               }
            }

         sizes.add(genes.size());
         for(String gene:genes)
            {
            int test=1;
            if ( (!gene.equals("null"))  &&  (hm.get(gene)==null) )
               {
               hm.put(gene,new Integer(test));
               }
            else if( (hm.get(gene)!=null) )
              {
              test=(Integer)hm.get(gene)+1;
              hm.put(gene,new Integer(test));
              }
            }

         }
       //set the vector that represents the
       //sizes of the genes retrieved
       this.setGenesSeen(sizes);
       Vector<String> tmpSignGenes = new Vector<String>();
       Iterator it = hm.keySet().iterator();
       while(it.hasNext())
          {
          String key    = (String) it.next();
          int val    = (Integer)hm.get(key);

          if(val>=selected.size())
             {
             System.out.println(key+" : "+val);
             tmpSignGenes.add(key);
             }
          }
       this.setSignGenes(tmpSignGenes);
       }

      public DataManager getDm() {
          return dm;
      }

      public void setDm(DataManager dm) {
          this.dm = dm;
      }


     public boolean validate()
       {
       boolean test=false;
       if(selected.size()>1)
         {
         test=true;
         }
       return test;
       }

      public Vector<String> getSelected() {
          return selected;
      }

      public void setSelected(Vector<String> selected) {




          this.selected = selected;
      }

}
```

BinarySearch.java

```
/*
 * To change this template, choose Tools | Templates
 * and open the template in the editor.
 */
package oga;

/**
 *
 * @author jesuse
 */
import java.io.IOException;
import java.io.RandomAccessFile;
import java.nio.ByteBuffer;
import java.util.Vector;
import java.util.logging.Level;
import java.util.logging.Logger;



public class BinarySearch {



RandomAccessFile fixed_input_file;
int fixed_input_length;

RandomAccessFile dis_ord;
int byte_count;

/*
 * binary search to support a suffix array
 */

BinarySearch()
   {
   }

  public void setInput(RandomAccessFile input,int fileLength)
    {
    this.fixed_input_file=input;
    this.fixed_input_length=fileLength;
    }



  public void setOrdering(RandomAccessFile dis_ord,int byte_count)
    {
    this.dis_ord=dis_ord;
    this.byte_count=byte_count;
    }

  private int getOrdering(int index)
     {
     byte[] myint = new byte[byte_count];
        try {
            dis_ord.seek((long)index * (byte_count) );
            for(int i=(byte_count-1);i>-1;i--)
              {
              myint[i]=dis_ord.readByte();
              }
                      ByteBuffer bb = ByteBuffer.wrap(myint);
            } catch (IOException ex) {
            Logger.getLogger(OgaMainFrame.class.getName()).log(Level.SEVERE, null, ex);
            }
```

```
private char getInputChar( int ind, int displ, int stx) throws IOException
   {
     int loc = stx +  ind;
     //int index= ordering.get( loc)+displ;
     int index = this.getOrdering(loc)+displ;
     if(index>=fixed_input_length)
        {
          index=index%this.fixed_input_length;
        }

    fixed_input_file.seek((long)index);
    char rslt  =  (char) fixed_input_file.readByte() ;

    return rslt;
   }




Vector<Integer> find_pattern(String pattern) throws IOException
   {
   int start =0;
   int end   = fixed_input_length;

   for(int i =0;i<pattern.length();i++)
      {

      int seg_len = end-start;

      int  left  =  find_left_boundary(pattern.charAt(i),0,seg_len,i,start);
      int  right  = find_right_boundary(pattern.charAt(i),0,seg_len,i,start);

      end   = start+right+1;
      start = start+left;
      if( (left==-1) || (right==-1)
        )
         {
             i=pattern.length();
             end=0;
             start=0;
             }
      }


    Vector<Integer> matchings= new Vector<Integer>();

    for(int i =start;i<end;i++)
      {
          matchings.add(getOrdering(i));

      }
   return matchings;
   }


}
```

```
    ByteBuffer bb = ByteBuffer.wrap(myint);
    return bb.getInt();
    }


public int find_right_boundary(char to_test, int ls, int le, int disp, int stx) throws IOException
   {

   int t1 =  (int) Math.floor( ls + ((le-ls)/2) ) ;
   int next =  t1+1;
   int answer  = -1;
   char input_t1 ;
   char input_t1_minus ;
   char input_next ;

   char calc_input_t1 = getInputChar(t1, disp, stx);
   char calc_input_t1_minus =  getInputChar(t1-1, disp, stx);
   char calc_input_next =  getInputChar(next, disp, stx);

   input_t1 = calc_input_t1;
   input_t1_minus = calc_input_t1_minus;
   input_next = calc_input_next;


   if (    (next==le)  &&  (input_t1==to_test)
      )
      {
          return t1;
          }
   if (    (next==le)  &&  (input_t1_minus==to_test)
      )
      {
          return t1-1;
          }

   else if ( (next<le) &&(input_next<to_test) && (input_t1<to_test)
       )
      {
           ls=t1;
           answer = find_right_boundary(to_test,ls,le,disp,stx);
      }
 else if (   (input_t1==to_test) && (input_next==to_test)
       )
      {
           ls=t1;
           answer = find_right_boundary(to_test,ls,le,disp,stx);
      }
  else if (   (input_t1<to_test) && (input_next==to_test)
       )
      {
           ls=t1;

           answer = find_right_boundary(to_test,ls,le,disp,stx);
      }
   else if( (input_t1==to_test) && (input_next>to_test)
          )
      {
      answer = t1;
      }
   else if ( (next<le) && (input_t1>to_test)  && (input_next>to_test)
       )
      {
           le=t1;
           answer = find_right_boundary(to_test,ls,le,disp,stx);
      }
```

```
    return answer;
   }

int  find_left_boundary(char to_test, int ls,  int le, int disp,  int stx) throws IOException
   {

    int t1 =  (int)Math.floor( ls + ((le-ls)/2) );
    int prior =  t1-1;
    int answer  = -1;



   char input_t1 = this.getInputChar(t1, disp, stx);



   char input_prior = getInputChar(prior, disp, stx);

   if( (le-ls)<1 )
      {
           return answer;
         }
   else if   (  (t1==ls) && ( input_t1==to_test ) )
      {
       return t1;
         }
   else if   (  (t1==ls) && ( input_t1!=to_test ) )
      {
       return answer;
         }
   else if ( (input_prior<to_test) && (input_t1<to_test)
       )
      {
           ls=t1;
           answer = find_left_boundary(to_test,ls,le,disp,stx);
      }
   else if(  (input_prior<to_test) && (input_t1==to_test)
          )
      {
      answer = t1;
      }
   else if ( (input_prior>to_test) && (input_t1>to_test)
       )
      {
           le=t1;
           answer = find_left_boundary(to_test,ls,le,disp,stx);
      }
    else if ( (input_prior==to_test) && (input_t1==to_test)
       )
      {
           le=t1;
           answer = find_left_boundary(to_test,ls,le,disp,stx);
      }
     else if ( (input_prior==to_test) && (input_t1>to_test)
       )
      {
           le=t1;
           answer = find_left_boundary(to_test,ls,le,disp,stx);
      }
    return answer;
   }
```

DataManager.java

