dbAARD: DATABASE OF AGING AND AGE-RELATED DISORDERS

VAIBHAV MATHUR

18/BINF/2010

SUPERVISOR

Dr. YASHA HASIJA

DEPARTMENT OF BIOTECHNOLOGY

DELHI TECHNOLOGICAL UNIVERSITY

DELHI

This thesis is submitted in partial requirement for the degree of Master of Technology in Bioinformatics

CERTIFICATE

This is to certify that Mr. Vaibhav Mathur, Roll no. 18/BINF/2010, has completed this project entitled "*dbAARD: Database of Aging and Age-Related Disorders*" under my guidance in partial fulfillment of his M.Tech Bioinformatics degree at Delhi Technological University, Delhi.

Date:

Dr. Yasha Hasija
Asst. Professor
Dept of Biotechnology
Delhi Technological University
Delhi

DECLARATION

This is to certify that the Thesis entitled "*dbAARD: Database of Aging and Age-Related Disorders*" which is submitted by me in partial fulfillment of the requirement for the award of degree M.Tech. in Bioinformatics to Delhi Technological University, Delhi comprises only my original work and due acknowledgement has been made in the text to all other material used.

Date:

Vaibhav Mathur

Roll no. 18/BINF/2010

Department of Biotechnology

Delhi Technological University,

Delhi

ACKNOWLEDGEMENT

Firstly, I would like to thank Dr. S. Maji, Head, Department of Biotechnology, Delhi Technological University, Delhi, who has always encouraged and supported me. Next, I would like to thank my project guide Dr. Yasha Hasija, for her invaluable guidance which helped me a lot during the course of the project. She was instrumental in providing all the technical and moral support. I would also like to thank Nitish Sharma, Heena Dhiman, Deepika Jaggi, Sonam Arora and Unnati Goel, students of the department, who helped me in compiling the database of disease-SNP associations. Lastly, I wish to thank my family and colleagues at the department of Biotechnology who have always encouraged me to give my best.

Vaibhav Mathur

Roll no. 18/BINF/2010

Department of Biotechnology

Delhi Technological University,

Delhi

CONTENTS

LIST OF ABBREVIATIONS

AGE	:	Advanced Glycation End products
AMAR	:	Anti-aging Medicine And Research
AMD	:	Age related Macular Degeneration
ARD	:	Age Related Disorders
CGI	:	Common Gateway Interface
CSS	:	Cascading Style Sheet
CSV	:	Comma Separated Values
DNA	:	Deoxy riboNucleic Acid
ER	:	Entity Relation
GUI	:	Graphical User Interface
GWAS	:	Genome Wide Association Study
HTML	:	HyperText Markup Language
ICMR	:	Indian Council of Medical Research
IGF	:	Insulin-like Growth Factor
IIS	:	Insulin/Insulin-like growth factor 1 Signaling
NHGRI	:	National Human Genome Research Institute
OMIM	:	Online Mendelian Inheritance in Man
PCR	:	Polymerase Chain Reaction
PMID	:	PubMed-Indexed for Medline
ROS	:	Reactive Oxygen Species
SNP	:	Single Nucleotide Polymorphism
SQL	:	Structured Query Language
TOR	:	Target Of Rapamycin
UTR	:	UnTranslated Region
WTCCC	:	Wellcome Trust Case-Control Consortium
XAMPP	:	X (any one of 4 operating systems) Apache, MySQL, PHP, Perl

LIST OF FIGURES

LIST OF TABLES

LIST OF APPENDICES

ABSTRACT

Most old people suffer and die not of aging *per se*, but various age-related disorders (ARDs) like cancers, Alzheimer's disease, osteoporosis etc. It is well known that though the average global life expectancy has increased, the quality of life in old age has not. Hence researchers are now focusing not on increasing lifespan or "curing" aging itself, but finding the cause and treatment for different age-related disorders that afflict a growing number of people. The completion of the Human Genome Project and the HapMap project has provided us with the data necessary to link genome variations to susceptibility to complex disorders. Many genome-wide association studies (GWAS) have been conducted on the different age-related disorders to find single nucleotide polymorphisms (SNPs) that are responsible for causing these diseases. Not all these studies are consistent with each other, but we now have a far greater knowledge of disease-variation associations than we did before. However, a repository for storing all the associations detected by these GWAS has been sorely lacking in the field of aging research. This work is aimed at filling that void. A disease-SNP list of 34 age-related disorders and the corresponding genetic variations (SNPs) that have been associated with them was prepared. This list includes information on the location and ethnicity of the population under study and the p value or odds- ratio that has been reported in the study. The variants were mapped to their reference sequences at the genomic, transcriptomic as well as protein levels. In the case of missense mutations, the codon and amino acid change as well as the position were also mentioned. The data was compiled from various publicly available databases like the NHGRI GWAS catalogue, GWAS Central, OMIM, dbSNP and others, as well as literature searches. A relational data model was then prepared to store the information in the list. The resulting database created in MySQL is in the third normal form (3NF) to make it more efficient and reduce data redundancy. A web-based graphical-user interface was developed using HTML to query the database. The interface is easy-to-use and enables users to find the required information by using different filters. The interface connects to the database by a Perl CGI script.

This work is aimed at facilitating analysis of the large number of variants associated with age- related disorders. Such analysis may provide cues for deciphering the biology of ageing, thereby prioritizing drug targets for age-related disorders. The database is available as a free resource on the web at http://dbaard.dce.edu.

INTRODUCTION

Aging is an inevitable part of life. Throughout history, humans have tried to delay aging and achieve immortality. A well-known example of this effort is the search for the elixir of youth or the elixir of immortality, whose mention can be found in the mythologies of all ancient civilizations, like the Greek [1], Chinese [2] or Middle-Eastern [3]. In India, this mythical elixir is popularly known as *amrit*. Even now, long living populations continue to fascinate us, and have been the subject of various studies [4]. Prominent among these populations are the Hunza in Pakistan [5], the Vilcabamba in Ecuador [6] and the Abkhasia in Russia [7].

Aging is increasingly becoming a hot topic for research around the world as aging is becoming a worldwide social and economic problem. Population ageing is unprecedented, without parallel in human history—and the twenty-first century will witness even more rapid ageing than did the twentieth century (Appendix I). If the predictions of a recent United Nations report come true, the global population 60 or over will reach nearly 2 billion by the year 2050 [8] (Appendix II).

For developing countries, the problem is even more acute. The elderly population in India is expected to grow at more than 2.5% a year, on an average, for the next 40 years. India will have more than 300 million people over the age of 60 years by 2050 (Appendix III). However, India has still not shown much interest in the field of aging research. A government organization, The Indian Council of Medical Research (ICMR) has taken at least a minor initiative to promote aging research. Another organization, AMAR-India (Anti-aging Medicine And Research India), is a national non-profit educational medical organization based in Mumbai. It was incorporated in year 2007 in Mumbai to achieve the goal of increasing awareness in the field of Age management/Anti-aging medicine for early detection, prevention, reversal, and treatment of age related diseases and disorders not only to increase the life span but also to improve the quality of life so that the ideal of Healthy India is finally realized.

Population aging has caused a shift in age structure of the population, and the impact of this shift is profound enough to affect all spheres of human activity- economic, political and social. For instance, traditional support systems, like the intergenerational family, are breaking down. It has been shown that these systems are important for the well-being of both the older and younger generations [9]. This is even truer when the family size decreases and women take up jobs outside the home. Also, the social security systems need reforms as pensions and other retirement benefits are extended to more people and for longer periods of time [10, 11]. Moreover, since older people are more vulnerable to chronic disease [12, 13], medical costs will also increase and so will the demand for medical services.

Due to advancements in healthcare, average life expectancy at birth of the world is now more than 67 years [14]. It is quite common to see individuals living to 80 years or more, a fact that would have been a cause of awe and disbelief in the not-so-distant past. However, the quality of life in old age is usually far worse than in youth, because the elderly often suffer from various age-related disorders (ARDs), like arteriosclerosis, diabetes, dementia, osteoporosis, osteoarthritis and cancer [15]. Age-related disorders are the most common form of death in old age, and even in those who escape disease, the cause of death may be traced to subtle tissue atrophies, neuropathies or microvascular leakage [16]. Therefore, even as the proportion of old people (>60 years) rises in the world [17] (Appendix II), the focus of research has shifted from trying to delay aging or achieving immortality, to achieving "healthy" or "successful" aging. Healthy aging is said to comprise of the following [18]:

1. Low probability of disease or disability;
2. High cognitive and physical function capacity;
3. Active engagement with life

though a greater number of people self-report successful ageing than those that strictly meet these criteria [19].

Keeping the definition of "successful" aging in mind, it is pertinent to ask why some people live longer, and more "successfully", than others. Centenarians are persons who live to an age of 100 or above. Surprisingly, many of them also seem to be vibrant and full of life even in this advanced age [20]. Research on centenarians has been going on for some years and has focused on finding what biological, sociological and psychological factors they possess that enables them to survive for so long [21, 22].

One of the factors that play is crucial role in the aging process and determines how successfully an individual reaches old age is his genetic constitution. However, the role of genetics in aging and longevity is complex. It is long known that humans live significantly longer than lower animals like mice or fruit flies [23, 24], but until recently the reasons were obscure. Also, the relationship between aging and ARDs was debatable [25, 26]. Due to this, it may not be enough to successfully cure overt disease to "cure" aging. To fully understand how the pathogenesis of age-related relates to the basic molecular processes of aging, it is important to unravel how disease processes intersect with the basic aging cause(s). Designing effective aging interventions or "cures" will necessarily require multidisciplinary teams of clinicians and basic scientists working in cooperation with the pharmaceutical industry and regulatory agencies due to the complexity of the task [27].

Since the completion of the Human Genome Project researchers have a powerful new way at looking at the underlying causes of aging and age-related disorders. Before the completion of the project, knowledge about human genetic variation was limited mainly to the heterochromatin polymorphisms, large enough to be visible in the light microscope, and the single nucleotide polymorphisms (SNPs) identified by traditional PCR-based DNA sequencing [28]. The human genome project acted as a catalyst and greatly accelerated the advancement of sequencing technologies [29]. In the past five-seven years, the rapid development and expanded use of microarray technologies, including oligonucleotide array comparative genomic hybridization and SNP genotyping arrays, as well as next-generation sequencing with "paired-end" methods, has enabled a whole-genome analysis with almost unlimited resolution [28]. This has made it easier for genome-wide association studies to be conducted to search for variants found in various complex disorders [30]. Though these studies have greatly increased our knowledge about the variations responsible for disease, the mechanisms underlying these diseases still remain unclear and all the variations have not been cataloged [31]. Such studies generate a large amount of data in the form of SNPs or markers that are found to be associated to a particular disease, along-with the p-value or odds-ratio which quantifies the significance of the prediction. Ever-improving sequencing technologies make it possible to conduct such studies on larger scales, involving more patients and covering the genome more extensively. This also leads to an increase in the amount and rate of data being produced. A need was felt for a database to hold all the information from the various genome-wide association studies being conducted worldwide on the different age-related disorders. A graphical user interface was also needed to enable non-programmers to query the database for relevant information. Finally, this database had to be a free resource so that the information could be easily accessed by anybody.

Since a satisfactory resource could not be found freely available on the internet, this work is an effort to compile all the SNPs found to be involved in the different ARDs in GWA studies in one place to facilitate analysis. An effort has been made to make the database as comprehensive as possible. A graphical user interface has also been designed to enable easy querying of the database. Mining the database for clues to the genetic causes of ARDs may allow us to shed more light on the similarities and differences between the various ARDs and also between ARDs and aging. This information may help us to identify novel drug targets for ARDs that are found to play a critical role in the development of these diseases. The links between ARDs and aging may eventually help us unravel the biology of aging itself.

REVIEW OF LITERATURE

Though there has been a great interest in populations that have been traditionally known to be long-living, like the Hunza, but there was no systematic study of the aging process till the 20th century. Modern ageing research is considered to have begun more than a century ago with Max Rubner and Raymond Pearl who discovered that metabolic rate is the key regulator of aging in 1908 [32]. It was their finding that triggered a new interest in the biology of aging. Since then, various theories of ageing have been proposed. Some of them are mentioned below:

1. **Mutation accumulation** (1952) [33]: This theory, formulated by Peter Medawar in 1952, is believed to be the first modern, successful theory of mammal ageing. According to this, the mechanism of aging involves random, detrimental mutations of a kind that happen to show their effect only late in life. Unlike most detrimental mutations, these are not efficiently weeded out by natural selection. This is because nature is highly competitive, and almost all animals in nature die before they attain old age. Therefore, there is not much reason why the body should remain fit for a very long time as there is not much selection pressure for traits that would maintain viability past the time when most animals would be dead anyway, killed by predators or by accident or disease. Hence these mutations would 'accumulate' and, perhaps, cause all the decline and damage that we associate with ageing.

2. **Free radical theory** (1956) [34]: According to this theory, somatic damage is mainly caused by the accumulation of reactive oxygen species (ROS). ROS are the mainly the by-products of respiration and other metabolic processes. Also, the reducing sugars tend to react with free amino groups and carbohydrates, thus forming advanced glycation end products (AGEs) which are extremely difficult to degrade. They accumulate in proteins like collagen and elastin, thereby increasing the stiffness of blood vessels, joints and the bladder, and impair function in the kidney, heart, retina and other organs.

3. **Antagonistic pleiotropy** (1957) [35]: Pleiotropy refers to the phenomenon of one gene having two or more effects on the phenotype. According to G.C. Williams who postulated this theory, one of these effects is beneficial and another is detrimental. Basically this refers to genes that offer benefits early in life, but exact a cost later on. For example, enhanced early fertility could be selected even if it came with a cost that included decline and death later on.

4. **Hayflick Limit Theory** (1961) [36]: It suggests that the human cell is limited in the number of times it can divide. Dr. Hayflick showed that the human cells ability to divide is limited to approximately 50-times, after which they simply stop dividing (and hence die).

5. **Disposable soma theory** (1977) [37]: This theory was formulated by T.B. Kirkwood and proposes that the body must budget the amount of energy available to it. The body uses food energy for many purposes like metabolism, reproduction, and repair and maintenance. With a finite supply of food, the body must compromise, and do none of these things quite as well as it would like. It is the compromise in allocating energy to the repair function that causes the body gradually to deteriorate with age.

While these theories have been around for the last 30 years or more, two new theories have lately been discussed. They are the telomere shortening theory [38] and the stem cell theory [39].

According to the telomere shortening theory, cell senescence and death is due to telomere loss. Telomeres are the “caps” found on the end of chromosomes in somatic cells. They consist of long repeated lengths of what appears to be usually inactive DNA. Telomeres do not encode genetic information. They serve to preserve the integrity of the information encoded in chromosomes during the process of cell division. With each cell division, the telomeres become a bit shorter in the daughter cells. After a point (the Hayflick limit) the telomeres are too short for reliable chromosome duplication. Telomere shortening is observed in mitotic (dividing) human cells during aging. Also, almost all chronic diseases increase the rate of cell turnover and therefore telomere shortening. This theory seems to have been validated as a research group at Harvard Medical School succeeded in reversing ageing in mice by reactivating telomerase, the enzyme that synthesizes telomeres [40].

The latest theory is the stem cell theory, whose main proponents are Norman E. Sharpless, Ronald A. DePinho, Huber Warner, Alessandro Testori and others. This theory says that the aging process is due to the inability of various types of stem cells to continue to replenish the tissues of an organism with functionally differentiated cells capable of maintaining the original function of that tissue (or organ).

Model organisms like fruit flies, nematodes and mammals like mice have been used to study aging as the aging pathways in these organisms are similar to those in humans [41]. Pioneering work has been done on model organisms by scientists like Cynthia Kenyon [42], Leonard Guarente [43, 44], Coleen Murphy [45, 46], Matt Kaeberlein [44, 47]. It is now known that mutations in certain genes significantly alter the life span of fruit flies [48-52] and nematodes [49, 53-55]. These discoveries have led to the aging process being viewed as malleable by the same methods that are used to understand and manipulate development and disease [4]. At present, hundreds of mutant genes are known that can increase longevity in model organisms, including nematodes, yeast (*Saccharomyces cerevisiae*), fruitflies (*Drosophila melanogaster*) and mice (*Mus musculus*). Most act in evolutionarily conserved pathways that regulate growth, energy metabolism, nutrition sensing and/or reproduction [56]. Prominent examples include genes encoding components of the insulin/insulin-like growth factor 1 (IGF-I) signaling (IIS) pathway [57], the target of rapamycin (TOR) pathway [58], and the mitochondrial electron transport chain [59]. Work on model organisms has also shown how mechanisms like caloric restriction affect the aging process [46, 47].

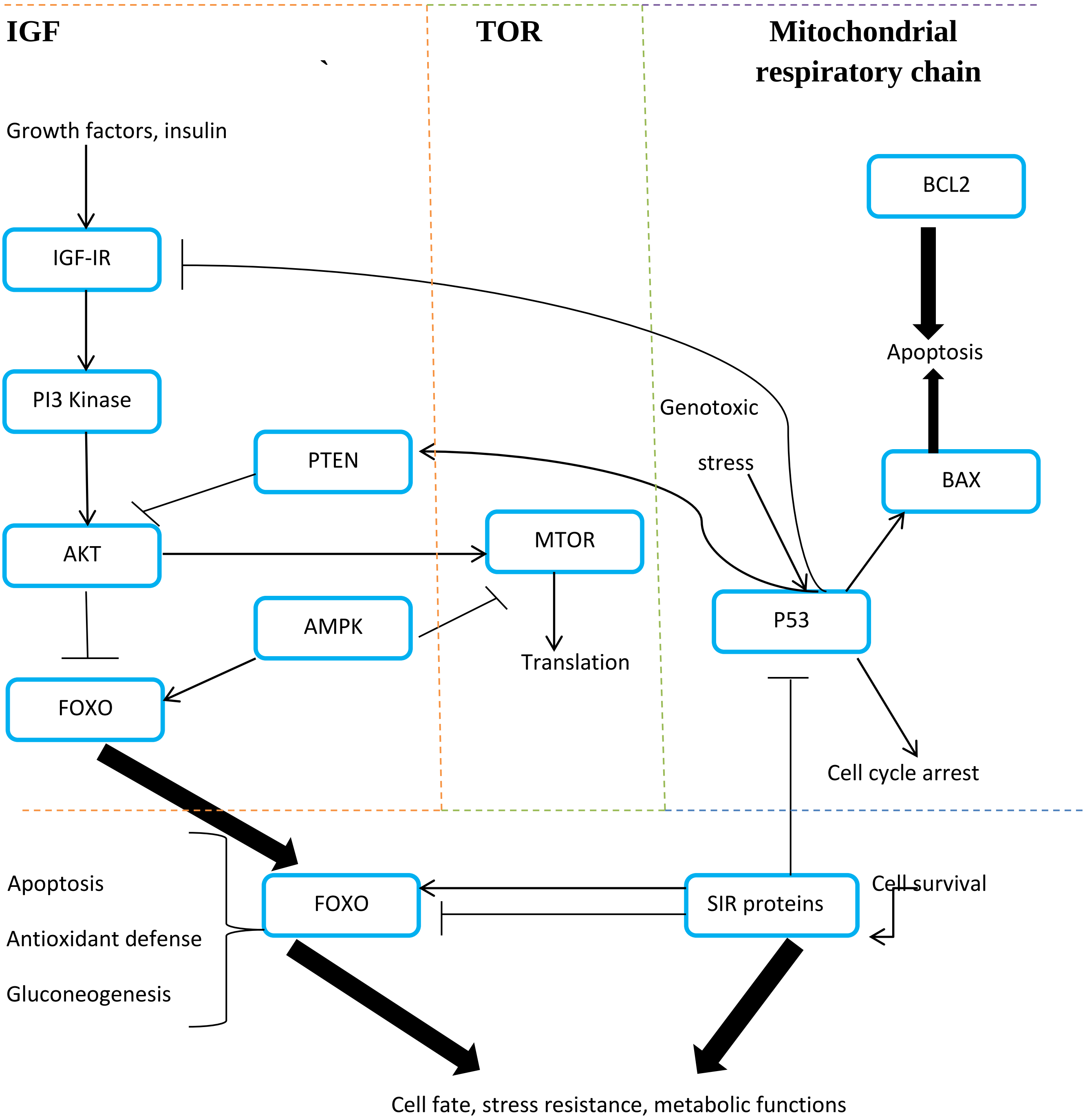


Figure 1. Potentially conserved pro-ageing pathways and their interconnections. Adapted from [16]

In humans, research on long-lived individuals may shed light on the mechanisms behind aging. It is seen that the longest lived people, called centenarians, are not usually diagnosed with any common age associated illness until very late in life, or they escape such diseases altogether [60]. Also, cardiovascular disease profiles are known to be better in centenarians, their siblings, and their offspring, than in controls [60-62]. Evert and colleagues [63] classified centenarians into three morbidity profiles: “Survivors” had a diagnosis of an age-associated illness prior to the age of 80, “Delayers” delayed the onset of age-associated illness until at least the age of 80, and “Escapers” attained their 100th year of life without the diagnosis of common age-associated illnesses. This suggests that the genetic study of longevity should focus on the most prevalent cause of morbidity and mortality, such as cardiovascular disease, in older age [64, 65]. Hence, for most biomedical scientists, the real concern is not ageing or immortality, but the diseases related to ageing, like Type-II Diabetes, obesity, arthritis, osteoporosis, cancer, Alzheimer’s, Parkinson’s etc., which now afflict a growing number of individuals.