```
/*
 * To change this template, choose Tools | Templates
 * and open the template in the editor.
 */
package oga;

/**
 *
 * @author jesuse
 */
import java.sql.Connection;
import java.sql.DriverManager;
import java.sql.ResultSet;
import java.sql.SQLException;
import java.sql.Statement;
import java.util.Vector;
import java.util.logging.Level;
import java.util.logging.Logger;


public class DataManager
{
 Connection connection = null;
 Statement statement   = null;
 public DataManager() throws ClassNotFoundException
   {
   // load the sqlite-JDBC driver using the current class loader
   Class.forName("org.sqlite.JDBC");
    try {
        connection = DriverManager.getConnection("jdbc:sqlite:merge.db");
        statement = connection.createStatement();
         } catch (SQLException ex) {
            Logger.getLogger(DataManager.class.getName()).log(Level.SEVERE, null, ex);
        }
   }

 public String formatId(String id)
    {
    while(id.length()<7)
      {
      id="0"+id;
      }
    return "HP:"+id;
    }

 public Vector<Integer> getGadsLineage(String id)
   {
   Vector<Integer> gads = new Vector<Integer>();
   Vector<Integer> children = getChildren(id);
   gads.addAll(getGads(id));
   if(children.isEmpty())
     {
     gads.addAll(getGads(id));
     }
   else
      {
      for(int i:children)
         {
         gads.addAll(getGadsLineage(Integer.toString(i))
                 );
         }
      }
   return gads;
   }

 public Vector<Integer> getGads(String id)
    {
```

```
    Vector<Integer> gads = new Vector<Integer>();
     try {

        String query = "SELECT Gad FROM HOP_GAD where hop=\'"+id+"\'";
        ResultSet rs = statement.executeQuery(query);
           while(rs.next())
              {
              gads.add( rs.getInt("Gad") );
              }

        } catch (SQLException ex) {
           Logger.getLogger(DataManager.class.getName()).log(Level.SEVERE, null, ex);
        }

   return gads;
   }

 public Vector<Integer> getParents(String id)
   {
   Vector<Integer> parents = new Vector<Integer>();
     try {

        String query = "SELECT Parent FROM PARENT_CHILD where Child=\'"+id+"\'";
        ResultSet rs = statement.executeQuery(query);
           while(rs.next())
              {
              parents.add( rs.getInt("Parent") );
              }

        } catch (SQLException ex) {
           Logger.getLogger(DataManager.class.getName()).log(Level.SEVERE, null, ex);
        }

   return parents;
   }

public Vector<Integer> getChildren(String id)
   {
   Vector<Integer> children = new Vector<Integer>();
     try {

        String query = "SELECT Child FROM PARENT_CHILD where Parent=\'"+id+"\'";
        ResultSet rs = statement.executeQuery(query);
           while(rs.next())
              {
              children.add( rs.getInt("Child") );
              }

        } catch (SQLException ex) {
           Logger.getLogger(DataManager.class.getName()).log(Level.SEVERE, null, ex);
        }

   return children;
   }


public Vector<String> getGenesRand(int size)
   {
   Vector<String> genes = new Vector<String>();
    try {

        String query = "SELECT Gene from GAP ORDER BY RANDOM() limit "+size;
        ResultSet rs = statement.executeQuery(query);
           while(rs.next())
              {
              genes.add( rs.getString("Gene") );
              }
```

```
        } catch (SQLException ex) {
            Logger.getLogger(DataManager.class.getName()).log(Level.SEVERE, null, ex);
        }

    return genes;
    }

  public String getGadGene(String id)
     {
     String gene="null";
     try {

         String query = "select gene from GAP where id=\'"+id+"\'";
         ResultSet rs = statement.executeQuery(query);
            while(rs.next())
               {
               gene= rs.getString("gene");
               if(gene.length()>0)
                  {
                  gene=rs.getString("gene");
                  }
               }
         } catch (SQLException ex) {
            Logger.getLogger(DataManager.class.getName()).log(Level.SEVERE, null, ex);
         }
    return gene;
     }

  public String getGadName(String id)
    {
    String name="";
     try {

         String query = "select phenotype,gene from GAP where id=\'"+id+"\'";
         ResultSet rs = statement.executeQuery(query);
            while(rs.next())
               {
               name= rs.getString("phenotype");
               String gene= rs.getString("gene");
               if(gene.length()>0)
                  {
                  name=name+" "+rs.getString("gene");
                  }

               }

         } catch (SQLException ex) {
            Logger.getLogger(DataManager.class.getName()).log(Level.SEVERE, null, ex);
         }
    return name;
    }

 public String getConceptName(String id)
    {
    String name="";
     try {

         String query = "select name from HOP where id=\'"+formatId(id)+"\'";
         ResultSet rs = statement.executeQuery(query);
            while(rs.next())
               {
               name= rs.getString("name");
               }

         } catch (SQLException ex) {
            Logger.getLogger(DataManager.class.getName()).log(Level.SEVERE, null, ex);
```

```
        }
     return name;
     }

 public void runQuery()
   {
   try {
            ResultSet rs = statement.executeQuery("select * from GAP");
            while(rs.next())
               {
                 // read the result set
                 System.out.println("name = " + rs.getString("phenotype"));
                 System.out.println("id = " + rs.getInt("id"));
               }
         } catch (SQLException ex) {
            Logger.getLogger(DataManager.class.getName()).log(Level.SEVERE, null, ex);
         }

   }


  public static void main(String[] args) throws ClassNotFoundException
     {
     DataManager dm = new DataManager();
     dm.runQuery();
     }
}
```

Intermediator.java

```
/*
 * To change this template, choose Tools | Templates
 * and open the template in the editor.
 */
package oga;

/**
 *
 * @author jesuse
 */
public interface Intermediator {
    public void processChoice(int test);
}
```

OgaMainFrame.java

```
/*
 * To change this template, choose Tools | Templates
 * and open the template in the editor.
 */
package oga;

import java.awt.GridLayout;
import java.awt.Toolkit;
import java.io.DataInputStream;
import java.io.File;
import java.io.FileInputStream;
import java.io.FileNotFoundException;
import java.io.IOException;
import java.io.RandomAccessFile;
import java.nio.ByteBuffer;
import java.util.Vector;
import java.util.logging.Level;
import java.util.logging.Logger;

/**
 *
 * @author jesuse
 */
public class OgaMainFrame extends javax.swing.JFrame {


 FileInputStream ordering_file;
 DataInputStream dis_ord;
 int bit_size ;
 int byte_count;
 final OgaBrowser oga = new OgaBrowser();;
 SearchDialog sd;
    /**
     * Creates new form AlignerJFrame
     */
    public OgaMainFrame() {
        //Toolkit toolkit = Toolkit.getDefaultToolkit();
        //this.setPreferredSize(toolkit.getScreenSize());
        initComponents();



        //open the files and load them
        //open text
                // Stream to read file
        //READ TEXT FILE
        try
          {
          // Open an input stream
          ordering_file = new FileInputStream ("Concepts.data.bis");
          bit_size = Integer.parseInt( System.getProperty("sun.arch.data.model"));
          byte_count = bit_size/8;

           }           // Catches any error conditions
        catch (IOException e)
          {
                       System.err.println ("Unable to read from file");
                       System.exit(-1);
          }

    }

    /**
     * This method is called from within the constructor to initialize the form.
     * WARNING: Do NOT modify this code. The content of this method is always
     * regenerated by the Form Editor.
```

```
     */
    @SuppressWarnings("unchecked")
    // <editor-fold defaultstate="collapsed" desc="Generated Code">//GEN-BEGIN:initComponents
    private void initComponents() {

        jScrollPane1 = new javax.swing.JScrollPane();
        jTextPane1 = new javax.swing.JTextPane();
        jButton1 = new javax.swing.JButton();
        jTabbedPane1 = new javax.swing.JTabbedPane();
        jMenuBar1 = new javax.swing.JMenuBar();
        jMenu1 = new javax.swing.JMenu();
        jMenu2 = new javax.swing.JMenu();
        jMenu3 = new javax.swing.JMenu();

        setDefaultCloseOperation(javax.swing.WindowConstants.EXIT_ON_CLOSE);
        setTitle("Ontology of Genetic Associations");

        jTextPane1.setText("Enter search term here...");
        jScrollPane1.setViewportView(jTextPane1);

        jButton1.setText("FIND");
        jButton1.addActionListener(new java.awt.event.ActionListener() {
            public void actionPerformed(java.awt.event.ActionEvent evt) {
                jButton1ActionPerformed(evt);
            }
        });

        oga.setVisible(true);
        jTabbedPane1.add("oga",oga);

        jMenu1.setText("File");
        jMenuBar1.add(jMenu1);

        jMenu2.setText("Edit");

        jMenu3.setText("About");
        jMenu3.addActionListener(new java.awt.event.ActionListener() {
            public void actionPerformed(java.awt.event.ActionEvent evt) {
                jMenu3ActionPerformed(evt);
            }
        });
        jMenu2.add(jMenu3);

        jMenuBar1.add(jMenu2);

        setJMenuBar(jMenuBar1);

        javax.swing.GroupLayout layout = new javax.swing.GroupLayout(getContentPane());
        getContentPane().setLayout(layout);
        layout.setHorizontalGroup(
            layout.createParallelGroup(javax.swing.GroupLayout.Alignment.LEADING)
            .addGroup(layout.createSequentialGroup()
                .addContainerGap()
                .addGroup(layout.createParallelGroup(javax.swing.GroupLayout.Alignment.LEADING)
                    .addGroup(layout.createSequentialGroup()
                        .addComponent(jTabbedPane1)
                        .addContainerGap())
                    .addGroup(layout.createSequentialGroup()
                        .addComponent(jScrollPane1, javax.swing.GroupLayout.PREFERRED_SIZE, 277,
javax.swing.GroupLayout.PREFERRED_SIZE)
                        .addPreferredGap(javax.swing.LayoutStyle.ComponentPlacement.RELATED)
                        .addComponent(jButton1)
                        .addGap(0, 410, Short.MAX_VALUE))))
        );
        layout.setVerticalGroup(
            layout.createParallelGroup(javax.swing.GroupLayout.Alignment.LEADING)
            .addGroup(javax.swing.GroupLayout.Alignment.TRAILING, layout.createSequentialGroup()
```

```
                .addGroup(layout.createParallelGroup(javax.swing.GroupLayout.Alignment.LEADING, false)
                    .addComponent(jScrollPane1, javax.swing.GroupLayout.DEFAULT_SIZE, 33, Short.MAX_VALUE)
                    .addComponent(jButton1, javax.swing.GroupLayout.DEFAULT_SIZE,
javax.swing.GroupLayout.DEFAULT_SIZE, Short.MAX_VALUE))
                .addPreferredGap(javax.swing.LayoutStyle.ComponentPlacement.RELATED)
                .addComponent(jTabbedPane1, javax.swing.GroupLayout.DEFAULT_SIZE, 452, Short.MAX_VALUE)
                .addContainerGap())
        );

        pack();
    }// </editor-fold>//GEN-END:initComponents

private char getInputChar( int index, RandomAccessFile fixed_input_file) throws IOException
   {

    fixed_input_file.seek((long)index);
    char rslt  =  (char) fixed_input_file.readByte() ;
    return rslt;
   }

private String getInputSubstring(int stx,int end, RandomAccessFile raf) throws IOException
  {
  String matched="";
  raf.seek((long)stx);
  for(int i=stx;i<end;i++)
     {
     matched+=(char)raf.readByte();
     }
  return matched;
  }

   public void updateOga(int index)
     {
      oga.updateBrowser(Integer.toString(index));
      this.getContentPane().repaint();
     }

    private void jButton1ActionPerformed(java.awt.event.ActionEvent evt) {//GEN-
FIRST:event_jButton1ActionPerformed
        // TODO add your handling code here:

        String myInput = jTextPane1.getText();

        BinarySearch bs = new BinarySearch();


          RandomAccessFile dis;
          try {

            dis= new RandomAccessFile("Concepts.data.bis","r");
            bs.setOrdering(dis, byte_count);
            } catch (FileNotFoundException ex) {
            Logger.getLogger(OgaMainFrame.class.getName()).log(Level.SEVERE, null, ex);
            }

          RandomAccessFile input_file=null;
        try {
            input_file = new RandomAccessFile("Concepts.data","r");
            int file_length = (int) input_file.length();
            bs.setInput(input_file, file_length);

            } catch (IOException ex) {
            Logger.getLogger(OgaMainFrame.class.getName()).log(Level.SEVERE, null, ex);
           }

       String[] matchingsText=null ;
       String output="";
```

```
        Vector<Integer> matchings;
        try {
            matchings = bs.find_pattern(myInput);

            matchingsText = new String[matchings.size()];
            int index=0;

        for(int i=0;i<matchings.size();i++)
           {
           int start = matchings.get(i);
           //I know that this string is separated by \n

           int stx=start;
           int endx=start;

           while( (getInputChar(endx,input_file)!='\n') && (endx< input_file.length() ))
              {
                   endx++;
              }

           while( (getInputChar(stx,input_file)!=':') && (stx>=0) )
              {
                  stx--;
              }
           if( stx<0  ){stx=0;}
           String matched = this.getInputSubstring(stx,endx,input_file);
           output+="HP"+matched+"\n";
           matchingsText[index]="HP"+matched+"\n";
           index++;
           }
         } catch (IOException ex) {
            Logger.getLogger(OgaMainFrame.class.getName()).log(Level.SEVERE, null, ex);
        }
       Intermediator im=new Intermediator() {
               public void processChoice(int id) {
               updateOga(id);
               }
            };
        sd = new SearchDialog(this,false,im);
        sd.setText(output);
        sd.setMachings(matchingsText);
        sd.setVisible(true);
    }//GEN-LAST:event_jButton1ActionPerformed

    private void jMenu3ActionPerformed(java.awt.event.ActionEvent evt) {//GEN-
FIRST:event_jMenu3ActionPerformed
        // TODO add your handling code here:
    }//GEN-LAST:event_jMenu3ActionPerformed