A big impetus for research into the genetics of age-related disorders came from the Human Genome Project. The first draft was published in 2001 [66, 67], and revealed some surprising facts about the human genome. Only 1.1% of the genome was found to consist of protein-coding regions or exons, whereas 24% is in introns or parts of genes that are not protein-coding, and 75% of the genome is intergenic DNA. About 2.1 million SNP locations were found scattered throughout the genome, but less than 1% of these SNPs caused variations in proteins. The International HapMap project was set up to determine the common patterns of DNA sequence variation in the human genome and to make this information freely available in the public domain [68]. The HapMap was designed to facilitate identification of commonly occurring disease-causing variants based upon the “common disease, common variant” hypothesis [69]. This hypothesis suggests that at least some of the genetic influences on many common diseases are attributable to a limited number of common allelic variants that are present in more than 5% of the population. The latest data-set of this project, called HapMap 3, genotyped 1.6 million common single nucleotide polymorphisms (SNPs) in 1,184 reference individuals from 11 global populations [70]. Researchers now have a set of research tools that make it possible to find the genetic contributions to common diseases. The tools include computerized databases that contain the reference human genome sequence, a map of human genetic variation and a set of new technologies that can quickly and accurately analyze whole-genome samples for genetic variations that contribute to the onset of a disease.

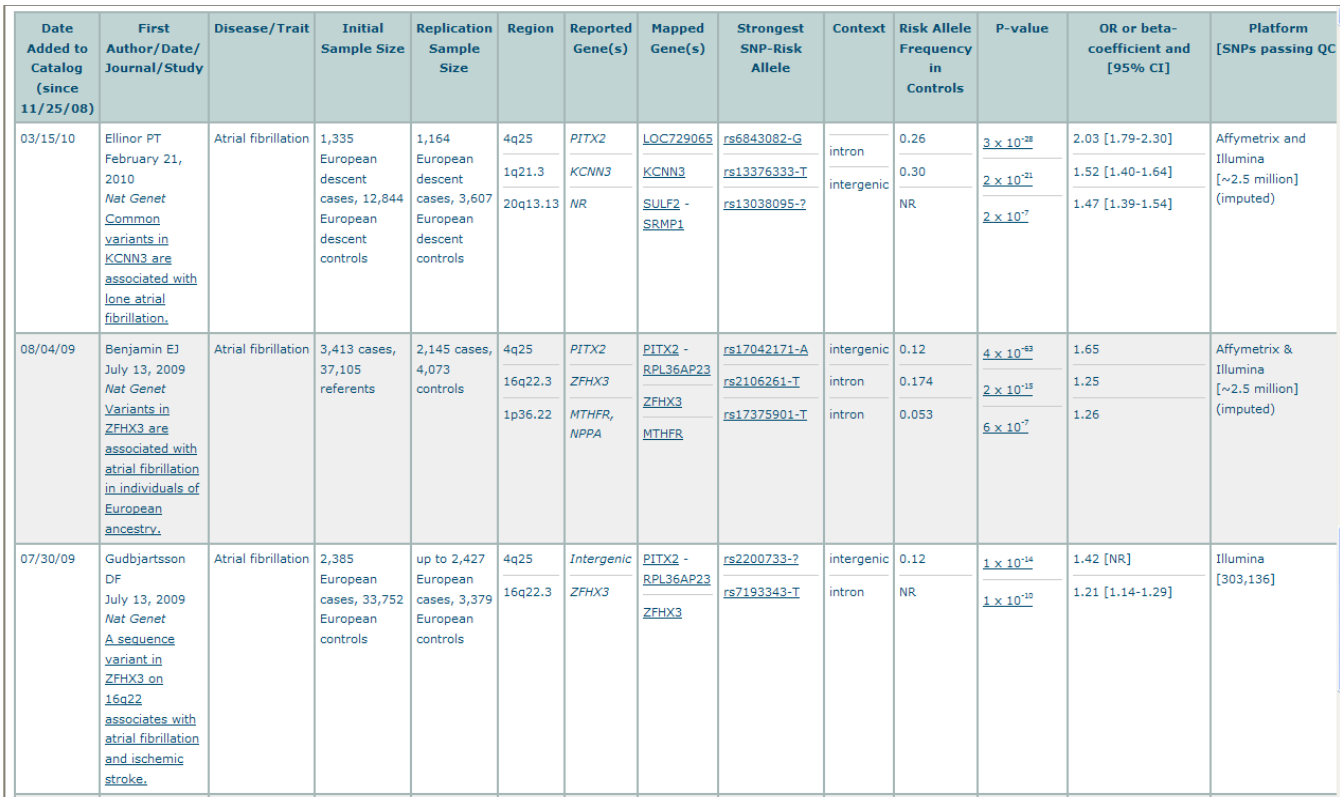
These tools have led to a large number of genome-wide association studies to be conducted for various diseases. GWA studies have been defined by the National Institutes of Health (NIH) as studies of common genetic variation across the entire human genome designed to identify genetic associations with observable traits [71]. GWA studies attempt to identify these common disease-causing variants by using high-throughput genotyping technologies to assay hundreds of thousands of common SNPs throughout the genome and relate them to clinical conditions and measurable traits. The first notable GWAS was published in 2005 [72] which implicated the gene for complement factor H (CFH) in age-related macular degeneration (AMD).

Many GWA studies have been conducted on age-related disorders. Examples include Alzheimer’s disease [73-76], schizophrenia [77, 78], Crohn’s disease [79-81], prostate cancer [82-84], breast cancer [85-87], AMD [72, 88-89], coronary artery disease [90, 91] and others. Larger scale studies include the deCODE database [92] or the National Heart, Lung, and Blood Institute’s Framingham Study [91]. Now, collaborating groups of researchers are combining data from multiple studies to identify associations that no single study could identify on its own. The Diabetes Genetics Replication and Meta-analysis (DIAGRAM) Consortium [93] is a progress in this regard. It combines 10 different GWA studies on type 2 diabetes for an effective population size of >50,000. Collaboration has even taken place across diseases, as in the case of the Wellcome Trust Case-Control Consortium (WTCCC) [94], which conducted a landmark study of 2000 cases of each of seven common diseases and 3000 shared controls [95]. This study provided many fundamental methodologic advances, including demonstration of the robustness of a single control group, the value of using cases of some diseases as controls for others, the greater power provided by increased sample size (numbers of subjects) rather than increased genomic coverage (numbers of SNPs), the critical need for manual review of automated genotyping calls, and the reliability of imputed genotypes for SNPs that were not actually typed by the genotyping platform [96].

GWA studies have been successful in predicting unusual associations. Many of the associations identified to date, such as CFH in macular degeneration [72] and TCF7L2 in type 2 diabetes [97, 98] have been surprising—the genes were not previously suspected of being related to the disease. Some, such as the strong associations of prostate cancer with SNPs in the 8q24 region [82] and Crohn's disease with the 5p13 region [95], have been in genomic regions containing no known genes at all. Variants or regions implicated in multiple diseases, such as the 8q24 region in prostate, breast, and colorectal cancers and the PTPN2 gene in type-1 diabetes and Crohn's disease [99] have also aroused a lot of interest. However, all findings of these studies have not been confirmed. For example, results of GWAS conducted in 2006 on Parkinson's disease and obesity [100, 101] could not be replicated [102, 103].

The data being generated from these GWAS studies needs to be stored in a database for easy access to researchers worldwide. Some of these databases are given below:

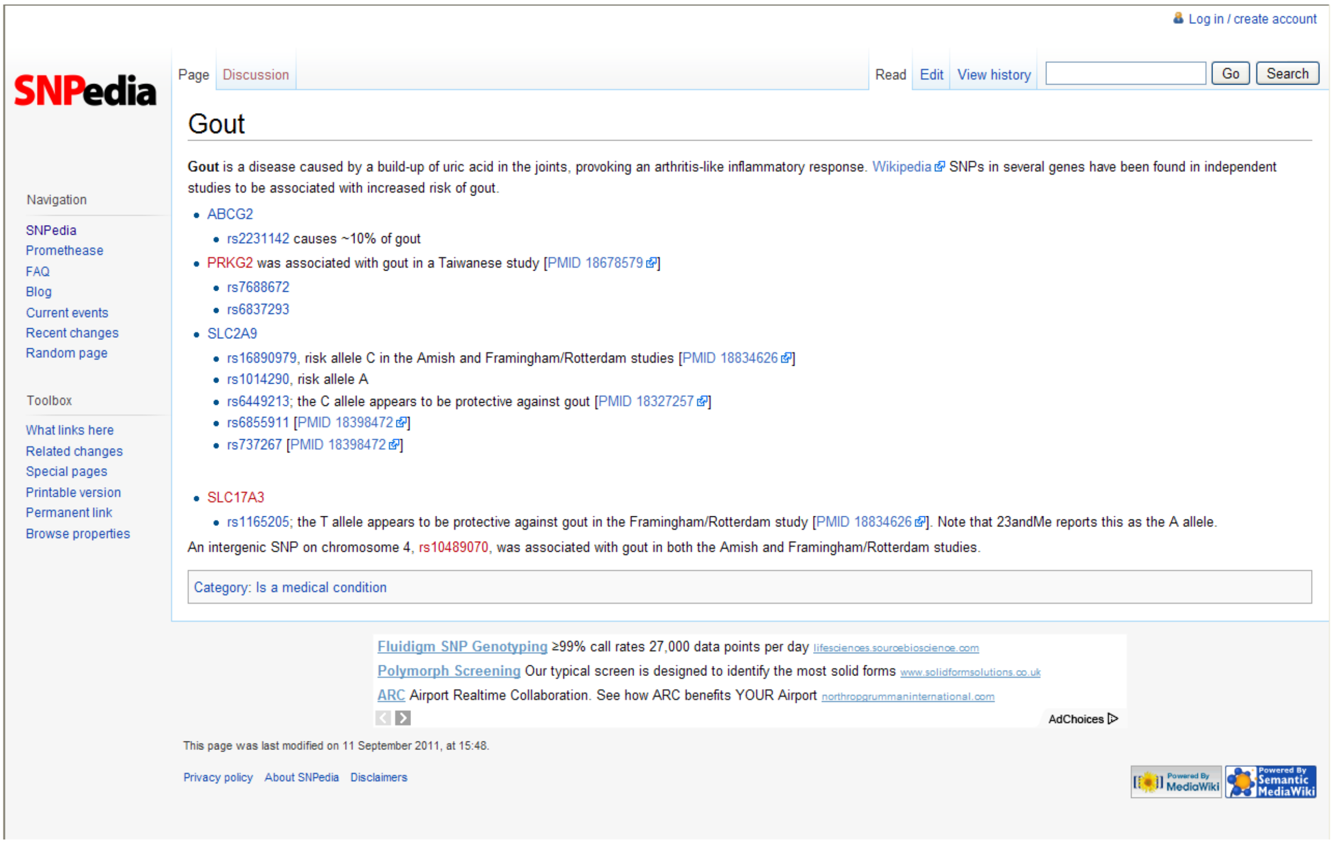
1. **NHGRI GWAS catalog**: The National Human Genome Research Institute (NHGRI) (www.genome.gov/) maintains a catalog of all GWAS conducted that attempted to assay at least 100,000 SNPs in the initial stage [104]. This catalog is downloadable and can also be searched via its online interface at (http://www.genome.gov/gwastudies) using various criteria like disease, gene, SNP, journal, author, p-value or combinations of these. The results are displayed in the format shown in Figure 2.



Date Added to Catalog (since 11/25/08)	First Author/Date/ Journal/Study	Disease/Trait	Initial Sample Size	Replication Sample Size	Region	Reported Gene(s)	Mapped Gene(s)	Strongest SNP-Risk Allele	Context	Risk Allele Frequency in Controls	P-value	OR or beta-coefficient and [95% CI]	Platform [SNPs passing QC
03/15/10	Ellinor PT February 21, 2010 *Nat Genet* Common variants in KCNN3 are associated with lone atrial fibrillation.	Atrial fibrillation	1,335 European descent cases, 12,844 European descent controls	1,164 European descent cases, 3,607 European descent controls	4q25 1q21.3 20q13.13	*PITX2* *KCNN3* *NR*	LOC729065 KCNN3 SULF2 - SRMP1	rs6843082-G rs13376333-T rs13038095-?	intron intergenic	0.26 0.30 NR	3×10^{-28} 2×10^{-21} 2×10^{-7}	2.03 [1.79-2.30] 1.52 [1.40-1.64] 1.47 [1.39-1.54]	Affymetrix and Illumina [~2.5 million] (imputed)
08/04/09	Benjamin EJ July 13, 2009 *Nat Genet* Variants in ZFHX3 are associated with atrial fibrillation in individuals of European ancestry.	Atrial fibrillation	3,413 cases, 37,105 referents	2,145 cases, 4,073 controls	4q25 16q22.3 1p36.22	*PITX2* *ZFHX3* *MTHFR, NPPA*	PITX2 - RPL36AP23 ZFHX3 MTHFR	rs17042171-A rs2106261-T rs17375901-T	intergenic intron intron	0.12 0.174 0.053	4×10^{-63} 2×10^{-15} 6×10^{-7}	1.65 1.25 1.26	Affymetrix & Illumina [~2.5 million] (imputed)
07/30/09	Gudbjartsson DF July 13, 2009 *Nat Genet* A sequence variant in ZFHX3 on 16q22 associates with atrial fibrillation and ischemic stroke.	Atrial fibrillation	2,385 European cases, 33,752 European controls	up to 2,427 European cases, 3,379 European controls	4q25 16q22.3	*Intergenic* *ZFHX3*	PITX2 - RPL36AP23 ZFHX3	rs2200733-? rs7193343-T	intergenic intron	0.12 NR	1×10^{-14} 1×10^{-10}	1.42 [NR] 1.21 [1.14-1.29]	Illumina [303,136]

Figure 2. NHGRI GWAS catalog results for Atrial fibrillation

2. **SNPedia:** SNPedia [105] is a wiki that provides information about the effects of variations in DNA, citing peer-reviewed scientific publications. It can be searched by disease, gene or SNP. It is available on the web at www.snpedia.org/. SNPedia results for gout are displayed in Figure 3.



Gout

Gout is a disease caused by a build-up of uric acid in the joints, provoking an arthritis-like inflammatory response. Wikipedia SNPs in several genes have been found in independent studies to be associated with increased risk of gout.

- ABCG2
  - rs2231142 causes ~10% of gout
- PRKG2 was associated with gout in a Taiwanese study [PMID 18678579]
  - rs7688672
  - rs6837293
- SLC2A9
  - rs16890979, risk allele C in the Amish and Framingham/Rotterdam studies [PMID 18834626]
  - rs1014290, risk allele A
  - rs6449213; the C allele appears to be protective against gout [PMID 18327257]
  - rs6855911 [PMID 18398472]
  - rs737267 [PMID 18398472]
- SLC17A3
  - rs1165205; the T allele appears to be protective against gout in the Framingham/Rotterdam study [PMID 18834626]. Note that 23andMe reports this as the A allele.

An intergenic SNP on chromosome 4, rs10489070, was associated with gout in both the Amish and Framingham/Rotterdam studies.

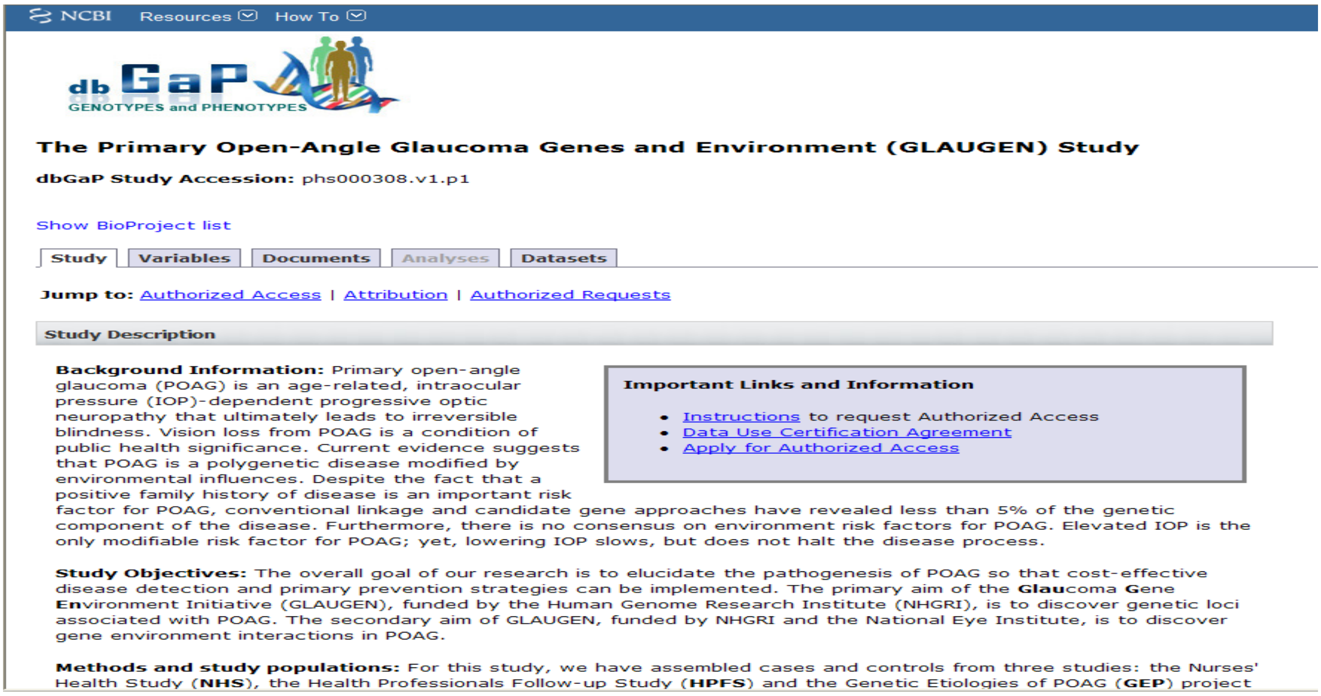
Category: Is a medical condition

This page was last modified on 11 September 2011, at 15:48.