    /**
     * @param args the command line arguments
     */
    public static void main(String args[]) {
        /* Set the Nimbus look and feel */
        //<editor-fold defaultstate="collapsed" desc=" Look and feel setting code (optional) ">
        /* If Nimbus (introduced in Java SE 6) is not available, stay with the default look and feel.
         * For details see http://download.oracle.com/javase/tutorial/uiswing/lookandfeel/plaf.html
         */
        try {
            for (javax.swing.UIManager.LookAndFeelInfo info :
javax.swing.UIManager.getInstalledLookAndFeels()) {
                if ("Nimbus".equals(info.getName())) {
                    javax.swing.UIManager.setLookAndFeel(info.getClassName());
                    break;
                }
            }
        } catch (ClassNotFoundException ex) {
```

```
            java.util.logging.Logger.getLogger(OgaMainFrame.class.getName()).log
(java.util.logging.Level.SEVERE, null, ex);
        } catch (InstantiationException ex) {
            java.util.logging.Logger.getLogger(OgaMainFrame.class.getName()).log
(java.util.logging.Level.SEVERE, null, ex);
        } catch (IllegalAccessException ex) {
            java.util.logging.Logger.getLogger(OgaMainFrame.class.getName()).log
(java.util.logging.Level.SEVERE, null, ex);
        } catch (javax.swing.UnsupportedLookAndFeelException ex) {
            java.util.logging.Logger.getLogger(OgaMainFrame.class.getName()).log
(java.util.logging.Level.SEVERE, null, ex);
        }
        //</editor-fold>

        /* Create and display the form */
        java.awt.EventQueue.invokeLater(new Runnable() {
            public void run() {
                new OgaMainFrame().setVisible(true);
            }
        });
    }
    // Variables declaration - do not modify//GEN-BEGIN:variables
    private javax.swing.JButton jButton1;
    private javax.swing.JMenu jMenu1;
    private javax.swing.JMenu jMenu2;
    private javax.swing.JMenu jMenu3;
    private javax.swing.JMenuBar jMenuBar1;
    private javax.swing.JScrollPane jScrollPane1;
    private javax.swing.JTabbedPane jTabbedPane1;
    private javax.swing.JTextPane jTextPane1;
    // End of variables declaration//GEN-END:variables
}
```

SearchDialog.java

```
/*
 * To change this template, choose Tools | Templates
 * and open the template in the editor.
 */
package oga;

/**
 *
 * @author jesuse
 */
public class SearchDialog extends javax.swing.JDialog {

    /**
     * Creates new form SearchDialog
     */
    Intermediator im;
    String[] matchings;
    String output;
    public SearchDialog(java.awt.Frame parent, boolean modal,Intermediator im) {
        super(parent, modal);
        initComponents();
        this.im=im;
    }

   public void setMachings(String[] matchs)
      {
      matchings=matchs;
      }

   public void append(String st)
     {
     this.jTextArea1.append(st);
     }

    public void setText(String rslts)
     {
     this.jTextArea1.setText(rslts);
     output=rslts;
     }

    /**
     * This method is called from within the constructor to initialize the form.
     * WARNING: Do NOT modify this code. The content of this method is always
     * regenerated by the Form Editor.
     */
    @SuppressWarnings("unchecked")
    // <editor-fold defaultstate="collapsed" desc="Generated Code">//GEN-BEGIN:initComponents
    private void initComponents() {

        jScrollPane1 = new javax.swing.JScrollPane();
        jTextArea1 = new javax.swing.JTextArea();

        setDefaultCloseOperation(javax.swing.WindowConstants.DISPOSE_ON_CLOSE);
        setTitle("Concept Search Results");

        jTextArea1.setColumns(20);
        jTextArea1.setRows(5);
        jTextArea1.addMouseListener(new java.awt.event.MouseAdapter() {
            public void mouseClicked(java.awt.event.MouseEvent evt) {
                jTextArea1MouseClicked(evt);
            }
        });
        jScrollPane1.setViewportView(jTextArea1);

        javax.swing.GroupLayout layout = new javax.swing.GroupLayout(getContentPane());
        getContentPane().setLayout(layout);
        layout.setHorizontalGroup(
```

```
            layout.createParallelGroup(javax.swing.GroupLayout.Alignment.LEADING)
            .addGroup(layout.createSequentialGroup()
                .addContainerGap()
                .addComponent(jScrollPane1, javax.swing.GroupLayout.DEFAULT_SIZE, 363, Short.MAX_VALUE)
                .addContainerGap())
        );
        layout.setVerticalGroup(
            layout.createParallelGroup(javax.swing.GroupLayout.Alignment.LEADING)
            .addGroup(javax.swing.GroupLayout.Alignment.TRAILING, layout.createSequentialGroup()
                .addGap(24, 24, 24)
                .addComponent(jScrollPane1, javax.swing.GroupLayout.DEFAULT_SIZE, 240, Short.MAX_VALUE)
                .addGap(36, 36, 36))
        );

        pack();
    }// </editor-fold>//GEN-END:initComponents

    private void jTextArea1MouseClicked(java.awt.event.MouseEvent evt) {//GEN-
FIRST:event_jTextArea1MouseClicked
        // TODO add your handling code here:
         if(jTextArea1.getCaretPosition()<output.length())
        {
        int caretPos = jTextArea1.getCaretPosition();
        int stringPosition = 0;
        int index=0;
        while((stringPosition<caretPos)) {
            stringPosition=stringPosition+matchings[index].length();
            index++;
            }
        index--;
        String[] vals = matchings[index].split(":");
        String[] vals2 = vals[1].split("\t");
        String newSelected = vals2[0].trim();
        im.processChoice(Integer.parseInt(newSelected));
        }
    }//GEN-LAST:event_jTextArea1MouseClicked

    /**
     * @param args the command line arguments
     */
    public static void main(String args[]) {
        /* Set the Nimbus look and feel */
        //<editor-fold defaultstate="collapsed" desc=" Look and feel setting code (optional) ">
        /* If Nimbus (introduced in Java SE 6) is not available, stay with the default look and feel.
         * For details see http://download.oracle.com/javase/tutorial/uiswing/lookandfeel/plaf.html
         */
        try {
            for (javax.swing.UIManager.LookAndFeelInfo info :
javax.swing.UIManager.getInstalledLookAndFeels()) {
                if ("Nimbus".equals(info.getName())) {
                    javax.swing.UIManager.setLookAndFeel(info.getClassName());
                    break;
                }
            }
        } catch (ClassNotFoundException ex) {
            java.util.logging.Logger.getLogger(SearchDialog.class.getName()).log
(java.util.logging.Level.SEVERE, null, ex);
        } catch (InstantiationException ex) {
            java.util.logging.Logger.getLogger(SearchDialog.class.getName()).log
(java.util.logging.Level.SEVERE, null, ex);
        } catch (IllegalAccessException ex) {
            java.util.logging.Logger.getLogger(SearchDialog.class.getName()).log
(java.util.logging.Level.SEVERE, null, ex);
        } catch (javax.swing.UnsupportedLookAndFeelException ex) {
            java.util.logging.Logger.getLogger(SearchDialog.class.getName()).log
(java.util.logging.Level.SEVERE, null, ex);
        }
```

```
        //</editor-fold>

        /* Create and display the dialog */
        java.awt.EventQueue.invokeLater(new Runnable() {
            public void run() {
            Intermediator im=new Intermediator() {
               public void processChoice(int test) {
               }
            };
                SearchDialog dialog = new SearchDialog(new javax.swing.JFrame(), true,im);
                dialog.addWindowListener(new java.awt.event.WindowAdapter() {
                    @Override
                    public void windowClosing(java.awt.event.WindowEvent e) {
                        System.exit(0);
                    }
                });
                dialog.setVisible(true);
            }
        });
    }
    // Variables declaration - do not modify//GEN-BEGIN:variables
    private javax.swing.JScrollPane jScrollPane1;
    private javax.swing.JTextArea jTextArea1;
    // End of variables declaration//GEN-END:variables
}
```

REFERENCES

Allones, J. L., Taboada, M., Martinez, D., Lozano, R., & Sobrido, M. J. (2013). SNOMED CT module-driven clinical archetype management. *J Biomed Inform*.

Altmuller, J., Palmer, L. J., Fischer, G., Scherb, H., & Wjst, M. (2001). Genomewide scans of complex human diseases: true linkage is hard to find. *Am J Hum Genet, 69*(5), 936-950.

Altshuler, D., Pollara, V. J., Cowles, C. R., Van Etten, W. J., Baldwin, J., Linton, L., et al. (2000). An SNP map of the human genome generated by reduced representation shotgun sequencing. [Research Support, Non-U.S. Gov't]. *Nature, 407*(6803), 513-516.

Amberger, J., Bocchini, C., & Hamosh, A. (2011). A new face and new challenges for Online Mendelian Inheritance in Man (OMIM(R)). [Research Support, N.I.H., Extramural]. *Hum Mutat, 32*(5), 564-567.

Amberger, J., Bocchini, C. A., Scott, A. F., & Hamosh, A. (2009). McKusick's Online Mendelian Inheritance in Man (OMIM). [Research Support, N.I.H., Extramural]. *Nucleic Acids Res, 37*(Database issue), D793-796.

Amos, C. I. (1994). Robust variance-components approach for assessing genetic linkage in pedigrees. *Am J Hum Genet, 54*(3), 535-543.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., et al. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. [Research Support, Non-U.S. Gov't Research Support, U.S. Gov't, P.H.S.]. *Nat Genet, 25*(1), 25-29.

Balding, D. J. (2006). A tutorial on statistical methods for population association studies. [Research Support, Non-U.S. Gov't Review]. *Nat Rev Genet, 7*(10), 781-791.

Banerjee-Basu, S., & Packer, A. (2010). SFARI Gene: an evolving database for the autism research community. [News]. *Dis Model Mech, 3*(3-4), 133-135.

Barrett, J. C. (2009). Haploview: Visualization and analysis of SNP genotype data. *Cold Spring Harb Protoc, 2009*(10), pdb ip71.

Barrett, J. C., Hansoul, S., Nicolae, D. L., Cho, J. H., Duerr, R. H., Rioux, J. D., et al. (2008). Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. [Meta-Analysis Research Support, N.I.H., ExtramuralResearch Support, Non-U.S. Gov't]. *Nat Genet, 40*(8), 955-962.

Baxevanis, A. D. (2012). Searching Online Mendelian Inheritance in Man (OMIM) for information on genetic loci involved in human disease. *Curr Protoc Hum Genet, Chapter 9*, Unit 9 13 11-10.

Becker, K. G., Barnes, K. C., Bright, T. J., & Wang, S. A. (2004). The genetic association database. [Letter]. *Nat Genet, 36*(5), 431-432.

Bedard, T., Lowry, R. B., & Sibbald, B. (2012). ICD-10 coding for congenital anomalies: a Canadian experience. *J Registry Manag, 39*(1), 4-7.

Bertram, L., Blacker, D., Mullin, K., Keeney, D., Jones, J., Basu, S., et al. (2000). Evidence for genetic linkage of Alzheimer's disease to chromosome 10q. [Research Support, Non-U.S. Gov'tResearch Support, U.S. Gov't, P.H.S.]. *Science, 290*(5500), 2302-2303.

Boyadjiev, S. A., & Jabs, E. W. (2000). Online Mendelian Inheritance in Man (OMIM) as a knowledgebase for human developmental disorders. [Research Support, U.S. Gov't, P.H.S.
Review]. *Clin Genet, 57*(4), 253-266.

Bray, P. F., Mathias, R. A., Faraday, N., Yanek, L. R., Fallin, M. D., Herrera-Galeano, J. E., et al. (2007). Heritability of platelet function in families with premature coronary artery disease. [Research Support, N.I.H., Extramural
Research Support, N.I.H., Intramural
Research Support, Non-U.S. Gov't]. *J Thromb Haemost, 5*(8), 1617-1623.

Chen, L., & Friedman, C. (2004). Extracting phenotypic information from the literature via natural language processing. [Evaluation StudiesResearch Support, N.I.H., Extramural Research Support, U.S. Gov't, Non-P.H.S.Research Support, U.S. Gov't, P.H.S.]. *Stud Health Technol Inform, 107*(Pt2),758 762.

Chian, L. T., & Wilgram, G. F. (1967). Tyrosinase inhibition: its role in suntanning and in albinism. *Science, 155*(3759), 198-200.

Christ, S. E. (2003). Asbjorn Folling and the discovery of phenylketonuria. [Biography Historical ArticlePortraits]. *J Hist Neurosci, 12*(1), 44-54.

Cunningham, J. M., Sellers, T. A., Schildkraut, J. M., Fredericksen, Z. S., Vierkant, R. A., Kelemen, L. E., et al. (2008). Performance of amplified DNA in an Illumina GoldenGate BeadArray assay. [Comparative StudyResearch Support, N.I.H.,

ExtramuralResearch Support, Non-U.S. Gov't]. *Cancer Epidemiol Biomarkers Prev, 17*(7), 1781-1789.

Day-Richter, J., Harris, M. A., Haendel, M., Gene Ontology, O. B. O. E. W. G., & Lewis, S. (2007). OBO-Edit--an ontology editor for biologists. *Bioinformatics, 23*(16), 2198-2200.

Du, Y., Xie, J., Chang, W., Han, Y., & Cao, G. (2012). Genome-wide association studies: inherent limitations and future challenges. [Research Support, Non-U.S. Gov't]. *Front Med, 6*(4), 444-450.

Easton, D. F., Bishop, D. T., Ford, D., & Crockford, G. P. (1993). Genetic linkage analysis in familial breast and ovarian cancer: results from 214 families. The Breast Cancer Linkage Consortium. [Research Support, Non-U.S. Gov't]. *Am J Hum Genet, 52*(4), 678-701.

Eden, E., Navon, R., Steinfeld, I., Lipson, D., & Yakhini, Z. (2009). GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. [Research Support, Non-U.S. Gov't]. *BMC Bioinformatics, 10*, 48.