Privacy policy About SNPedia Disclaimers

Figure 3. SNPedia results for Gout

3. **dbGap**: The database of Genotypes and Phenotypes (dbGaP) (http://www.ncbi.nlm.nih.gov/gap) is maintained by the National Centre of Biotechnology Information (NCBI) and maintains a list of GWAS conducted till date. It contains detailed information about the study like the sample sizes, study method etc. It also gives information about the variables and datasets used.



The Primary Open-Angle Glaucoma Genes and Environment (GLAUGEN) Study

dbGaP Study Accession: phs000308.v1.p1

Show BioProject list

Jump to: Authorized Access | Attribution | Authorized Requests

Study Description

Important Links and Information

- Instructions to request Authorized Access
- Data Use Certification Agreement
- Apply for Authorized Access

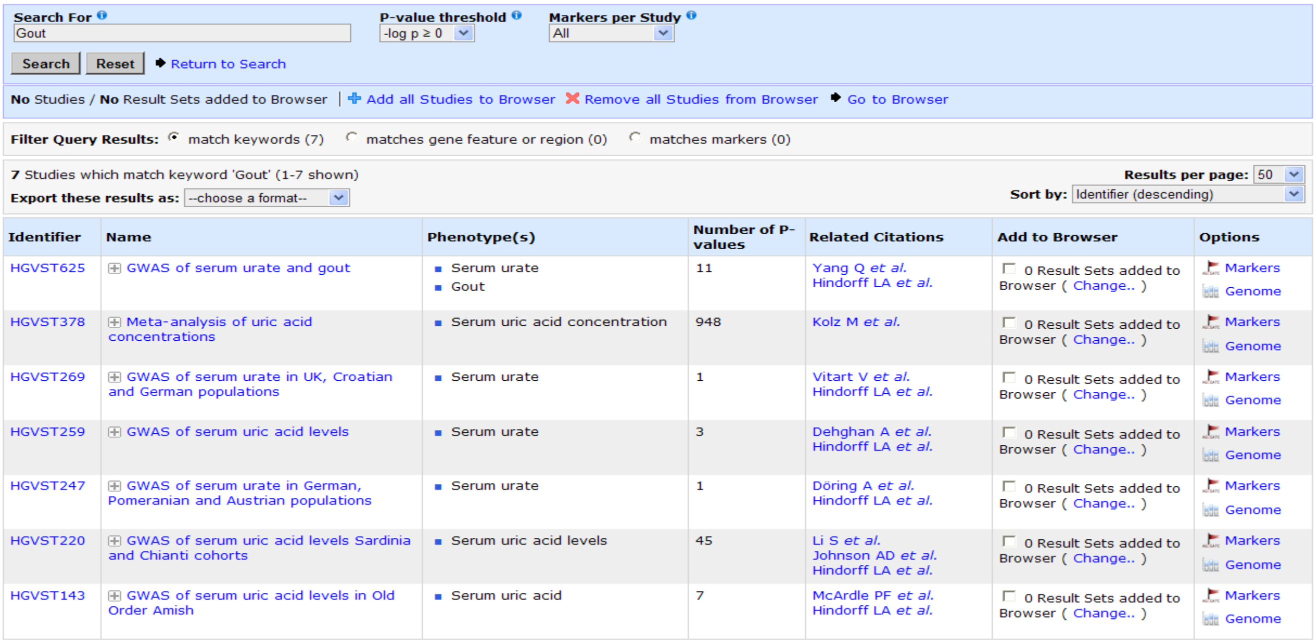
Background Information: Primary open-angle glaucoma (POAG) is an age-related, intraocular pressure (IOP)-dependent progressive optic neuropathy that ultimately leads to irreversible blindness. Vision loss from POAG is a condition of public health significance. Current evidence suggests that POAG is a polygenetic disease modified by environmental influences. Despite the fact that a positive family history of disease is an important risk factor for POAG, conventional linkage and candidate gene approaches have revealed less than 5% of the genetic component of the disease. Furthermore, there is no consensus on environment risk factors for POAG. Elevated IOP is the only modifiable risk factor for POAG; yet, lowering IOP slows, but does not halt the disease process.

Study Objectives: The overall goal of our research is to elucidate the pathogenesis of POAG so that cost-effective disease detection and primary prevention strategies can be implemented. The primary aim of the **Glau**coma **Gen**e **En**vironment Initiative (GLAUGEN), funded by the Human Genome Research Institute (NHGRI), is to discover genetic loci associated with POAG. The secondary aim of GLAUGEN, funded by NHGRI and the National Eye Institute, is to discover gene environment interactions in POAG.

Methods and study populations: For this study, we have assembled cases and controls from three studies: the Nurses' Health Study (**NHS**), the Health Professionals Follow-up Study (**HPFS**) and the Genetic Etiologies of POAG (**GEP**) project

Figure 4. dbGaP result for a glaucoma GWAS

4. **GWAS Central**: It was earlier known as HGVbaseG2P [106]. It can now be accessed at (http://www.gwascentral.org/index). It provides a list of all GWAS conducted on a particular disease along-with the markers that were found to be significantly associated.



Search For Gout
P-value threshold -log p ≥ 0
Markers per Study All

No Studies / **No** Result Sets added to Browser | Add all Studies to Browser ✗ Remove all Studies from Browser Go to Browser

Filter Query Results: ● match keywords (7) ○ matches gene feature or region (0) ○ matches markers (0)

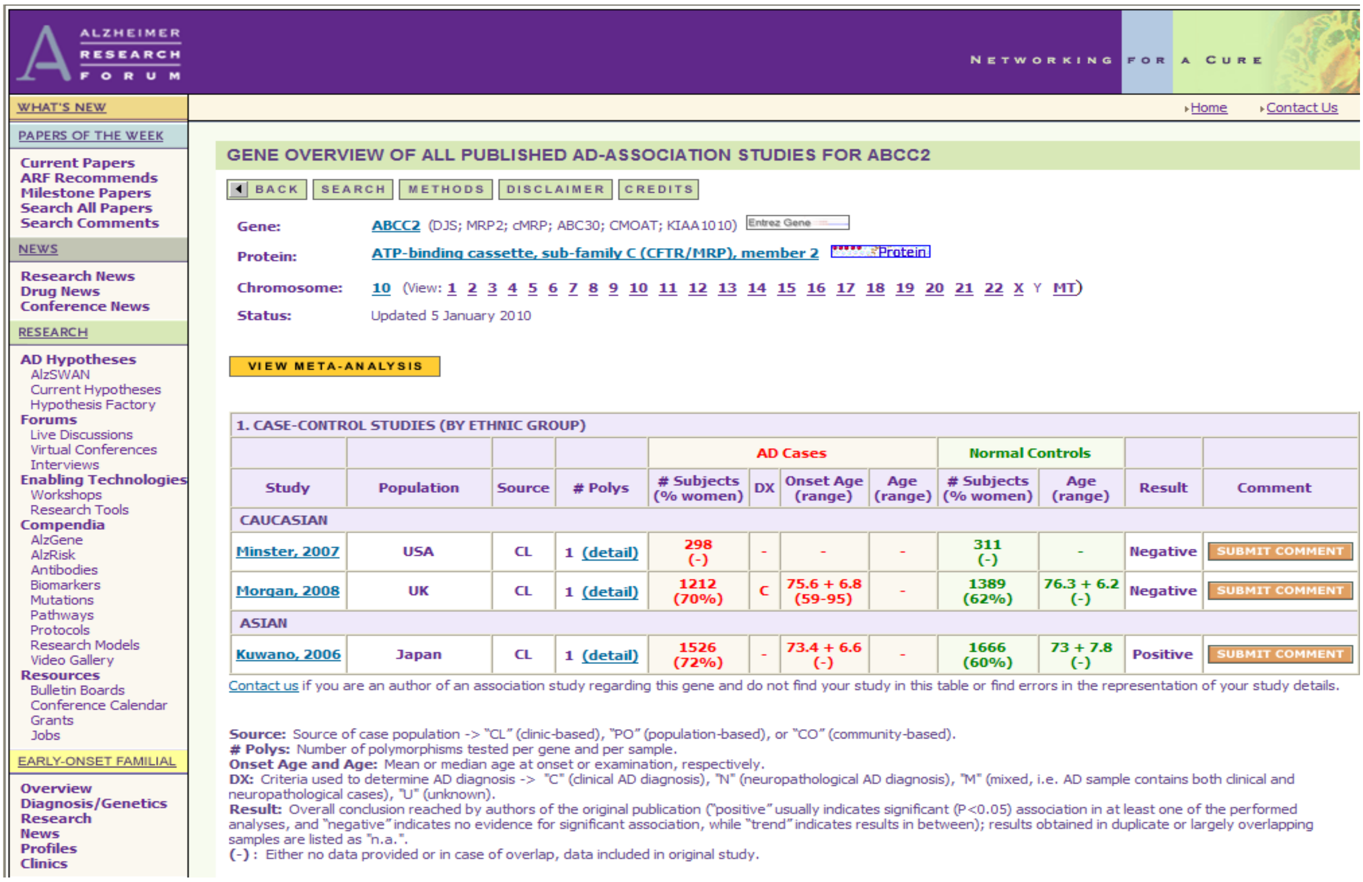
7 Studies which match keyword 'Gout' (1-7 shown)
Results per page: 50
Export these results as: --choose a format--
Sort by: Identifier (descending)

Identifier	Name	Phenotype(s)	Number of P-values	Related Citations	Add to Browser	Options
HGVST625	⊞ GWAS of serum urate and gout	Serum urate Gout	11	Yang Q *et al.* Hindorff LA *et al.*	□ 0 Result Sets added to Browser (Change..)	Markers Genome
HGVST378	⊞ Meta-analysis of uric acid concentrations	Serum uric acid concentration	948	Kolz M *et al.*	□ 0 Result Sets added to Browser (Change..)	Markers Genome
HGVST269	⊞ GWAS of serum urate in UK, Croatian and German populations	Serum urate	1	Vitart V *et al.* Hindorff LA *et al.*	□ 0 Result Sets added to Browser (Change..)	Markers Genome
HGVST259	⊞ GWAS of serum uric acid levels	Serum urate	3	Dehghan A *et al.* Hindorff LA *et al.*	□ 0 Result Sets added to Browser (Change..)	Markers Genome
HGVST247	⊞ GWAS of serum urate in German, Pomeranian and Austrian populations	Serum urate	1	Döring A *et al.* Hindorff LA *et al.*	□ 0 Result Sets added to Browser (Change..)	Markers Genome
HGVST220	⊞ GWAS of serum uric acid levels Sardinia and Chianti cohorts	Serum uric acid levels	45	Li S *et al.* Johnson AD *et al.* Hindorff LA *et al.*	□ 0 Result Sets added to Browser (Change..)	Markers Genome
HGVST143	⊞ GWAS of serum uric acid levels in Old Order Amish	Serum uric acid	7	McArdle PF *et al.* Hindorff LA *et al.*	□ 0 Result Sets added to Browser (Change..)	Markers Genome

Figure 5. GWAS Central results for gout

Each of these databases is linked to NCBI databases like dbSNP (www.ncbi.nlm.nih.gov/projects/SNP/) and PubMed (http://www.ncbi.nlm.nih.gov/pubmed/). Hence, for any additional information required by the user, he/she has to visit the links provided.

Unfortunately, there is no such database dedicated to all age-related disorders, though specialized databases focused on a single disease exist. The AlzGene database for Alzheimer's [107] and the SzGene database for Schizophrenia [108] are collections of GWAS results on these two disorders. They can be accessed at (http://www.alzgene.org/) and (http://www.szgene.org/). They can be searched by gene, protein, polymorphism or study. A sample of the results obtained by using AlzGene is shown in Figure 6.



GENE OVERVIEW OF ALL PUBLISHED AD-ASSOCIATION STUDIES FOR ABCC2

Gene: ABCC2 (DJS; MRP2; cMRP; ABC30; CMOAT; KIAA1010)

Protein: ATP-binding cassette, sub-family C (CFTR/MRP), member 2

Chromosome: 10 (View: 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X Y MT)

Status: Updated 5 January 2010

1. CASE-CONTROL STUDIES (BY ETHNIC GROUP)

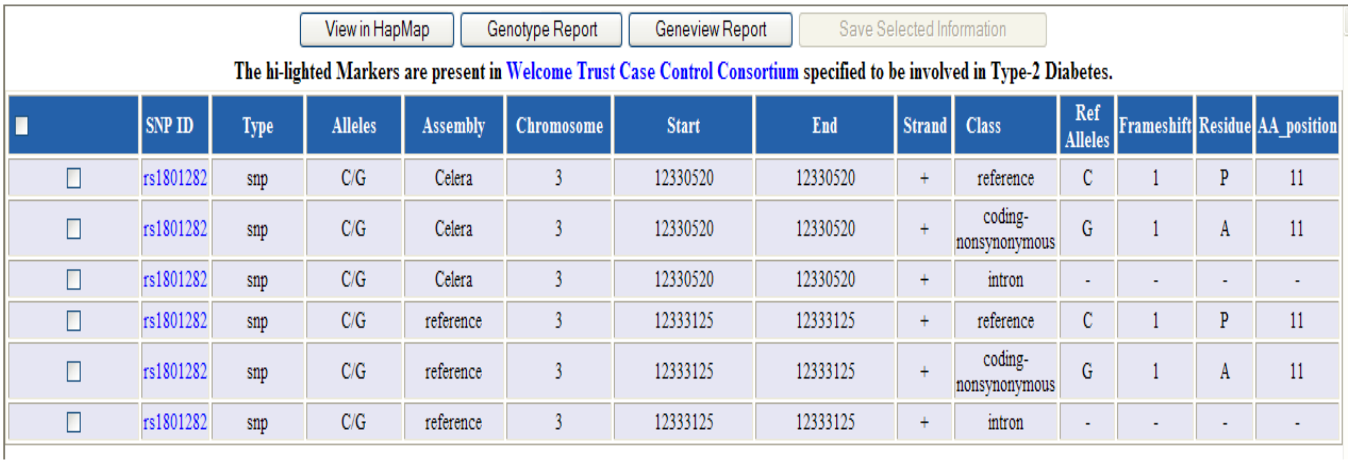
				AD Cases				Normal Controls			
Study	Population	Source	# Polys	# Subjects (% women)	DX	Onset Age (range)	Age (range)	# Subjects (% women)	Age (range)	Result	Comment
CAUCASIAN											
Minster, 2007	USA	CL	1 (detail)	298 (-)	-	-	-	311 (-)	-	Negative	SUBMIT COMMENT
Morgan, 2008	UK	CL	1 (detail)	1212 (70%)	C	75.6 + 6.8 (59-95)	-	1389 (62%)	76.3 + 6.2 (-)	Negative	SUBMIT COMMENT
ASIAN											
Kuwano, 2006	Japan	CL	1 (detail)	1526 (72%)	-	73.4 + 6.6 (-)	-	1666 (60%)	73 + 7.8 (-)	Positive	SUBMIT COMMENT

Contact us if you are an author of an association study regarding this gene and do not find your study in this table or find errors in the representation of your study details.

Source: Source of case population -> "CL" (clinic-based), "PO" (population-based), or "CO" (community-based).
# Polys: Number of polymorphisms tested per gene and per sample.
Onset Age and Age: Mean or median age at onset or examination, respectively.
DX: Criteria used to determine AD diagnosis -> "C" (clinical AD diagnosis), "N" (neuropathological AD diagnosis), "M" (mixed, i.e. AD sample contains both clinical and neuropathological cases), "U" (unknown).
Result: Overall conclusion reached by authors of the original publication ("positive" usually indicates significant ($P<0.05$) association in at least one of the performed analyses, and "negative" indicates no evidence for significant association, while "trend" indicates results in between); results obtained in duplicate or largely overlapping samples are listed as "n.a.".
(-) : Either no data provided or in case of overlap, data included in original study.

Figure 6. AlzGene results for ABCC2 gene

Also, T2D-Db [109] (http://t2ddb.ibab.ac.in/home.shtml) is a database of all molecular factors reported to be involved in the pathogenesis of Type 2 diabetes in human, mouse and rat. It provides information on candidate genes, gene description, genomic loci, aliases, gene and protein sequences and the corresponding literature, SNP markers and transcript information. It also caters information on genes candidates for the risk factors/complications reported to be associated with Type 2 diabetes.



The hi-lighted Markers are present in Welcome Trust Case Control Consortium specified to be involved in Type-2 Diabetes.

□	SNP ID	Type	Alleles	Assembly	Chromosome	Start	End	Strand	Class	Ref Alleles	Frameshift	Residue	AA_position
□	rs1801282	snp	C/G	Celera	3	12330520	12330520	+	reference	C	1	P	11
□	rs1801282	snp	C/G	Celera	3	12330520	12330520	+	coding-nonsynonymous	G	1	A	11
□	rs1801282	snp	C/G	Celera	3	12330520	12330520	+	intron	-	-	-	-
□	rs1801282	snp	C/G	reference	3	12333125	12333125	+	reference	C	1	P	11
□	rs1801282	snp	C/G	reference	3	12333125	12333125	+	coding-nonsynonymous	G	1	A	11
□	rs1801282	snp	C/G	reference	3	12333125	12333125	+	intron	-	-	-	-

Figure 7. T2D-Db results for SNP rs1801282

These databases are popular as they are linked to NCBI databases like dbSNP and Pubmed, and also have a user-friendly interface. It may be of even more help to users if the important information found in the NCBI databases is contained within the database of age-related disorders. There should also be a better interface so that the user is able to find the required information without any trouble. Hence there is a pressing need for a comprehensive database of ARDs that stores the information about disease-SNP links in one place. This work is a step in that direction.

OBJECTIVES

1. To compile and classify all age-related disorders.
2. To curate all SNPs associated with the selected age-related disorders from the available literature of genome-wide association studies (GWAS).
3. To create a relational data model to store information about the SNPs and their disease associations
4. To use MySQL to create a database of disease-SNP associations based on this data model.
5. To create a web-based graphical user interface for querying the database.
6. To connect the database to the interface using Perl CGI scripting.

METHODOLOGY

Firstly, the age-related disorders to be analyzed were selected. Then the SNPs associated with them in various genome-wide association studies (GWAS) were mined from the available literature and already existing databases. The information provided in the GWAS like the population under study, its geographical location and the p-value associated with the prediction were all included in preparing a disease-SNP list. Also, information related to the SNP like its location, type and nucleotide/codon change was added from dbSNP. MySQL (www.mysql.com) was used to convert this disease-SNP list into a normalized relational database and the information was loaded into this database. A free interactive HTML webpage designing software called Nvu was used to design the interface for the database. Finally, the database and interface were connected using Perl CGI. Each of these steps is described in more detail below.