Falconer, D. (1993). Quantitative genetics in Edinburgh: 1947-1980. [Autobiography Biography Historical Article]. *Genetics, 133*(2), 137-142.

Gayther, S. A., Warren, W., Mazoyer, S., Russell, P. A., Harrington, P. A., Chiano, M., et al. (1995). Germline mutations of the BRCA1 gene in breast and ovarian cancer families provide evidence for a genotype-phenotype correlation. [Research Support, Non-U.S. Gov't]. *Nat Genet, 11*(4), 428-433.

Groth, P., Pavlova, N., Kalev, I., Tonov, S., Georgiev, G., Pohlenz, H. D., et al. (2007). PhenomicDB: a new cross-species genotype/phenotype resource. [Research Support, Non-U.S. Gov't]. *Nucleic Acids Res, 35*(Database issue), D696-699.

Groth, P., Weiss, B., Pohlenz, H. D., & Leser, U. (2008). Mining phenotypes for gene function prediction. *BMC Bioinformatics, 9*, 136.

Gu, L., Long, J., Yan, Y., Chen, Q., Pan, R., Xie, X., et al. (2013). HTR2A-1438A/G polymorphism influences the risk of schizophrenia but not bipolar disorder or major depressive disorder: A meta-analysis. *J Neurosci Res*.

Hamosh, A., Scott, A. F., Amberger, J. S., Bocchini, C. A., & McKusick, V. A. (2005). Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. [Research Support, Non-U.S. Gov'tResearch Support, U.S. Gov't, P.H.S.]. *Nucleic Acids Res, 33*(Database issue), D514-517.

Hanis, C. L., Boerwinkle, E., Chakraborty, R., Ellsworth, D. L., Concannon, P., Stirling, B., et al. (1996). A genome-wide search for human non-insulin-dependent (type 2) diabetes genes reveals a major susceptibility locus on chromosome 2. [Research Support, Non-U.S. Gov't Research Support, U.S. Gov't, P.H.S.]. *Nat Genet, 13*(2), 161-166.

Harris, M., Martin, J. M., Peden, J. F., & Rawlings, C. J. (2003). SNP cherry picker: maximizing the chance of finding an association with a disease SNP. *Bioinformatics, 19*(16), 2141-2143.

Haseman, J. K., & Elston, R. C. (1972). The investigation of linkage between a quantitative trait and a marker locus. *Behav Genet, 2*(1), 3-19.

Hearne, C. M., Ghosh, S., & Todd, J. A. (1992). Microsatellites for linkage analysis of genetic traits. [Research Support, Non-U.S. Gov'tReview]. *Trends Genet, 8*(8), 288-294.

Herrera-Galeano, J. E., Becker, D. M., Wilson, A. F., Yanek, L. R., Bray, P., Vaidya, D., et al. (2008). A novel variant in the platelet endothelial aggregation receptor-1 gene is associated with increased platelet aggregability. [Research Support, N.I.H., Extramural]. *Arterioscler Thromb Vasc Biol, 28*(8), 1484-1490.

Ho, P. S., Ho, K. K., Huang, W. S., Yen, C. H., Shih, M. C., Shen, L. H., et al. (2012). Association study of serotonin transporter availability and SLC6A4 gene polymorphisms in patients with major depression. *Psychiatry Res*.

Jeffreys, A. J., Wilson, V., & Thein, S. L. (1985). Individual-specific 'fingerprints' of human DNA. [Research Support, Non-U.S. Gov't]. *Nature, 316*(6023), 76-79.

Jin, C., Xu, W., Yuan, J., Wang, G., & Cheng, Z. (2013). Meta-analysis of association between the -1438A/G (rs6311) polymorphism of the serotonin 2A receptor gene and major depressive disorder. [Research Support, Non-U.S. Gov't]. *Neurol Res, 35*(1), 7-14.

Johnson, A. D., Yanek, L. R., Chen, M. H., Faraday, N., Larson, M. G., Tofler, G., et al. (2010). Genome-wide meta-analyses identifies seven loci associated with platelet aggregation in response to agonists. [Research Support, N.I.H., Extramural Research Support, Non-U.S. Gov't]. *Nat Genet, 42*(7), 608-613.

Kang, H. J., Kim, J. M., Stewart, R., Kim, S. Y., Bae, K. Y., Kim, S. W., et al. (2013). Association of SLC6A4 methylation with early adversity, characteristics and outcomes in depression. *Prog Neuropsychopharmacol Biol Psychiatry, 44C*, 23-28.

Kathiresan, S., Melander, O., Guiducci, C., Surti, A., Burtt, N. P., Rieder, M. J., et al. (2008). Six new loci associated with blood low-density lipoprotein cholesterol, high-density lipoprotein cholesterol or triglycerides in humans. [Comparative Study Meta-Analysis Research Support, N.I.H., Extramural
Research Support, Non-U.S. Gov't]. *Nat Genet, 40*(2), 189-197.

Kwon, J. M., & Goate, A. M. (2000). The candidate gene approach. [Review]. *Alcohol Res Health, 24*(3), 164-168.

Lamprecht, A. L., Naujokat, S., Margaria, T., & Steffen, B. (2011). Semantics-based composition of EMBOSS services. *J Biomed Semantics, 2 Suppl 1*, S5.

Laulederkind, S. J., Tutaj, M., Shimoyama, M., Hayman, G. T., Lowry, T. F., Nigam, R., et al. (2012). Ontology searching and browsing at the Rat Genome Database. [Research Support, N.I.H., Extramural]. *Database (Oxford), 2012*, bas016.

Lee, J. S., Katari, G., & Sachidanandam, R. (2005). GObar: a gene ontology based analysis and visualization tool for gene sets. [Research Support, Non-U.S. Gov't]. *BMC Bioinformatics, 6*, 189.

Lewis, C. M., & Knight, J. (2012). Introduction to genetic association studies. [Review]. *Cold Spring Harb Protoc, 2012*(3), 297-306.

Li, M. J., Wang, P., Liu, X., Lim, E. L., Wang, Z., Yeager, M., et al. (2012). GWASdb: a database for human genetic variants identified by genome-wide association studies. [Research Support, N.I.H., Intramural Research Support, Non-U.S. Gov't]. *Nucleic Acids Res, 40*(Database issue), D1047-1054.

Lin, H. Y., Chen, Y. A., Tsai, Y. Y., Qu, X., Tseng, T. S., & Park, J. Y. (2012). TRM: a powerful two-stage machine learning approach for identifying SNP-SNP interactions. [Comparative StudyEvaluation Studies Research Support, N.I.H., Extramural
Research Support, Non-U.S. Gov't]. *Ann Hum Genet, 76*(1), 53-62.

Lin, Y., Xiang, Z., & He, Y. (2011). Brucellosis Ontology (IDOBRU) as an extension of the Infectious Disease Ontology. *J Biomed Semantics, 2*(1), 9.

Liu, T. F., Sung, W. K., Li, Y., Liu, J. J., Mittal, A., & Mao, P. L. (2005). Effective algorithms for tag SNP selection. [Comparative Study]. *J Bioinform Comput Biol, 3*(5), 1089-1106.

Lohoff, F. W., Narasimhan, S., & Rickels, K. (2012). Interaction between polymorphisms in serotonin transporter (SLC6A4) and serotonin receptor 2A (HTR2A) genes predict treatment response to venlafaxine XR in generalized anxiety disorder. *Pharmacogenomics J*.