Selection of diseases:

Complex genetic disorders that are believed to be prevalent in old age, like Alzheimer's and Parkinson's disease, have been considered age-related disorders for the purpose of this work. The list of these disorders was made as comprehensive as possible. The disease names were mostly kept consistent with OMIM nomenclature. Susceptibility to a disease and various subtypes of the disease were merged into the disease name. For example, Autosomal dominant and Late-onset Parkinson's disease have been included under the term Parkinson's disease, and susceptibility to osteoarthritis is included under osteoarthritis. The list of diseases is given in Appendix IV.

Further the diseases were classified using the classification scheme used by Goh et al [110] for constructing the diseasome. This classification was based on the body part or system affected by the disorder. For example, atrial fibrillation is classified as a cardiovascular disease while osteoporosis is a bone disorder. The list of disease classes along with the diseases present in each class is given in Appendix V.

Mining literature for disease-SNP associations:

Single nucleotide polymorphisms (SNPs) related to each disease were mined from literature. GWA studies conducted on the disease were used for predicting association between the disease and an SNP. Only SNPs that were found significantly associated ($p<0.05$) in a GWAS were included. The list of SNPs was curated from online resources like the NHGRI GWAS catalog, SNPedia, OMIM, dbGaP and PubMed. The rsIDs of the SNPs were noted. The PMID reference of the disease-SNP link was also noted.

Creation of disease-SNP list:

The PMIDs noted earlier were used to obtain information about the GWAS, like the population under study and the geographical location in which the study was conducted. The p-value associated with the disease-SNP link was also included in the list. Also, information related to the SNP, like its type (missense, synonymous, non-coding), location (exon, intron, 5'-UTR etc.), gene, nucleotide/amino acid change, and others was compiled from dbSNP and included in the list. The details about information under each heading are given below:

Class	Assigned on the basis of classification of diseases as mentioned in the section "Selection of diseases".
Disease	The disease name. It has been kept consistent with OMIM nomenclature as far as possible. Also see the section "Selection of diseases".
rsID	The rsID of the SNP as given in dbSNP build 134.
PMID	The PubMed reference number of the article discussing the GWAS.
Population	The population which was studied by the GWAS. If the study was a meta-analysis, this information is provided under Population. Also if the study used patients from a large consortium like WTCCC, the name of the consortium is mentioned instead of the population.
Geographical location	The location (country name) in which the GWAS was conducted. This field may be empty if the study was a meta-analysis or if it was conducted by an international consortium. The location may be given as "Various" if a large number of countries were involved.
p-value	The p-value assigned to the disease-SNP association by the GWAS, derived from the largest sample size, typically a combined analysis (initial plus replication studies). If this value is not reported, the p-value from the initial study sample is recorded. If the GWAS did not assign any p-value, the information is given as NR (Not Reported).
Odds-ratio (OR)	The odds-ratio of the disease-SNP association as given by the GWAS, derived from the largest sample size, typically a combined analysis (initial plus replication studies). If this value is not reported, the OR from the initial study sample is recorded. If the study reported ORs for the different disease models (dominant, heterozygous, recessive), then the highest odds-ratio is mentioned. If the GWAS did not assign any OR, the information is given as NR (Not Reported).
Gene	If the SNP is found in the exon, intron or untranslated region (UTR) of any gene as given by dbSNP, the name of that gene is given. If the SNP is in an intergenic region, the names of the two closest genes are given, that is, the two closest genes in between which the SNP is found. This information is taken from the GeneView feature of dbSNP. For example, the gene for SNP rs2076756 is NOD2. However rs2542151 lies between PSMG2 - PTPN2.
Chromosome	The chromosome on which the SNP is found. For example, the SNP rs2076756 belongs to chromosome 16. This information is taken from dbSNP.

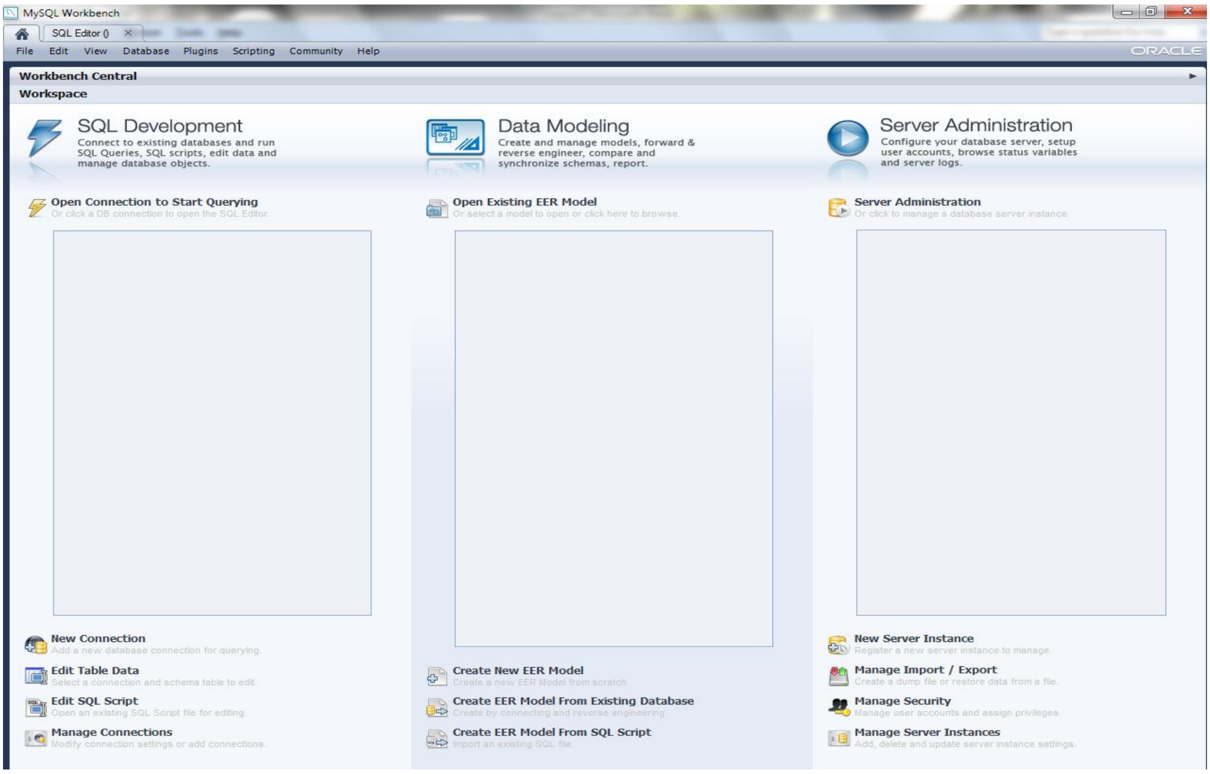
Ref_contig	The accession number of the reference contig in which the SNP is found. The NT_ accession numbers are reported. For example, the Ref_contig of rs12035082 is given as NT_004487.19. The mapping is done according to the NCBI scheme GRCh37.
Ref_mRNA	The accession number of the reference transcript in which the SNP is found. If the SNP is not found in any transcripts, this is left empty. For example, the Ref_mRNA of rs2076756 is given as NM_022162.1. However, no accession number is given for rs2542151 as it is an intergenic SNP and hence is not a part of any transcript.
Ref_protein	The accession number of the protein sequence in which the SNP is found. If the SNP is not found in any transcripts, this is left empty. For example, the Ref_protein of rs11209026 is given as NP_653302.2. However there is no such number provided for rs2076756 as it is an intronic SNP and hence not a part of the final protein product.
Type	The type of SNP. If the SNP is not found in a protein, the type is given as “Non coding”. For example, the SNP rs2076756 which is an intronic SNP is given as Non coding. The different types of SNPs are: Missense, Non coding and cds-synon.
Nucleotide/Codon Change	If the SNP is in a coding region, the codon change is given. Else, the nucleotide change taking place in the polymorphism is given. If more than one change may take place under a single rsID, each of these is included. For example, in the case of rs2076756 which is an intronic SNP, the nucleotide change is given, which is A>G. however, in the case of rs2241880 which is a missense SNP, the codon change is mentioned, that is, ACT>GCT.
Amino Acid change	If the SNP causes a change in a protein, the amino acids being changed are denoted by their single letter codes. For example, in the case of rs2241880 the amino acid change is given as T>A. In the case of synonymous SNPs, there is no change in amino acid, hence it is represented as such, for example, A>A.
Location	The location of the SNP. This information is taken from the function class in dbSNP. If the SNP is Missense, then its location is given as “Exon”. If it does not lie in a gene or a UTR, its location is reported to be “Intergenic”. The different locations of SNPs are: Exon, Intron, 3-UTR, 5-UTR and Intergenic.
Position	The position of the mutated amino acid in the protein. If the SNP is not in a protein, this field is left blank. For example, the position of rs1799983 is 298.
Remarks	Any additional information to be provided about the disease-SNP link. For example, one study [66] found the association of rs370409 to Grave’s disease in Han Chinese to be significant only among women. This fact is mentioned in Remarks.
Annotator	The name of the annotator of the record in this list.

Table I: Information included under each heading in disease-SNP list

All this information was compiled in an Excel worksheet as raw data. Care was taken to ensure that each row was unique, that is, there was duplication of data.

Creation of data model for disease-SNP list:

The data model was created using the free open-source version of the MySQL Workbench. MySQL Workbench is a visual database design application that can be used to efficiently design, manage and document database schemata.



Connect to existing databases and run SQL Queries, SQL scripts, edit data and manage database objects.

Create and manage models, forward & reverse engineer, compare and synchronize schemas, report.

Configure your database server, setup user accounts, browse status variables and server logs.

Or click a DB connection to open the SQL Editor.

Or select a model to open or click here to browse.

Or click to manage a database server instance.

Add a new database connection for querying.

Select a connection and schema table to edit.

Open an existing SQL Script file for editing.

Modify connection settings or add connections.

Create a new EER Model from scratch.

Create by connecting and reverse engineering.

Import an existing SQL file.

Register a new server instance to manage.

Create a dump file or restore data from a file.

Manage user accounts and assign privileges.

Add, delete and update server instance settings.

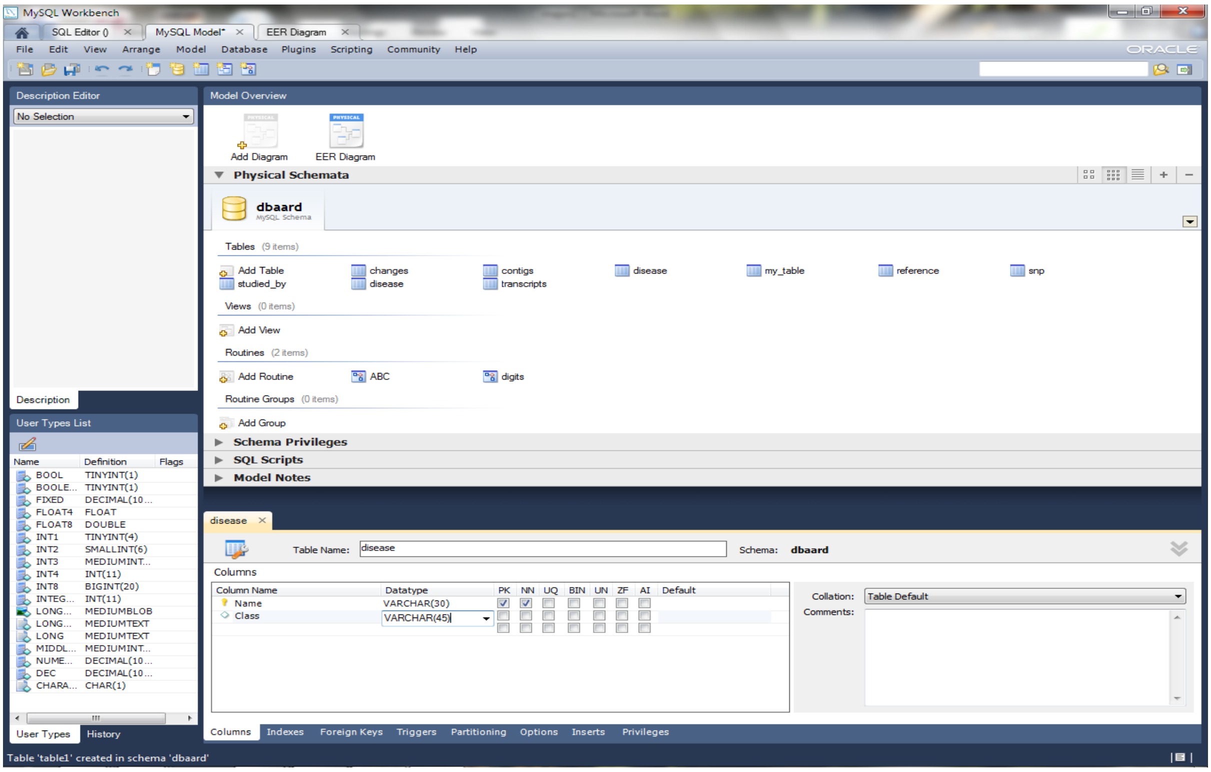
Figure 8. MySQL Workbench

A data model for the database called dbAARD was designed using MySQL Workbench version 5.2 (http://www.mysql.com/products/workbench/) to store the information in the disease-SNP list in the form of a relational database. To create a new data model, the following steps were performed:

1. Under the heading "Data Modeling", click on "Create new data model".
2. Double-click on the "mydb" tab to enter the name of the database.
3. Double-click on "Add table" to add a new table to the database. It is possible to enter the table name, column names and types as well as declare primary key, foreign keys and other constraints and triggers.

The fields were divided into tables in such a way as to minimize data redundancy. Each table was assigned one or more primary keys. Foreign keys were used to reference other tables in the database. To create a new table, the following steps were performed:

1. Double-click on “Add table” under the heading “Tables”.
2. Type the table name in the text box.
3. Double-click under the heading Column Name and type the name of the column that is the primary key. Ensure that the PK (Primary Key) and NN (Not Null) checkboxes are selected.
4. Select the data-type of the column.
5. Create the other fields in the same way.
6. To declare a column as a foreign key, click on the “Foreign Keys” tab at the bottom. Type the foreign key name and the referenced table and column. Also select Cascade option in the case of both Update and Delete. This will ensure that the changed value of the referenced column is reflected accurately in this table as well.



Name	Definition	Flags
BOOL	TINYINT(1)	
BOOLE...	TINYINT(1)	
FIXED	DECIMAL(10...	
FLOAT4	FLOAT	
FLOAT8	DOUBLE	
INT1	TINYINT(4)	
INT2	SMALLINT(6)	
INT3	MEDIUMINT...	
INT4	INT(11)	
INT8	BIGINT(20)	
INTEG...	INT(11)	
LONG...	MEDIUMBLOB	
LONG...	MEDIUMTEXT	
LONG	MEDIUMTEXT	
MIDDL...	MEDIUMINT...	
NUME...	DECIMAL(10...	
DEC	DECIMAL(10...	
CHARA...	CHAR(1)	

Table Name: disease Schema: dbaard

Columns

Column Name	Datatype	PK	NN	UQ	BIN	UN	ZF	AI	Default
Name	VARCHAR(30)	☑	☑	□	□	□	□	□	
Class	VARCHAR(45)	□	□	□	□	□	□	□	
		□	□	□	□	□	□	□	

Collation: Table Default
Comments:

Table 'table1' created in schema 'dbaard'

Figure 9. Creating a new table

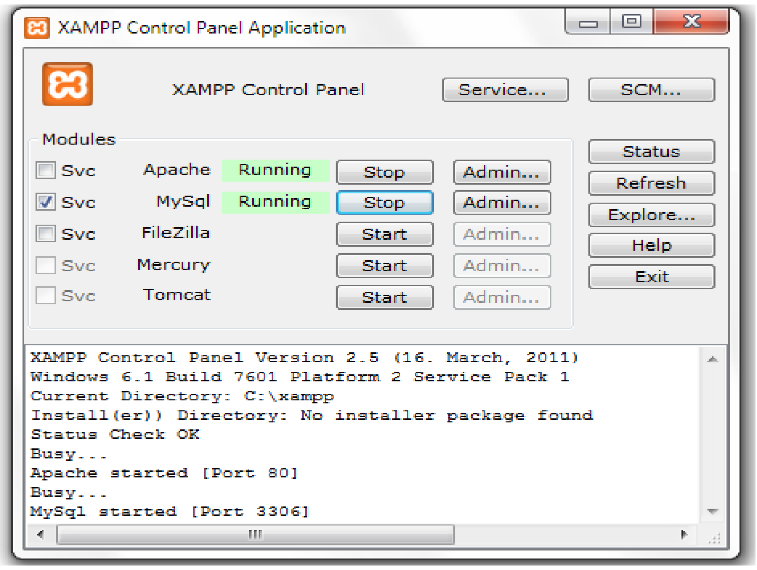
For example, the table SNP contains all the information about the SNP like rsID, gene, chromosome, type and location. Here, the field rsID is the primary key, which is also referenced by other tables. However, the other fields are not found anywhere else in the database, so there is no duplication of data and redundancy is minimized. The data model was designed in such a way that the database would be in the third normalized form (3NF).

An Entity-Relationship (ER) diagram was also created for this data model that showed the relations between the different tables, as well as the primary key, foreign key and Not Null columns in all the tables. This was done by selecting the Export option in the File menu and then clicking on “Export as single page PDF”.

The SQL Create script was generated automatically from this data model using the workbench. This was done by selecting the Export option in the File menu and then clicking on "Forward engineer SQL Create script".

Creation of database and loading of data:

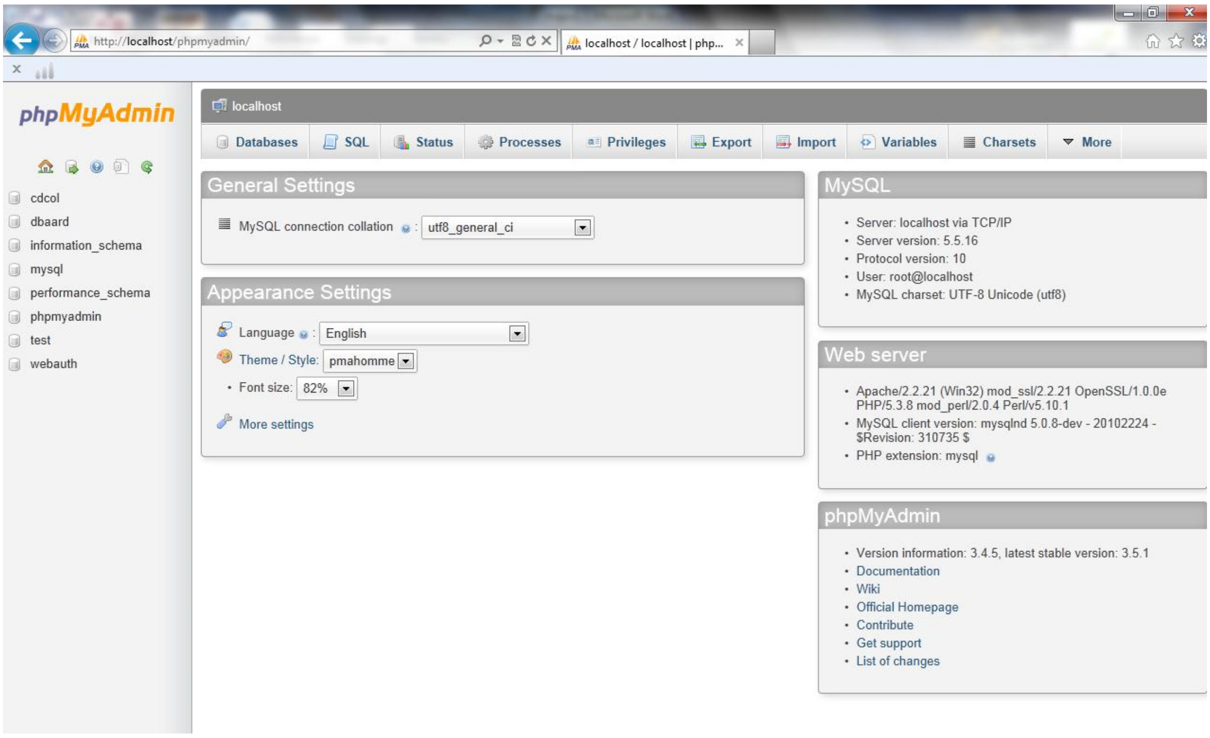
XAMPP version 1.7.7 (http://www.apachefriends.org/en/xampp.html) was downloaded from the website and installed. XAMPP is a free Apache distribution containing MySQL (version 5.5.16), PHP and Perl 5.10.


XAMPP Control Panel Application
XAMPP Control Panel
Service...
SCM...
Modules
□ Svc Apache Running Stop Admin...
☑ Svc MySql Running Stop Admin...
□ Svc FileZilla Start Admin...
□ Svc Mercury Start Admin...
□ Svc Tomcat Start Admin...
Status
Refresh
Explore...
Help
Exit

```
XAMPP Control Panel Version 2.5 (16. March, 2011)
Windows 6.1 Build 7601 Platform 2 Service Pack 1
Current Directory: C:\xampp
Install(er)) Directory: No installer package found
Status Check OK
Busy...
Apache started [Port 80]
Busy...
MySql started [Port 3306]
```

Figure 10. XAMPP Control Panel with the Apache web server and MySQL running

It also contains a module called phpMyAdmin (version 3.4.5) that makes creation and maintenance of MySQL databases simple. This can be accessed from http://localhost/phpmyadmin.



General Settings

- MySQL connection collation : utf8_general_ci

Appearance Settings

- Language : English
- Theme / Style: pmahomme
- Font size: 82%
- More settings

MySQL

- Server: localhost via TCP/IP
- Server version: 5.5.16
- Protocol version: 10
- User: root@localhost
- MySQL charset: UTF-8 Unicode (utf8)

Web server

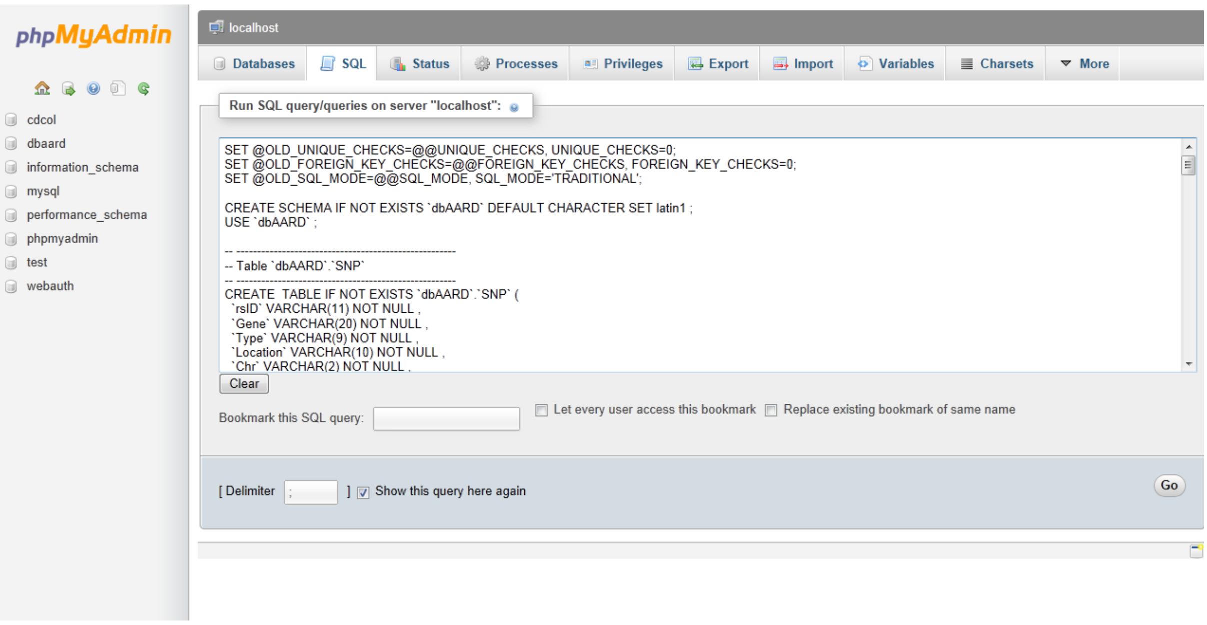
- Apache/2.2.21 (Win32) mod_ssl/2.2.21 OpenSSL/1.0.0e PHP/5.3.8 mod_perl/2.0.4 Perl/v5.10.1
- MySQL client version: mysqlnd 5.0.8-dev - 20102224 - $Revision: 310735 $
- PHP extension: mysql

phpMyAdmin

- Version information: 3.4.5, latest stable version: 3.5.1
- Documentation
- Wiki
- Official Homepage
- Contribute
- Get support
- List of changes

Figure 11. phpMyAdmin homepage

Using the SQL Create script generated earlier in the SQL tab of phpMyAdmin, the database was created.



Run SQL query/queries on server "localhost":

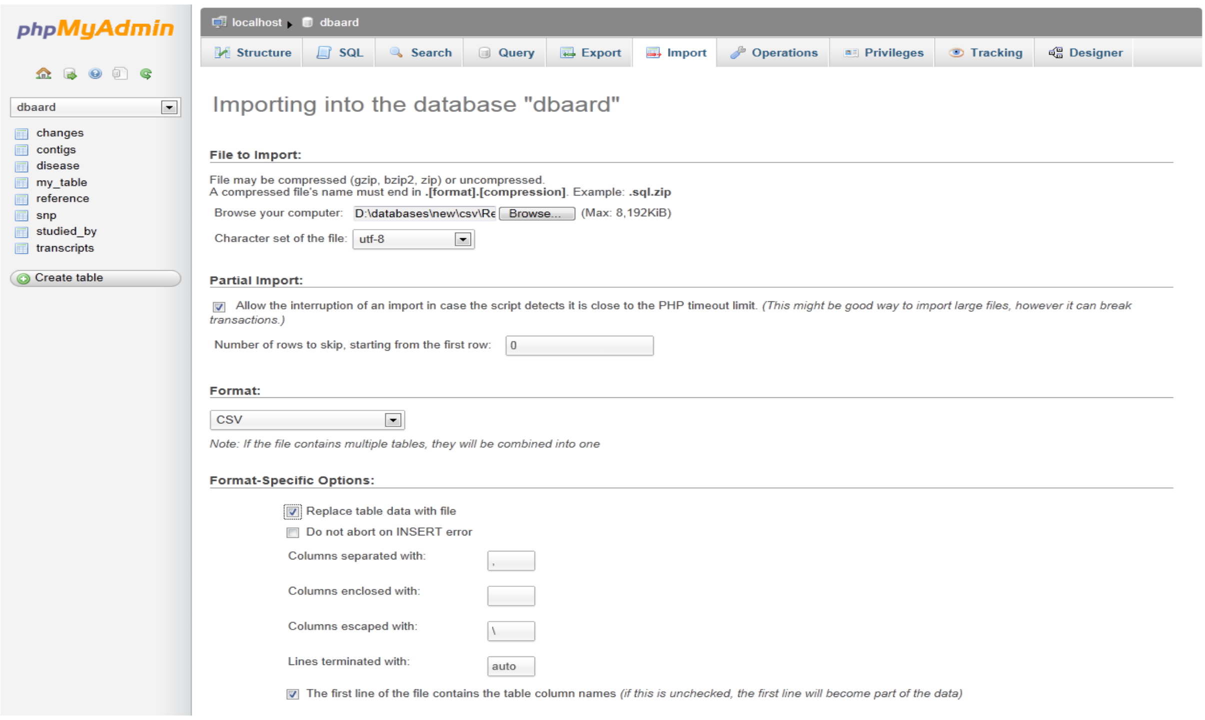
```
SET @OLD_UNIQUE_CHECKS=@@UNIQUE_CHECKS, UNIQUE_CHECKS=0;
SET @OLD_FOREIGN_KEY_CHECKS=@@FOREIGN_KEY_CHECKS, FOREIGN_KEY_CHECKS=0;
SET @OLD_SQL_MODE=@@SQL_MODE, SQL_MODE='TRADITIONAL';

CREATE SCHEMA IF NOT EXISTS `dbAARD` DEFAULT CHARACTER SET latin1 ;
USE `dbAARD` ;

-- -----------------------------------------------------
-- Table `dbAARD`.`SNP`
-- -----------------------------------------------------
CREATE  TABLE IF NOT EXISTS `dbAARD`.`SNP` (
  `rsID` VARCHAR(11) NOT NULL ,
  `Gene` VARCHAR(20) NOT NULL ,
  `Type` VARCHAR(9) NOT NULL ,
  `Location` VARCHAR(10) NOT NULL ,
  `Chr` VARCHAR(2) NOT NULL ,
```

Bookmark this SQL query: ☐ Let every user access this bookmark ☐ Replace existing bookmark of same name

[Delimiter ;] ☑ Show this query here again

Figure 12. The phpMyAdmin SQL tab with the dbAARD Create script

The raw data was divided into different Comma Separated Variable (CSV) files such that one CSV file corresponded to one table in the data model. The data was loaded into the tables using the Import tab in phpMyAdmin.



Importing into the database "dbaard"

File to Import:

File may be compressed (gzip, bzip2, zip) or uncompressed.
A compressed file's name must end in **.[format].[compression]**. Example: **.sql.zip**

Browse your computer: D:\databases\new\csv\Re Browse... (Max: 8,192KiB)

Character set of the file: utf-8

Partial Import:

☑ Allow the interruption of an import in case the script detects it is close to the PHP timeout limit. *(This might be good way to import large files, however it can break transactions.)*

Number of rows to skip, starting from the first row: 0

Format:

CSV

Note: If the file contains multiple tables, they will be combined into one

Format-Specific Options:

☑ Replace table data with file
☐ Do not abort on INSERT error

Columns separated with: ,
Columns enclosed with:
Columns escaped with: \
Lines terminated with: auto

☑ The first line of the file contains the table column names *(if this is unchecked, the first line will become part of the data)*

Figure 13. Importing a CSV file into the database

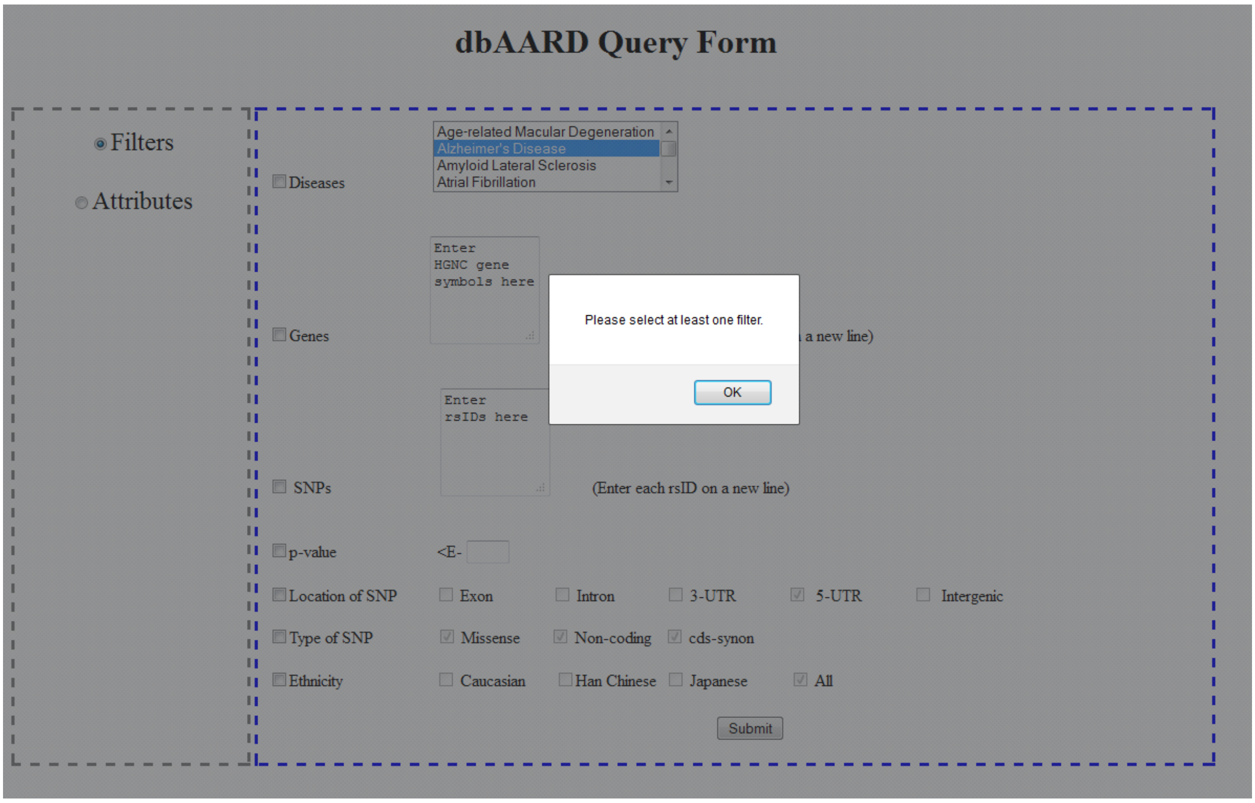
Designing the web interface:

A free interactive webpage designing software Nvu (http://net2.com/) was downloaded from the website and installed. Nvu is an HTML editor that is based on the Mozilla Application Suite 1.7. Therefore pages created in Nvu look exactly the same in Mozilla Firefox browser. Nvu supports HTML 4.01 as well as CSS styling. Both these features are used in the creation of the interface.

The development of the GUI (Graphical User Interface) is mainly menu-driven, as the software does not require the user to have in-depth knowledge of HTML. For example, the creation of a checkbox is shown below:

1. From the Insert menu, select Form and from the sub-menu, select Form Field.
2. In the resulting dialog box, select Check box in the dropdown menu for Field Type.
3. Type the name of the checkbox
4. The checkbox is created.

Javascript was used to as the client side script. This code is executed by the user's computer and hence Javascript needs to be enabled on the computer for the GUI to work properly. For example, it is used to validate the form before submission. If there are errors in the form, the appropriate message alert is shown and the form is not submitted.



dbAARD Query Form

◉ Filters

○ Attributes

☐ Diseases

Age-related Macular Degeneration
Alzheimer's Disease
Amyotrophic Lateral Sclerosis
Atrial Fibrillation

☐ Genes

Enter HGNC gene symbols here

a new line)

Please select at least one filter.

OK

☐ SNPs

Enter rsIDs here

(Enter each rsID on a new line)

☐ p-value <E-

☐ Location of SNP ☐ Exon ☐ Intron ☐ 3-UTR ☑ 5-UTR ☐ Intergenic

☐ Type of SNP ☑ Missense ☑ Non-coding ☑ cds-synon

☐ Ethnicity ☐ Caucasian ☐ Han Chinese ☐ Japanese ☑ All

Submit

Figure 14. Error message shown due to invalid form submitted

Writing the Perl CGI script:

The Common Gateway Interface, or CGI, is a set of standards that define how information is exchanged between the web server and a custom script.

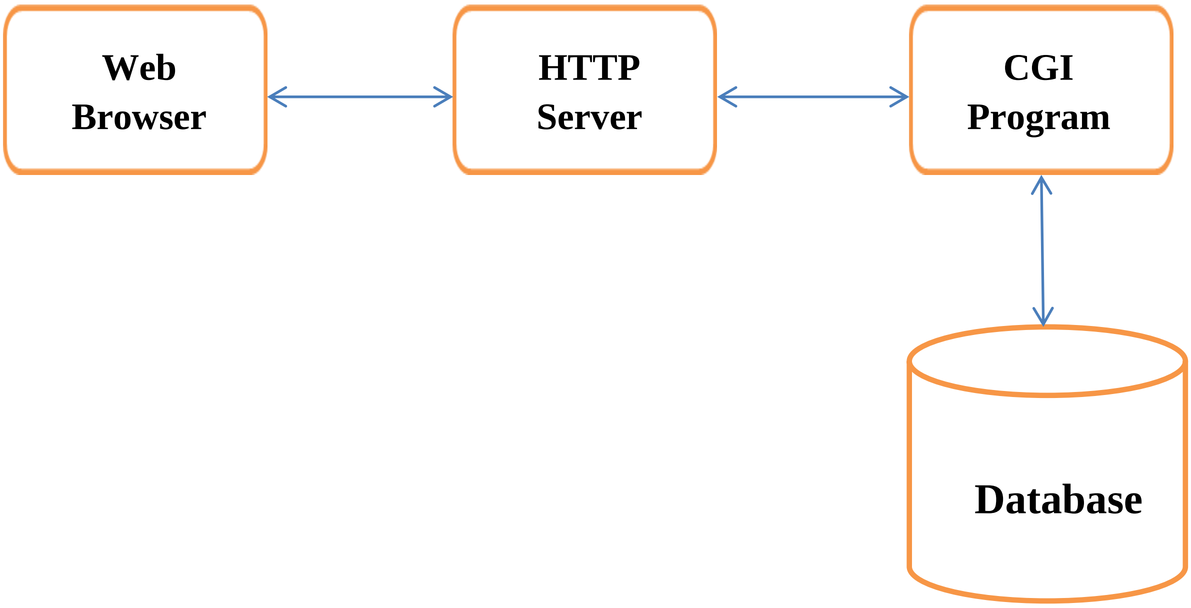


Figure 15. Information retrieval from database by web browser using CGI

Figure 14 shows the process of retrieving data from an online database through a web browser. The process is as follows:

1. The browser sends the query to the server. This may be done using the GET or POST methods.
2. The server executes the CGI program which connects to the database and gets the information required.
3. This information is passed to the server.
4. The server sends the information back to the browser where it is displayed to the user.

Perl 5.10 was used as the server-side scripting language. This code is executed by the server when the form is submitted by the user. The POST method is used to submit the form, so the form data is sent inside a message to the server, which makes it easier to extract and parse.

Three external modules were used to create the CGI script:

- The Perl module CGI.pm was used to create the script.
- The module DBI.pm was used to connect to the database.
- The module HTML::Template was used to create the final results table.