Lussier, Y., Borlawsky, T., Rappaport, D., Liu, Y., & Friedman, C. (2006). PhenoGO: assigning phenotypic context to gene ontology annotations with natural language processing. [Evaluation Studies Research Support, N.I.H., Extramural]. *Pac Symp Biocomput*, 64-75.

Mailman, M. D., Feolo, M., Jin, Y., Kimura, M., Tryka, K., Bagoutdinov, R., et al. (2007). The NCBI dbGaP database of genotypes and phenotypes. [Research Support, N.I.H., Intramural]. *Nat Genet, 39*(10), 1181-1186.

Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., et al. (2009). Finding the missing heritability of complex diseases. [Review]. *Nature, 461*(7265), 747-753.

Mathias, R. A., Kim, Y., Sung, H., Yanek, L. R., Mantese, V. J., Hererra-Galeano, J. E., et al. (2010). A combined genome-wide linkage and association approach to find susceptibility loci for platelet function phenotypes in European American and African American families with coronary artery disease. [Research Support, N.I.H., Extramural Research Support, Non-U.S. Gov't]. *BMC Med Genomics, 3*, 22.

McKusick, V. A. (2006). A 60-year tale of spots, maps, and genes. [Historical Article Review]. *Annu Rev Genomics Hum Genet, 7*, 1-27.

Miki, Y., Katagiri, T., Kasumi, F., Yoshimoto, T., & Nakamura, Y. (1996). Mutation analysis in the BRCA2 gene in primary breast cancers. [Research Support, Non-U.S. Gov't]. *Nat Genet, 13*(2), 245-247.

Mullis, K., Faloona, F., Scharf, S., Saiki, R., Horn, G., & Erlich, H. (1986). Specific enzymatic amplification of DNA in vitro: the polymerase chain reaction. *Cold Spring Harb Symp Quant Biol, 51 Pt 1*, 263-273.

Musen, M. A., Noy, N. F., Shah, N. H., Whetzel, P. L., Chute, C. G., Story, M. A., et al. (2012). The National Center for Biomedical Ontology. [Research Support, N.I.H., Extramural]. *J Am Med Inform Assoc, 19*(2), 190-195.

Musen, M. A., Shah, N. H., Noy, N. F., Dai, B. Y., Dorf, M., Griffith, N., et al. (2008). BioPortal: ontologies and data resources with the click of a mouse. *AMIA Annu Symp Proc*, 1223-1224.

Musumba, C. O., Jorgensen, A., Sutton, L., Van Eker, D., Zhang, E., O'Hara, N., et al. (2013). CYP2C19*17 gain-of-function polymorphism is associated with peptic ulcer disease. [Multicenter StudyResearch Support, Non-U.S. Gov't]. *Clin Pharmacol Ther, 93*(2), 195-203.

Myocardial Infarction Genetics, C., Kathiresan, S., Voight, B. F., Purcell, S., Musunuru, K., Ardissino, D., et al. (2009). Genome-wide association of early-onset myocardial infarction with single nucleotide polymorphisms and copy number variants. [Research Support, N.I.H., ExtramuralResearch Support, Non-U.S. Gov't]. *Nat Genet, 41*(3), 334-341.

Nakamura, Y., Leppert, M., O'Connell, P., Wolff, R., Holm, T., Culver, M., et al. (1987). Variable number of tandem repeat (VNTR) markers for human gene mapping. [Research Support, Non-U.S. Gov't]. *Science, 235*(4796), 1616-1622.

Noy, N. F., Shah, N. H., Whetzel, P. L., Dai, B., Dorf, M., Griffith, N., et al. (2009). BioPortal: ontologies and integrated data resources at the click of a mouse. [Research Support, N.I.H., Extramural]. *Nucleic Acids Res, 37*(Web Server issue), W170-173.

O'Donoghue, M. L. (2011). CYP2C19 genotype and proton pump inhibitors in clopidogrel-treated patients: does it take two to tango? [Comment Editorial]. *Circulation, 123*(5), 468-470.

Oliphant, A., Barker, D. L., Stuelpnagel, J. R., & Chee, M. S. (2002). BeadArray technology: enabling an accurate, cost-effective approach to high-throughput genotyping. [Research Support, U.S. Gov't, P.H.S.]. *Biotechniques, Suppl*, 56-58, 60-51.

Palotie, A., Widen, E., & Ripatti, S. (2012). From genetic discovery to future personalized health research. *N Biotechnol*.

Papenberg, G., Backman, L., Nagel, I. E., Nietfeld, W., Schroder, J., Bertram, L., et al. (2013). Dopaminergic Gene Polymorphisms Affect Long-term Forgetting in Old Age: Further Support for the Magnification Hypothesis. *J Cogn Neurosci, 25*(4), 571-579.

Pearson, P., Francomano, C., Foster, P., Bocchini, C., Li, P., & McKusick, V. (1994). The status of online Mendelian inheritance in man (OMIM) medio 1994. *Nucleic Acids Res, 22*(17), 3470-3473.

Pickett, B. E., Sadat, E. L., Zhang, Y., Noronha, J. M., Squires, R. B., Hunt, V., et al. (2012). ViPR: an open bioinformatics database and analysis resource for virology research. [Research Support, N.I.H., Extramural]. *Nucleic Acids Res, 40*(Database issue), D593-598.

Portales-Casamar, E., Ch Ng, C., Lui, F., St-Georges, N., Zoubarev, A., Lai, A. Y., et al. (2013). Neurocarta: aggregating and sharing disease-gene relations for the neurosciences. *BMC Genomics, 14*(1), 129.

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., & Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. [Research Support, Non-U.S. Gov't]. *Nat Genet, 38*(8), 904-909.

Price, A. L., Zaitlen, N. A., Reich, D., & Patterson, N. (2010). New approaches to population stratification in genome-wide association studies. [Review]. *Nat Rev Genet, 11*(7), 459-463.

Province, M. A., & Borecki, I. B. (2008). Gathering the gold dust: methods for assessing the aggregate impact of small effect genes in genomic scans. [Research Support, N.I.H., Extramural]. *Pac Symp Biocomput*, 190-200.

Rebbeck, T. R., Friebel, T., Lynch, H. T., Neuhausen, S. L., van 't Veer, L., Garber, J. E., et al. (2004). Bilateral prophylactic mastectomy reduces breast cancer risk in BRCA1 and BRCA2 mutation carriers: the PROSE Study Group. [Multicenter Study Research Support, Non-U.S. Gov't Research Support, U.S. Gov't, Non-P.H.S. Research Support, U.S. Gov't, P.H.S.]. *J Clin Oncol, 22*(6), 1055-1062.

Robinson, P. N., & Mundlos, S. (2010). The human phenotype ontology. [Review]. *Clin Genet, 77*(6), 525-534.

Schriml, L. M., Arze, C., Nadendla, S., Chang, Y. W., Mazaitis, M., Felix, V., et al. (2012). Disease Ontology: a backbone for disease semantic integration. [Research Support, N.I.H., Extramural]. *Nucleic Acids Res, 40*(Database issue), D940-946.

Sham, P. C., & Purcell, S. (2001). Equivalence between Haseman-Elston and variance-components linkage analyses for sib pairs. [Research Support, Non-U.S. Gov't Research Support, U.S. Gov't, P.H.S.]. *Am J Hum Genet, 68*(6), 1527-1532.