RESULTS AND CONCLUSIONS

Disease-SNP list:

The list of disease-SNP associations was curated from GWAS literature and databases as mentioned in the Methodology. The information in the disease-SNP list so created is summarized below:

Diseases	34
SNPs	1338
Disease classes	14
Unique PMIDs	502

Table II. Summary of information in the disease-SNP list

The GWAS- and SNP- related data was also included in the list as discussed in the Methodology. A sample of the list records is given in Appendix VI.

The information in the database may be grouped on the basis of the headings Population, Type and Location. It can be seen from Table IV that most of the records belong to the Caucasian population, Non coding type and Intron location.

Headings	**Major groups**	**Percentage of records**
Population	Caucasian	53.2
	Japanese	7.8
	Han- Chinese	7.4
SNP Type	Non-coding	89.1
	Missense	8.8
	cds-synon	2.1
SNP Location	Intergenic	36.4
	Intron	43.0
	Exon	11.2
	UTR-3	3.1
	UTR-5	4.9

Table III. Records grouped under different headings

The majority of GWAS have been conducted on Caucasian populations in the USA and Western Europe. Hence, the database predominantly contains variations that are present in these populations. A user analyzing this data should be careful as a disease-associated SNP in one population may not be present in individuals from a different ethnicity having the same disease. Unless GWAS are conducted on Asian and African populations as well this data will not be truly representative of all the SNPs causing a disease. Also, many GWAS are conducted on multiple ethnicities and some are also meta-analyses.

Moreover, most of the SNPs are found to be non-coding. These are the SNPs found in intergenic, intronic and the 3' and 5'-untranslated regions. It can be seen that exonic mutations, which may be missense or synonymous, are much fewer than the mutations in the non-coding parts of the genome.

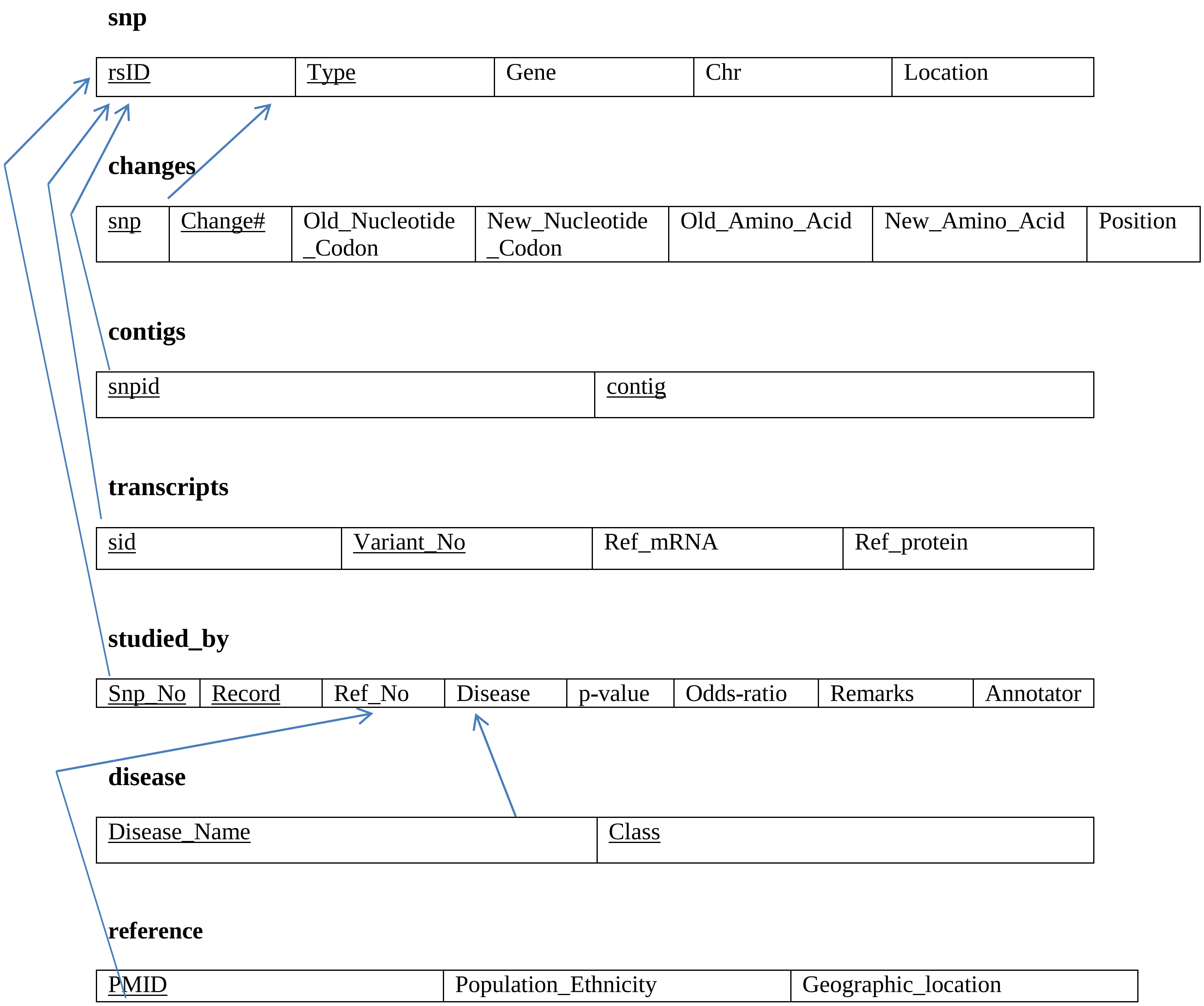
Data model:

A data model was created to store the information in the disease-SNP list. MySQL Workbench was used to create the data model as explained in the Methodology. It was normalized to conform to the following normalization standards:

- **1 NF** (First normal form): Relation should have no non-atomic attributes or nested relations. The values of the attributes in this data model cannot be broken down any further. Hence they are atomic and the model is in 1 NF.
- **2NF** (Second normal form): For relations where primary key contains multiple attributes, no non-key attribute should be functionally dependent on a part of the primary key. In this data model, the tables in which there are two attributes as primary key, the other attributes are not functionally dependent on any attribute of the primary key. Hence, the model is in 2NF.
- **3NF** (Third normal form): Relation should not have a non-key attribute functionally determined by another non-key attribute (or by a set of non-key attributes). There are no such attributes in the model which are dependent on other non-key attributes. Hence the model is in 3NF.

Normalization of the database ensures minimization of data redundancy and quick information retrieval. Hence the database operates efficiently in terms of both disk space occupied and time taken for retrieving a particular record.

The model is represented as follows (the underlined fields are the primary keys):



snp

rsID	Type	Gene	Chr	Location

changes

snp	Change#	Old_Nucleotide_Codon	New_Nucleotide_Codon	Old_Amino_Acid	New_Amino_Acid	Position

contigs

snpid	contig

transcripts

sid	Variant_No	Ref_mRNA	Ref_protein

studied_by

Snp_No	Record	Ref_No	Disease	p-value	Odds-ratio	Remarks	Annotator

disease

Disease_Name	Class

reference

PMID	Population_Ethnicity	Geographic_location

Figure 16. Tables in the data model. Arrows connect the foreign key to the primary key of parent table

MySQL Workbench was also used to generate the Entity-Relationship (ER) diagram. It is shown in the following figure:

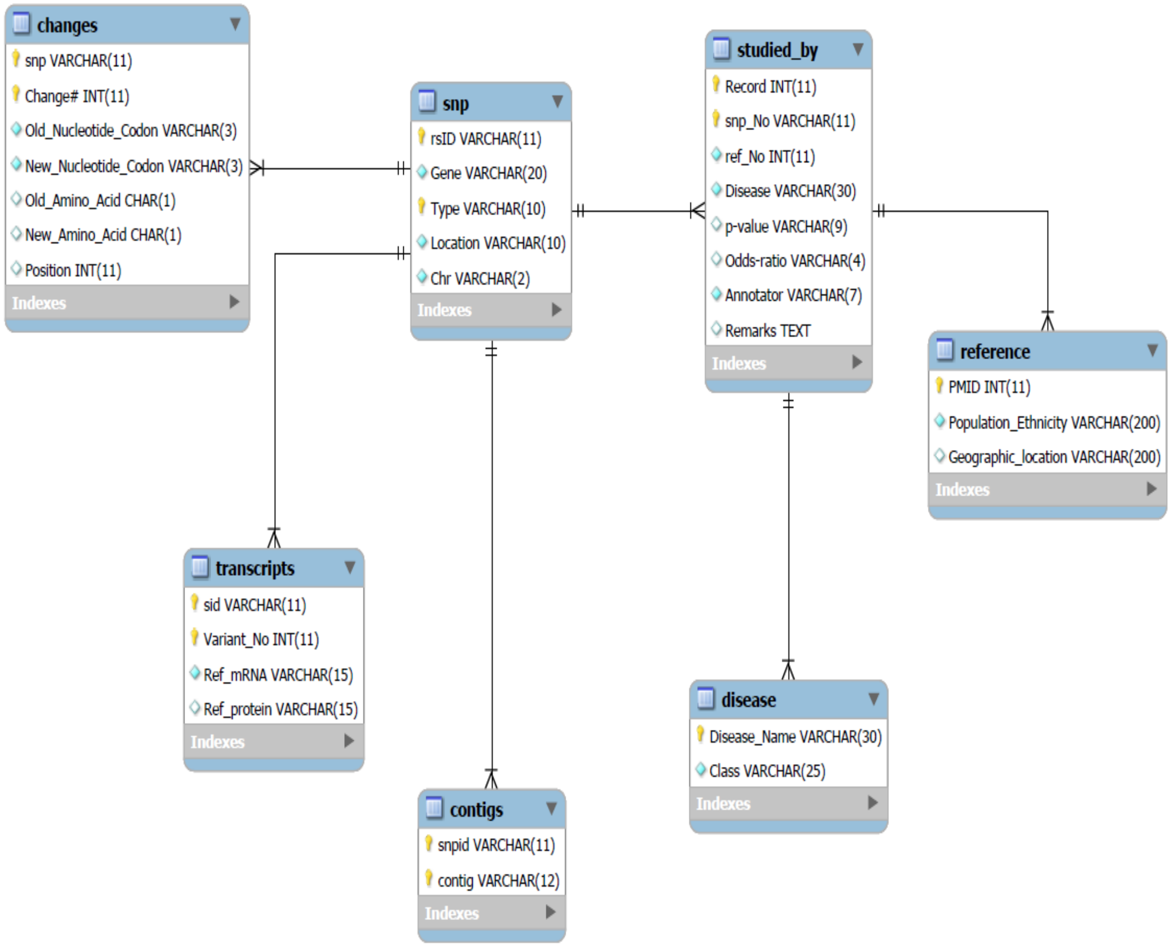


Figure 17. ER diagram of the data model in the crow-foot notation

🔑	Primary key
♦	Foreign key
◇	Not Null attribute
‖——————<	Identifying relationship (1:n)

Table IV. Key for ER diagram

Figure 17 shows the ER diagram in the crow-foot notation. In this diagram, the relations between the tables are represented by solid or dashed lines. Solid lines represent identifying relationships, that is, those relationships in which the foreign key in the child relation is also the primary key, for example, the snp-changes relationship. The relationships are 1: n, that is, each record in one relation (1 side of the relationship) references multiple records in the other relation (n side).

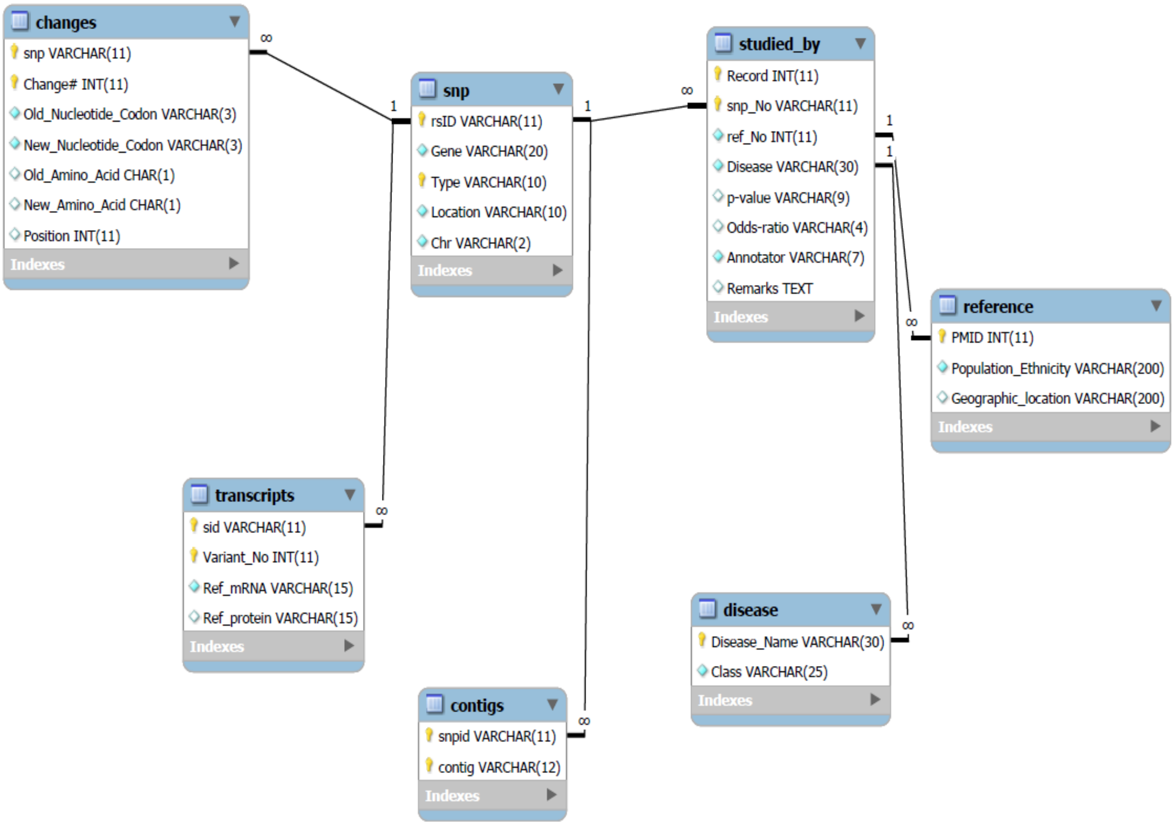


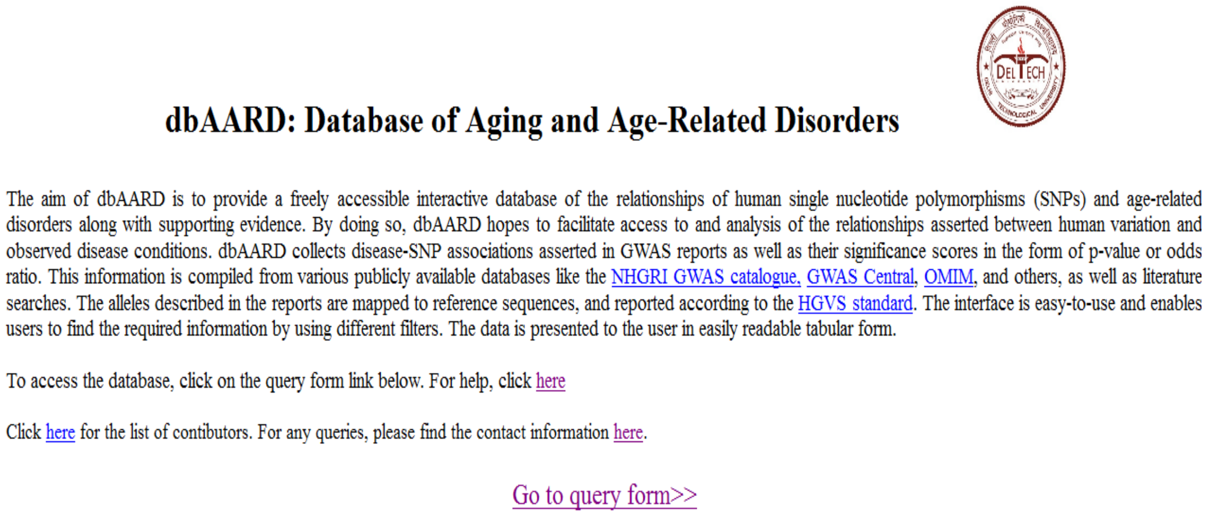
Figure 18. Another representation of the ER diagram

In figure 18, the ER diagram is shown in another notation. In this notation, the connections between the fields of the different relations are shown. For example, the attribute rsID of relation ‘snp’ is referred by the foreign key ‘snp’ of relation ‘changes’. In this notation the identifying and non-identifying relationships can be identified by solid ad dashed lines similar to the crow-foot notation. The 1: n relationships are written explicitly.

The CREATE script for the database was generated by the MySQL Workbench. It is given in Appendix X.

Interface:

A web-based graphical user interface has been designed using HTML. The software used for designing the interface is called Nvu. When the user visits the dbAARD website (http://dbaard.dce.edu) , the following homepage will be displayed in the browser:



dbAARD: Database of Aging and Age-Related Disorders

The aim of dbAARD is to provide a freely accessible interactive database of the relationships of human single nucleotide polymorphisms (SNPs) and age-related disorders along with supporting evidence. By doing so, dbAARD hopes to facilitate access to and analysis of the relationships asserted between human variation and observed disease conditions. dbAARD collects disease-SNP associations asserted in GWAS reports as well as their significance scores in the form of p-value or odds ratio. This information is compiled from various publicly available databases like the NHGRI GWAS catalogue, GWAS Central, OMIM, and others, as well as literature searches. The alleles described in the reports are mapped to reference sequences, and reported according to the HGVS standard. The interface is easy-to-use and enables users to find the required information by using different filters. The data is presented to the user in easily readable tabular form.

To access the database, click on the query form link below. For help, click here

Click here for the list of contibutors. For any queries, please find the contact information here.

Go to query form>>

Figure 19. Homepage of the database

The homepage provides the user with an overview of the database. It also links to the names of the contributors to the database. Links are also provided to the help page and contact details. At the end of the page, a link is given to proceed to the query form. Clicking on this link takes the user to the database interface.

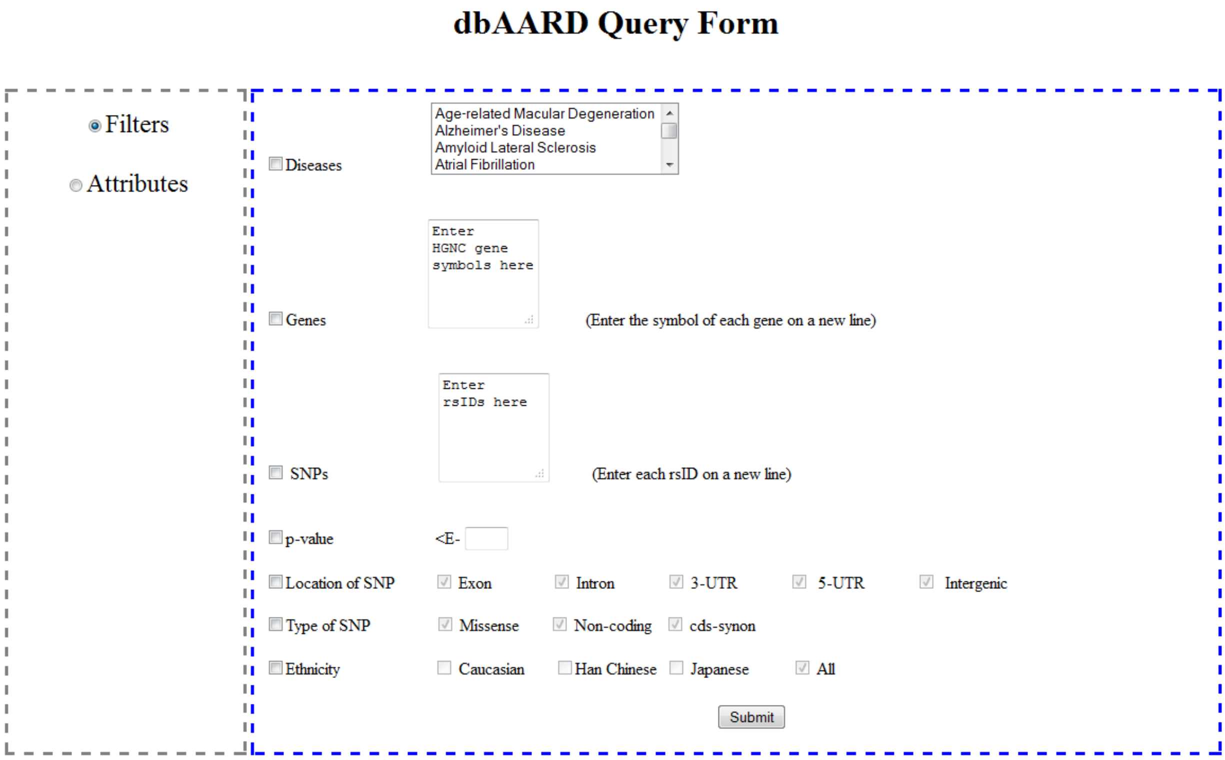
The query submission form consists of two parts:

Filters: These are the categories used to filter the records. They restrict the number of records that are shown in the result. The filters can be viewed by selecting the Filters radio button on the left. The filters are the following:

Diseases	The user may select the disease(s) for which he/she wants to know about the SNPs or GWAS. Multiple diseases may be selected from the list box by pressing the Ctrl key.
Genes	The user may filter records based on HGNC gene symbols. Multiple genes may be entered in the query box, one on each line.
SNPs	The user may filter records based on rsIDs as in dbSNP build 134. Multiple rsIDs may be entered in the query box, one on each line.
p-value	The user can restrict the records to only the subset that has disease-SNP associations below a certain p-value according to the literature. For example, if the user enters 9, all records with a p-value greater than 10^{-9} will be filtered out.
Location of SNP	This has 5 options: Exon, Intron, 3-UTR, 5-UTR and Intergenic. Initially all are selected. The user may uncheck one or more than one location. Only the SNPs in the selected location will be displayed.
Type of SNP	This has 3 options: Missense, Non coding and cds-synon. Initially all are selected. The user may uncheck one or more than one type. Only the SNPs of the selected type will be displayed.

Ethnicity	This has 4 options: Caucasian, Han Chinese, Japanese and ALL. Initially the option All is selected. The user may only select GWAS that have been conducted on a specific population by unchecking the All option and selecting one or more of the other options. Selecting the All option would also include Meta-analyses and studies conducted on multiple populations.

Table V. Filters in the interface



dbAARD Query Form

● Filters

○ Attributes

□ Diseases — Age-related Macular Degeneration
Alzheimer's Disease
Amyloid Lateral Sclerosis
Atrial Fibrillation

□ Genes — Enter HGNC gene symbols here — (Enter the symbol of each gene on a new line)

□ SNPs — Enter rsIDs here — (Enter each rsID on a new line)

□ p-value — <E-

□ Location of SNP — ☑ Exon ☑ Intron ☑ 3-UTR ☑ 5-UTR ☑ Intergenic

□ Type of SNP — ☑ Missense ☑ Non-coding ☑ cds-synon

□ Ethnicity — □ Caucasian □ Han Chinese □ Japanese ☑ All

Figure 20. Filters in the interface

To select a filter, the user must tick its checkbox. Similarly, to remove a filter, it must be unchecked. Only records that pass through all the selected filters will be displayed. For example, if the user selects AMD in the disease filter and Exon and Intergenic in the Location filter, only exonic and intergenic SNPs that cause AMD will be displayed.

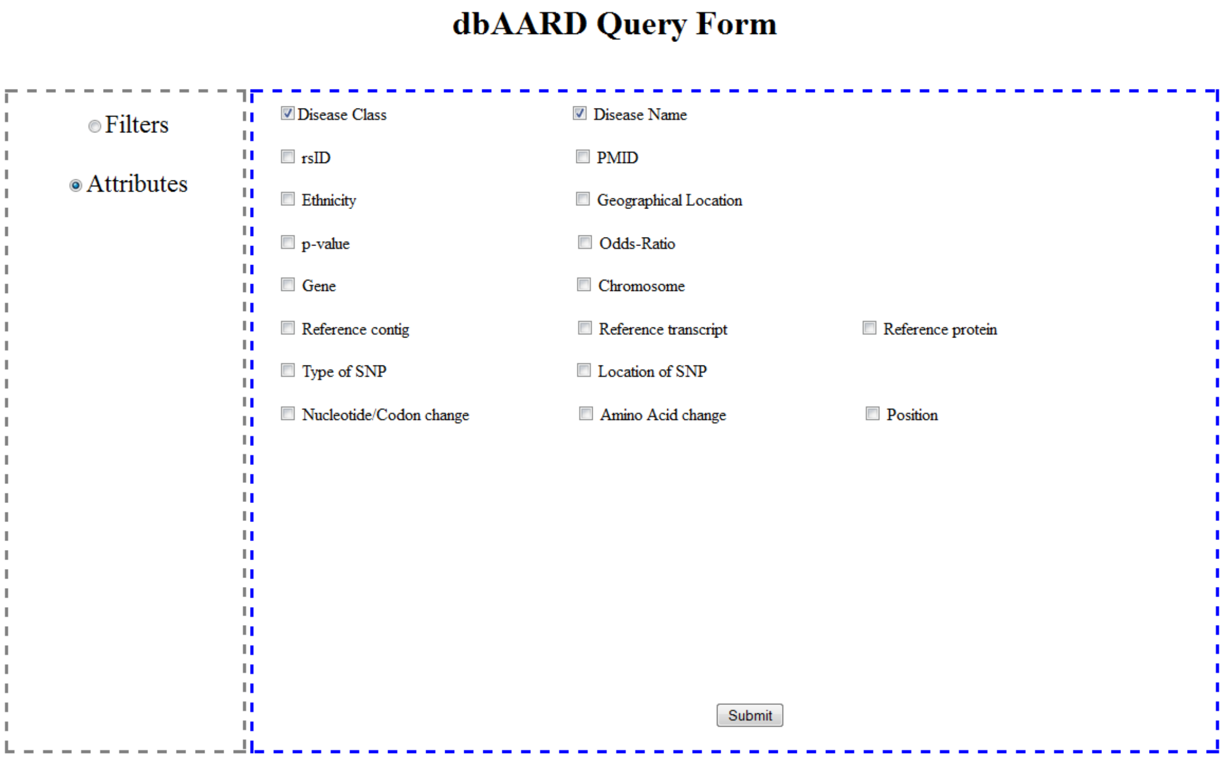
The user must exercise caution while selecting filters. For instance, if he/she selects Intron in the Location filter and Missense in the Type filter, no records will be displayed as there no missense mutations in introns.

If the user tries to submit the query form without selecting any filter, an error message will be displayed and the query form will not be submitted. Also, if the user selects the Gene, SNP or Disease filter and does not specify any gene, SNP or disease in the corresponding textbox or list box, an error message will be displayed on form submission and the filter will be automatically removed.

Attributes: These are the categories which will be displayed in the results. The attributes can be viewed by selecting the Attributes radio button on the left. The user can select an attribute by ticking its check box. Users may want to see only the information of their interest in the results instead of all the information present in a record. Selecting attributes allows them to view only the desired information without cluttering the screen with data which is of no interest to them. The following attributes can be selected (please see the section “Creation of disease-SNP list” in the Methodology for more details on each of these attributes):

- Disease Class
- Disease Name
- rsID
- PMID
- Ethnicity
- Geographical location
- p-value
- Odds-ratio
- Gene
- Chromosome
- Reference contig
- Reference transcript
- Reference protein
- Type of SNP
- Location of SNP
- Nucleotide/Codon change
- Amino Acid Change
- Position

The Disease Class and Disease Name attributes are selected by default. The records are displayed in a tabular manner with the attributes as columns in the same order as given above. The records are sorted by the first column of the table. So if a user has not selected the Disease Class column, the records will be ordered by Disease Name.



dbAARD Query Form

○ Filters

● Attributes

☑ Disease Class ☑ Disease Name

□ rsID □ PMID

□ Ethnicity □ Geographical Location

□ p-value □ Odds-Ratio

□ Gene □ Chromosome

□ Reference contig □ Reference transcript □ Reference protein

□ Type of SNP □ Location of SNP

□ Nucleotide/Codon change □ Amino Acid change □ Position

Figure 21. Attributes in the interface

Perl CGI script:

The CGI script receives the query form from the browser and retrieves the records from the database after filtering them as required by the user. The attributes that are selected by the user are then obtained from the filtered records and sent to the server which then sends them to the browser to be displayed. The detailed procedure is as follows:

1. Connect to the database using the DBI module. This requires knowing the database name, the hostname and port as well as the username and password that allows one to access the database.
2. Get all the checked items in the form as a hash. The keys of the hash are the names of the selected check boxes. This includes all the selected filters and attributes.
3. Create two hashes: *records* and *rsids*.
4. For Disease, p-value and Ethnicity filters, select the record numbers that satisfy the filtering criteria from the table *studied_by* and store them in the hash *records*. The keys of the hash are the filter names and the values are the record numbers stored as an array.
5. For Gene and SNP filters, first check if the genes and SNPs entered by the user are present in the database. If they are not, print the genes or SNPs that were not found before the rest of the results.

6. For the Gene, SNP, Type and Location filters, select the rsIDs that satisfy the filtering criteria from in the table *snp* and store them in the hash *rsids*. The keys of the hash are the filter names and the values are the rsIDs stored as an array.
7. Construct two arrays, one containing the elements common to all keys of *records* and the other containing elements common to all keys of *rsids*. Thus each of these arrays contains record numbers or rsIDs that have passed through all the filters.
8. Convert the rsIDs to record numbers. There may be many records associated with one rsID.
9. Create a final array of the elements common to both the arrays containing record numbers. These are the records that are finally selected. We now need to get the required attributes corresponding to the records.
10. Generate queries to retrieve the selected attributes from the given array of records. The queries are stored in an array *queries.*
11. Iterate through the array *queries* and run each query on the database. Store the retrieved results for each record in an array.
12. Format this row as a row of an HTML table by using the HTML::Template module and print it to the screen.

Appendix VII contains sample SQL queries for these steps.

The procedure for querying a database using the DBI module in Perl is as follows:

1. Prepare the query using the database handle. For example, if the database handle is *connxn*, use the connxn->prepare command.
2. Execute the prepared query. If the query has '?' as placeholders, then the appropriate variables should be included as parameters to the execute statement.
3. Bind the result of the query to a variable using the bind_columns command. The reference to the binding variable must be passed as a command parameter.
4. Use the fetch command to retrieve the results and store them in variables.

Appendix VIII shows an example of retrieving records from the database using placeholders in the query. The full code for the Perl CGI program can be found in Appendix IX.

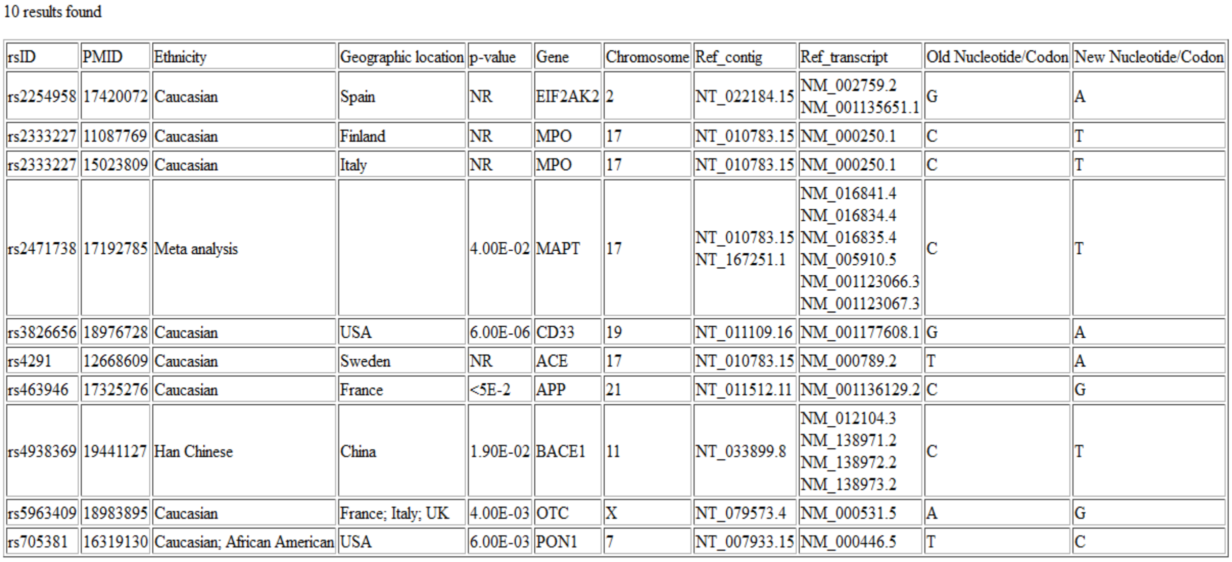
Example usage:

In conclusion, it is instructive to see a real-life example of the use of this database to solve a problem. Translational control is a prominent method of regulating eukaryotic gene expression. Translational control mostly occurs at the level of initiation, thus implicating the 5′ untranslated region as a major site of translational regulation. Suppose a user wants to see whether there is any support in the GWAS literature for association between SNPs in 5-UTR and Alzheimer's disease. The procedure is as follows:

1. Select Alzheimer's disease from the Diseases list box. The Diseases filter will be automatically selected.
2. Tick the check box for Location. From the options given, select 5-UTR and uncheck the rest.

3. Click on the radio button for attributes.
4. Remove the check marks against Disease Class and Disease Name since the user is only querying for one disease.
5. Select the attributes rsID, PMID, Ethnicity, Geographical location, p-value, Gene, Chromosome, Reference contig, Reference transcript and Nucleotide/Codon change. Since it is known that the SNPs are in the 5-UTR and are non-coding, we do not need to select the Location and Type attributes.
6. Click on Submit.

The following results will be displayed:



10 results found

rsID	PMID	Ethnicity	Geographic location	p-value	Gene	Chromosome	Ref_contig	Ref_transcript	Old Nucleotide/Codon	New Nucleotide/Codon
rs2254958	17420072	Caucasian	Spain	NR	EIF2AK2	2	NT_022184.15	NM_002759.2 NM_001135651.1	G	A
rs2333227	11087769	Caucasian	Finland	NR	MPO	17	NT_010783.15	NM_000250.1	C	T
rs2333227	15023809	Caucasian	Italy	NR	MPO	17	NT_010783.15	NM_000250.1	C	T
rs2471738	17192785	Meta analysis		4.00E-02	MAPT	17	NT_010783.15 NT_167251.1	NM_016841.4 NM_016834.4 NM_016835.4 NM_005910.5 NM_001123066.3 NM_001123067.3	C	T
rs3826656	18976728	Caucasian	USA	6.00E-06	CD33	19	NT_011109.16	NM_001177608.1	G	A
rs4291	12668609	Caucasian	Sweden	NR	ACE	17	NT_010783.15	NM_000789.2	T	A
rs463946	17325276	Caucasian	France	<5E-2	APP	21	NT_011512.11	NM_001136129.2	C	G
rs4938369	19441127	Han Chinese	China	1.90E-02	BACE1	11	NT_033899.8	NM_012104.3 NM_138971.2 NM_138972.2 NM_138973.2	C	T
rs5963409	18983895	Caucasian	France; Italy; UK	4.00E-03	OTC	X	NT_079573.4	NM_000531.5	A	G
rs705381	16319130	Caucasian; African American	USA	6.00E-03	PON1	7	NT_007933.15	NM_000446.5	T	C

Figure 22. Results for example query

It can be seen that there are various SNPs in the 5' untranslated region which have been implicated in GWAS for Alzheimer's disease. Hence the database saves the user the trouble of tedious literature searches and provides a quick way of finding relevant information.

DISCUSSION AND FUTURE PERSPECTIVES

Ageing and age-related disorders (ARDs) are complex phenotypes, involving many genes, variations and pathways that interact with the individual's environment. Till a few years ago, most of the genetic variations responsible for aging and ARDs were unknown. But since the completion of the Human Genome Project and the HapMap project, researchers have access to a large database of human variations found in different populations. The aim of GWA studies is to associate these variations with complex genetic disorders. In the last 5-6 years, GWA studies have grown increasingly larger in scale and are detecting variations at ever greater resolution. Since these studies involve a large number of SNPs and many of them have been found associated with various disorders, a vast amount of data is being generated. It is essential to compile and curate the data at a single location where it can be easily accessed so as to facilitate analysis.

This main objective of this work is to create a comprehensive database of variations associated with age-related disorders. Though there are databases like the NHGRI GWAS catalog that can be used to search for SNPs related to a particular disorder, this database has several points of difference from the existing ones. Firstly, it is focused on age-related disorders. Secondly, it contains information from not only the GWAS but also information about the SNP from dbSNP, so that the user gets all the information he/she needs from a single resource. Thirdly, since the disease-SNP list is compiled from several sources like the GWAS catalog, SNPedia, PubMed and others, it is more comprehensive than all of them. It was seen during curation that several GWAS studies that were found in, say GWAS Central were not found in the NHGRI catalog, for example [111, 112] in the case of glaucoma. This may be because the NHGRI catalog only contains those GWAS that attempted to assay at least 100,000 SNPs in the initial stage. In any case, this disease-SNP list contains more associations than any one of its sources. Also, by keeping only the SNPs having a p-value <0.05, it is ensured that only significant associations are included. Further, an easy-to-use graphical user allows users to analyze the information in the database without having any programming or database management experience.

This data allows the user to conduct an analysis of the genetic causes of ARDs on a genome-wide of systems level. It is possible to create networks of gene-disease as well as SNP-disease associations so as to make it easier to detect genes and SNPs that are significantly involved in different ARDS. For example, the figure below shows the disease-gene network. Table III shows that it contains ~1800 nodes and >6000 edges. The fact that the network diameter = 8 means that the graph is loosely connected. The diameter of this network has increased to 8 because a large number of genes are only involved in one disorder. Again, some facts can be clearly seen. Schizophrenia and Type 2 diabetes are the largest nodes again, as is expected from the figure. Both of them have a large number of genes that are exclusive only to these diseases, however, they also share many genes with other diseases. This graph also allows us to know which genes are most involved in age-related disorders. We see that the largest nodes in genes are present in the middle of the graph. These belong to genes like APOE, MTHFR, VEGFA, ACE, HLA-DQB1, NOS3, SOD2, TP53, PON1, IL1B and TNF. Each of these genes belongs to the 'mixed' class, which means that it is involved in more than one type of disorder. Each of these genes is well characterized. A summary of the gene functions can be found at the NCBI Gene database (http://www.ncbi.nlm.nih.gov/gene).