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., et al. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. [Research Support, Non-U.S. Gov't Research Support, U.S. Gov't, Non-P.H.S. Research Support, U.S. Gov't, P.H.S.]. *Genome Res, 13*(11), 2498-2504.

Smith, C. L., & Eppig, J. T. (2012). The Mammalian Phenotype Ontology as a unifying standard for experimental and high-throughput phenotyping data. [Research Support, N.I.H., Extramural]. *Mamm Genome, 23*(9-10), 653-668.

Spencer, T. J., Biederman, J., Faraone, S. V., Madras, B. K., Bonab, A. A., Dougherty, D. D., et al. (2012). Functional Genomics of Attention-Deficit/Hyperactivity Disorder (ADHD) Risk Alleles on Dopamine Transporter Binding in ADHD and Healthy Control Subjects. *Biol Psychiatry*.

Strofilas, A., Lagoudianakis, E. E., Seretis, C., Pappas, A., Koronakis, N., Keramidaris, D., et al. (2012). Association of helicobacter pylori infection and colon cancer. *J Clin Med Res, 4*(3), 172-176.

Stumpf, M. P., & Goldstein, D. B. (2003). Demography, recombination hotspot intensity, and the block structure of linkage disequilibrium. [Research Support, Non-U.S. Gov't]. *Curr Biol, 13*(1), 1-8.

Sun, H., Fang, H., Chen, T., Perkins, R., & Tong, W. (2006). GOFFA: gene ontology for functional analysis--a FDA gene ontology tool for analysis of genomic and proteomic data. *BMC Bioinformatics, 7 Suppl 2*, S23.

Suzuki, Y., Orita, M., Shiraishi, M., Hayashi, K., & Sekiya, T. (1990). Detection of ras gene mutations in human lung cancers by single-strand conformation polymorphism analysis of polymerase chain reaction products. [Research Support, Non-U.S. Gov't]. *Oncogene, 5*(7), 1037-1043.

Thorisson, G. A., Smith, A. V., Krishnan, L., & Stein, L. D. (2005). The International HapMap Project Web site. [Research Support, Non-U.S. Gov't Research Support, U.S. Gov't, Non-P.H.S.]. *Genome Res, 15*(11), 1592-1593.

Todd, J. A. (2006). Statistical false positive or true disease pathway? [Research Support, Non-U.S. Gov't]. *Nat Genet, 38*(7), 731-733.

Tortorella, A., Monteleone, P., Martiadis, V., Perris, F., & Maj, M. (2007). The 3111T/C polymorphism of the CLOCK gene confers a predisposition to a lifetime lower body weight in patients with anorexia nervosa and bulimia nervosa: a preliminary study. [Comparative Study]. *Am J Med Genet B Neuropsychiatr Genet, 144B*(8), 992-995.

Tweedie, S., Ashburner, M., Falls, K., Leyland, P., McQuilton, P., Marygold, S., et al. (2009). FlyBase: enhancing Drosophila Gene Ontology annotations. [Research Support, N.I.H., Extramural
Research Support, Non-U.S. Gov't]. *Nucleic Acids Res, 37*(Database issue), D555-559.

Vignini, A., Giulietti, A., Nanetti, L., Raffaelli, F., Giusti, L., Mazzanti, L., et al. (2013). Alzheimer's disease And Diabetes: New Insights and Unifying Therapies. *Curr Diabetes Rev*.

Visscher, P. M. (2008). Sizing up human height variation. [Comment News]. *Nat Genet, 40*(5), 489-490.

Walker, L., Starks, H., West, K. M., & Fullerton, S. M. (2011). dbGaP data access requests: a call for greater transparency. [Research Support, N.I.H., Extramural]. *Sci Transl Med, 3*(113), 113cm134.

Wang, Z., Wang, Y., Tan, K. L., Wong, L., & Agrawal, D. (2011). eCEO: an efficient Cloud Epistasis cOmputing model in genome-wide association study. [Evaluation Studies Research Support, Non-U.S. Gov't]. *Bioinformatics, 27*(8), 1045-1051.

Weedon, M. N., & Frayling, T. M. (2008). Reaching new heights: insights into the genetics of human stature. [Review]. *Trends Genet, 24*(12), 595-603.

Whetzel, P. L., Noy, N. F., Shah, N. H., Alexander, P. R., Nyulas, C., Tudorache, T., et al. (2011). BioPortal: enhanced functionality via new Web services from the National Center for Biomedical Ontology to access and use ontologies in software applications. [Research Support, N.I.H., Extramural]. *Nucleic Acids Res, 39*(Web Server issue), W541-545.

Wooten, E. C., & Huggins, G. S. (2011). Mind the dbGAP: the application of data mining to identify biological mechanisms. [Research Support, N.I.H., Extramural Research Support, Non-U.S. Gov't]. *Mol Interv, 11*(2), 95-102.

Yamamoto, N., Murakami, H., Nishina, T., Hirashima, T., Sugio, K., Muro, K., et al. (2013). The effect of CYP2C19 polymorphism on the safety, tolerability, and pharmacokinetics of tivantinib (ARQ 197): results from a phase I trial in advanced solid tumors. *Ann Oncol*.

Yu, W., Clyne, M., Khoury, M. J., & Gwinn, M. (2010). Phenopedia and Genopedia: disease-centered and gene-centered views of the evolving knowledge of human genetic associations. *Bioinformatics, 26*(1), 145-146.

Zagouri, F., Chrysikos, D. T., Sergentanis, T. N., Giannakopoulou, G., Zografos, C. G., Papadimitriou, C. A., et al. (2013). Prophylactic mastectomy: an appraisal. *Am Surg, 79*(2), 205-212.

Zeggini, E., Scott, L. J., Saxena, R., Voight, B. F., Marchini, J. L., Hu, T., et al. (2008). Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. [Meta-AnalysisResearch Support, N.I.H., Extramural Research Support, N.I.H., Intramural Research Support, Non-U.S. Gov't]. *Nat Genet, 40*(5), 638-645.

Zhang, Y., De, S., Garner, J. R., Smith, K., Wang, S. A., & Becker, K. G. (2010). Systematic analysis, comparison, and integration of disease based human genetic association data and mouse genetic phenotypic information. [Research Support, N.I.H., Intramural]. *BMC Med Genomics, 3*, 1.

Zheng, Q., & Wang, X. J. (2008). GOEAST: a web-based software toolkit for Gene Ontology enrichment analysis. [Research Support, Non-U.S. Gov't]. *Nucleic Acids Res, 36*(Web Server issue), W358-363.

CURRICULUM VIATE

Jesus Enrique Herrera-Galeano graduated from Johns Hopkins, Baltimore, Maryland, in 2005. He holds two master's degrees, the first one in genetics and the second one in Computer Science from Johns Hopkins. He was part of the faculty at the School of Medicine at Johns Hopkins for several years. He was part of the Bioinformatics team at the Center for Immune and Infectious diseases at Columbia University for two years and he is currently a Bioinformatics scientist at the Naval Medical Research Center in Frederick, MD.