The color coding scheme is given in Appendix IX.

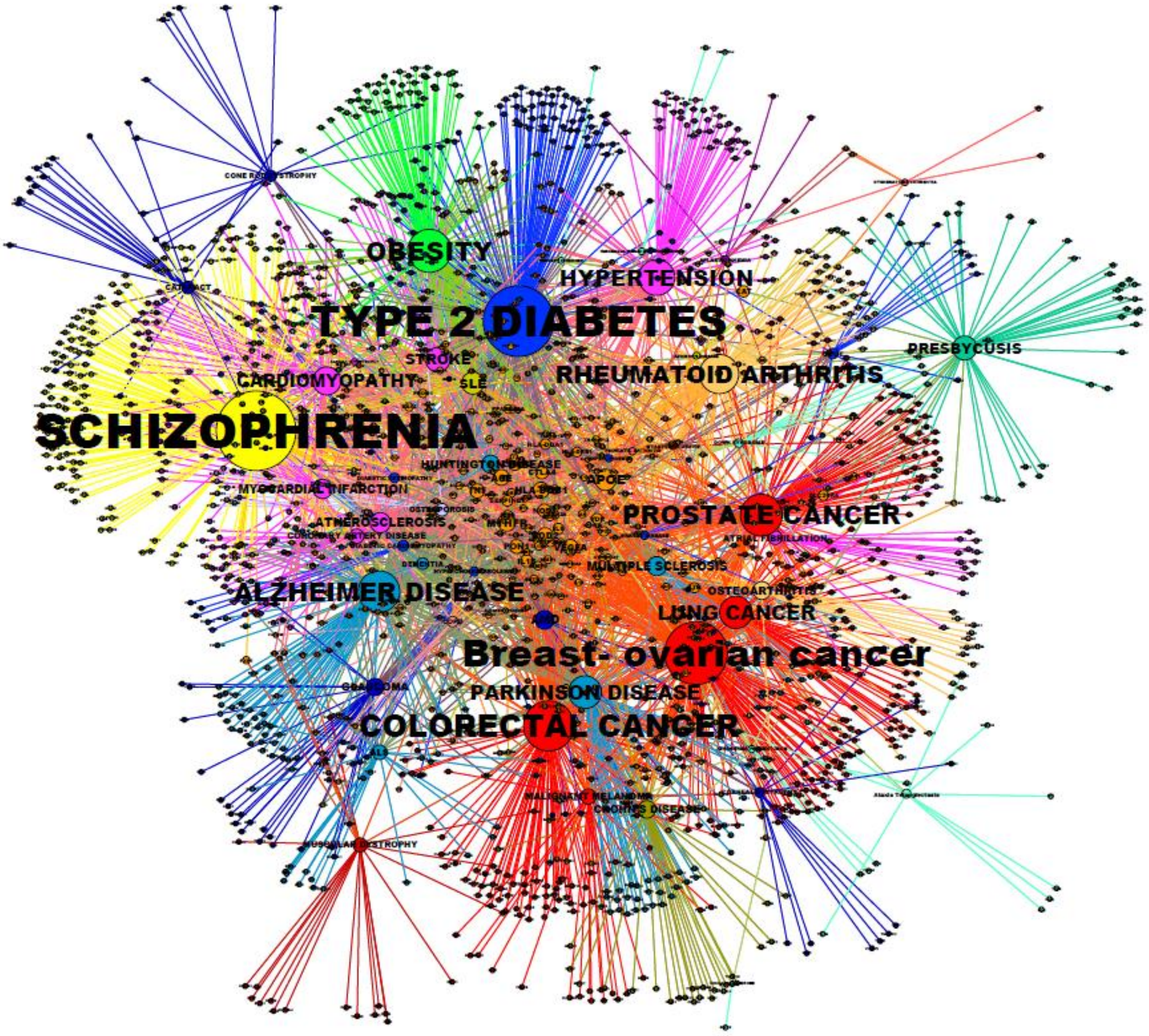


Figure 23. Disease-gene network

As another instance of the networks that can be constructed from this type of data, the figure below shows the disease-SNP network, which is the most loosely connected of all (diameter=18). This is because most SNPs are only involved in only one disease, and conversely, most diseases share only a few SNPs. One SNP, rs11868035, is implicated in as many as 5 different disorders, like Type 2 diabetes and Parkinson's disease, each belonging to a different class. Surprisingly, this is an intronic SNP in the retinoic acid induced (RAI1) gene, which is mainly expressed at high levels in neuronal tissues, which explains its association with Parkinson’s disease and schizophrenia at least. The other SNPs are involved in at most 2 diseases.

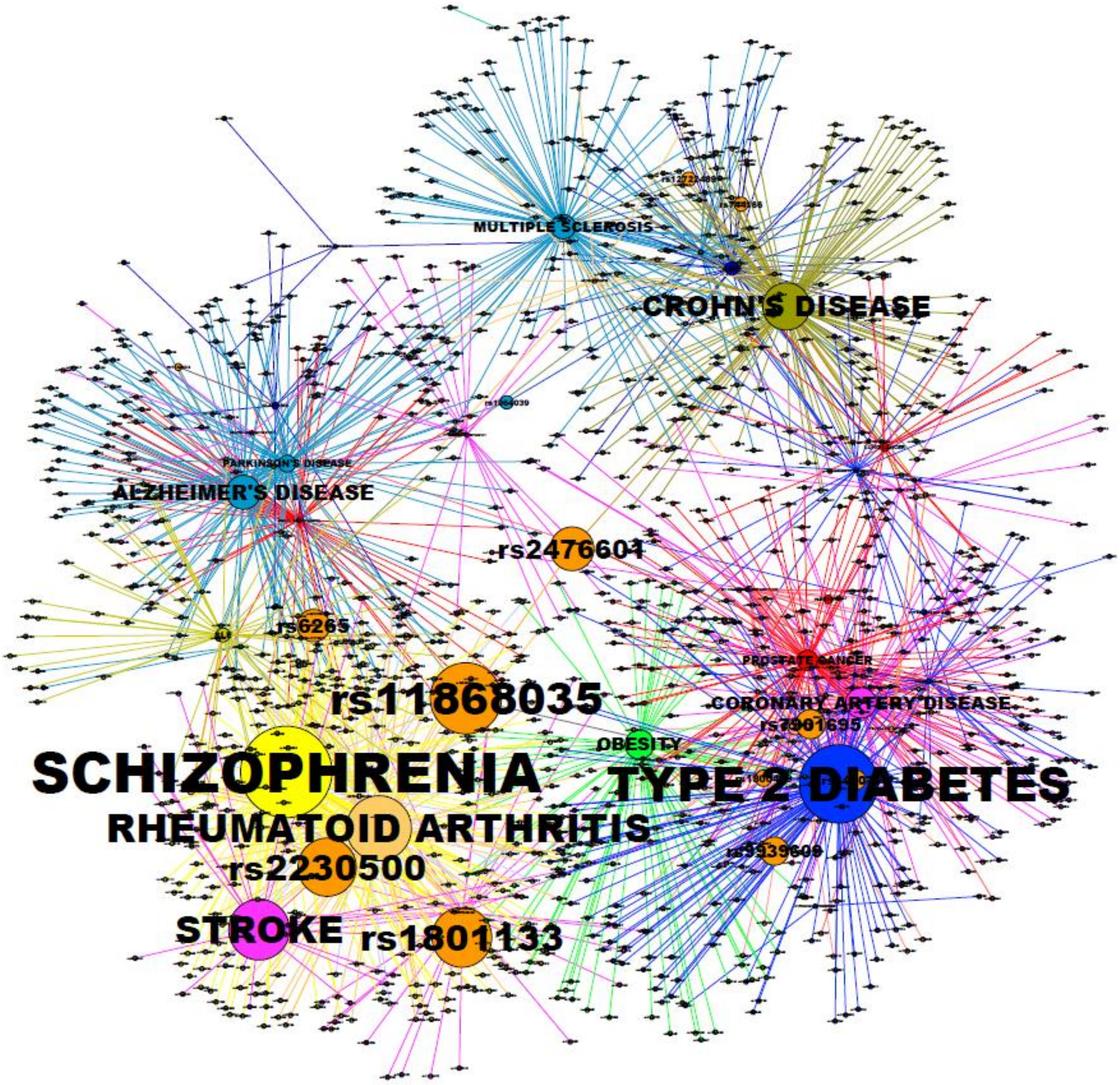


Figure 24. Disease-SNP network

These networks offer a rapid visual reference of the genetic links between disorders and disease genes and SNPs. The genetic networks are an easy way to see the genetic linkages between, say, type-2 diabetes and obesity, which are known to occur together in many individuals. They also reveal the genetic similarity underlying various neurological disorders and cancers.

Another way of analysis is at the level of individual genes and SNPs. It can be seen from Table III that a large number of SNPs implicated in ARDs are non-coding. It is a topic of great interest as to what role non-coding SNPs play in causing a disease. Many of these SNPs are in intergenic regions, that is, regions of the genome in which no gene exists. Again, it is important to know why these SNPs are found to be significantly associated with a disease. One reason could be that enhancer elements exist in these regions and when they are disrupted by mutation there is change in the expression of the regulated gene, which gives rise to disease. Since this database provides a long list of such SNPs, there is enough data at hand to attempt a thorough analysis. Moreover, it should be interesting to find the exact reason for a particular SNP in an intronic region being associated with a disease. Introns are known to contain splice sites and it is possible that a mutation disrupts such a site and causes translation of a malformed protein. Introns may also be part of micro RNAs that are important in gene expression regulation. Finally, the exonic SNPs present in the database may be further investigated as well, to find out the change they cause in protein structure that leads to its malfunctioning.

Though this work does provide a rich source of data for analysis, there is still scope for improvement. The database may be made more relevant by including more diseases and making the work more thorough. As GWA studies continue to be published at a fast rate, it is important to keep the database updated with the new information. It may also be needed to update the rsIDs in the database to keep them consistent with the newer builds of dbSNP. Moreover, it is important to take the feedback from biologists and bioinformaticians regarding the type of information that is present in the database and whether the GUI is suitable for their needs. Changes to the interface may be required from time to time according to the needs of the scientific community.

But important limitations of GWA studies should also be kept in mind while attempting an analysis of this data. Firstly, as seen from Table III, there is a predominance of GWAS conducted on Caucasian populations as opposed to others like Asian and African. This may lead to skewed analysis as the variations that are associated with disease in one population may not be present in another. Hence, unless more studies are conducted in other populations as well, we may not be able to catalog all the variations that give rise to a disease. Another problem is the great potential of GWAS for generating false-positive results. Because they test hundreds of thousands of statistical hypotheses—one for each allele or genotype assessed—GWA studies generate a large number of spurious associations. Another limitation of GWA studies is their lack of power for identifying associations with rare sequence variants, since these are poorly represented on current genotyping platforms. For example, in 2009, a medical team led by Nina P. Paynter of Brigham and Women’s Hospital in Boston collected 101 genetic variants that had been statistically linked to heart disease in various genome-scanning studies. But the variants turned out to have no value in forecasting disease among 19,000 women who had been followed for 12 years [113]. The limited information available on environmental exposures and other non-genetic risk factors in GWA studies makes it difficult to identify gene-environment interactions, or modification of gene-disease associations in the presence of environmental factors. However, even modest association can show the way to therapy for a disease.

Mining the database for biologically meaningful information keeping these points in mind is likely to reveal hitherto unknown facts about the underlying causes of age-related disorders. This knowledge may eventually be helpful in understanding the biology of aging itself.

REFERENCES

1. Fowden G. The Egyptian Hermes. Cambridge University Press, Cambridge, 1987.

2. Needham J, Ping-Yu Ho, Gwei-Djen Lu. Science and Civilisation in China, Volume V, Part III. Cambridge at the University Press, 1976.

3. Ragai J. The Philosopher's Stone: Alchemy and Chemistry. Journal of Comparative Poetics. 1992; 12 (Metaphor and Allegory in the Middle Ages): 58–77.

4. Leaf A. Long-lived populationa. J Am Geriatr Soc 1982; 30: 485-487.

5. Leaf A. The aging process: lessons from observations in man. Nutr Rev 1988; 46:40-4.

6. Mazess RB, Forman S. Longevity and age exaggeration in Vilcabamba, Ecuador. Journal of Gerontology 1979; 34:94-8.

7. Palmore EB. Longevity in Abkhazia: a re-evaluation. The Gerontologist 1984; 24: 95-96.

8. Population division, United Nations (2002). World Population Ageing: 1950-2050.

9. Bravo, J. Fiscal Implications of Ageing Societies Regarding Public and Private Pension Systems. 1999; In Population Ageing: Challenges for Policies and Programmes in Developed and Developing Countries, R. Cliquet and M. Nizamuddin, eds. New York: United Nations Population Fund; and Brussels: Centrum voor Bevolkings-en Gezinsstudiën (CBGS).

10. Cliquet, R., and Nizamuddin, M. eds. Population Ageing: Challenges for Policies and Programmes in Developed and Developing Countries. 1999; New York: United Nations Population Fund; and Brussels: Centrum voor Bevolkings-en Gezinsstudiën (CBGS).

11. Creedy, J. Pensions and Population Ageing: An Economic Analysis. 1998; Cheltenham, United Kingdom; and Northampton, Massachusetts: Edward Elgar Publishing.

12. de Jong-Gierveld J., and H. van Solinge Ageing and its Consequences for the Socio- Medical System. 1995; Population Studies, No. 29. Strasbourg, France: Council of Europe Press.

13. Holliday, R. Ageing in the 21st century. Lancet. 1999; 354 Suppl: SIV4.

14. United Nations. 2006. World Population Prospects: 2006 revision.

15. Holliday R. Ageing in the 21st century. Lancet. 1999; 354 Suppl: SIV4.

16. Vijg J, Campisi J. Puzzles, promises and a cure for ageing. Nature. 2008 August 28; 454(7208): 1065–1071.

17. Population division, United Nations. 2002. World Population Ageing: 1950-2050.

18. Rowe JW, Kahn RL. "Successful ageing". Gerontologist. 1997; 37 (4): 433–40.

19. Strawbridge WJ, Wallhagen MI, Cohen RD. "Successful ageing and well-being: self-rated compared with Rowe and Kahn". Gerontologist. 2002; 42 (6): 727–33.

20. Poon LW, Clayton GM, Martin P, Johnson MA, Courtenay BC, Sweaney AL et al. The Georgia Centenarian Study. Int J Aging Hum Dev. 1992; 34; (1)1-17.

21. Lehr U. [Centenarian--a contribution to longevity research]. Z Gerontol. 1991; 24; (5)227-32.

22. Martin P, Raiser V, Poon, LW Bramlett, MA & Johnson MA. Personality and activity in the oldest-old. In G. Huber (ed.), Healthy aging, activity, and sports (pp. 243-253). 1997; Werbach-Gamburg: Health Promotions Publications.

23. Vaupel J. in Between Zeus and the Salmon: The Biodemography of Aging. 1997. K. Wachter and C. E. Finch, Eds. National Academy of Sciences, Washington, DC, pp. 17–37.

24. Finch CE. Longevity, Senescence, and the Genome. 1990. Univ. of Chicago Press, Chicago, IL.

25. Holliday R. The close relationship between biological aging and age-associated pathologies in humans. J Gerontol A Biol Sci Med Sci. 2004. 59; (6) B543-6.

26. Hayflick L. The not-so-close relationship between biological aging and age-associated pathologies in humans. J Gerontol A Biol Sci Med Sci. 2004. 59; (6) B547-50; discussion 551-3.

27. Vijg, J., Campisi, J. Puzzles, promises and a cure for ageing. Nature. 2008 August 28; 454(7208): 1065–1071.

28. Stankiewicz P, Lupski JR. Structural variation in the human genome and its role in disease. Annu Rev Med. 2010. 61; 437-55.

29. Lander ES. Genome-sequencing anniversary. The accelerator. Science. 2011. 331; (6020)1024.

30. Hardy J, Singleton A. Genomewide association studies and human disease. N Engl J Med. 2009. 360; (17)1759-68.

31. Janssens AC, van Duijn CM. Genome-based prediction of common diseases: advances and prospects. Hum Mol Genet. 2008. 17; (R2)R166-73.

32. Rubner M. Das Problem der Lebensdauer und seine Beziehungen sum Wachstum und Ernahrung. Munich, Germany: Oldenbourg. 1908.

33. Medawar P.B. An Unsolved Problem of Biology. London: H.K. Lewis. Edney, E.B. and Gill, R.W. 1968.

34. Harman D. Aging: a theory based on free radical and radiation chemistry. Journal of Gerontology 1956; 11 (3): 298–300.

35. Williams G.C. Pleiotropy, natural selection and the evolution of senescence. Evolution 1957; 11 (4): 398–411.

36. Hayflick L. The limited in vitro lifetime of human diploid cell strains. Exp. Cell Res. 1965; 37 (3): 614–636.

37. Kirkwood T.B. Evolution of ageing. Nature 1977; 270 (5635): 301–4.

38. Harley CB, Futcher, AB & Greider CW. Telomeres shorten during ageing of human fibroblasts. Nature 1990; 345: 458–460.

39. Sharpless NE & DePinho RA. How stem cells age and why this makes us grow old. Nature Rev. Mol. Cell Biol. 2007; 8: 703–713.

40. DePinho RA. et al. Telomerase reactivation reverses tissue degeneration in aged telomerase-deficient mice. Nature 2011; 469(7328):102-6.

41. Kenyon CJ. The genetics of ageing. Nature. 2010. 464; (7288)504-12.

42. Guarente L, Kenyon C. Genetic pathways that regulate ageing in model organisms. Nature 2000; 408 (6809): 255–62.

43. Chen D., Guarente L. SIR2: a potential target for calorie restriction mimetics. Trends Mol Med 2007; 13 (2): 64–71.

44. Kaeberlein M., McVey M., and Guarente L. The SIR2/3/4 complex and SIR2 alone promote longevity in Saccharomyces cerevisiae by two different mechanisms. Genes Dev 1999; 13: 2570-2580.

45. Luo S., Murphy C.T. Caenorhabditis elegans reproductive aging: Regulation and underlying mechanisms. Genesis. 2011; 49(2):53-65.

46. Kaletsky R., Murphy C.T. The role of insulin/IGF-like signaling in C. elegans longevity and aging. Dis Model Mech. 2010; 3: 415-419.

47. Kaeberlein M., and Powers, R.W., 3rd. Sir2 and calorie restriction in yeast: A sceptical perspective. Ageing Res Rev 2007; 6: 128-140.

48. Rose MR. The Evolutionary Biology of Aging. 1991. Oxford Univ. Press, New York.

49. Martin GA, Austad SA, Johnson TE. Genetic analysis of ageing: role of oxidative damage and environmental stresses. Nature Genet. 1996; 13:25.

50. Wheeler JC, Bieschke ET, Tower J. Muscle-specific expression of Drosophila hsp70 in response to aging and oxidative stress. Proc. Natl. Acad. Sci. U.S.A. 1995; 92:10408.

51. Sohal RS, Weindruch R. Oxidative stress, caloric restriction, and aging. Science. 1996; 273:59.

52. Rogina B, Benzer S, Helfand SL. Drosophila drop-dead mutations accelerate the time course of age-related markers. Proc. Natl. Acad. Sci. U.S.A. 1997; 94:6303.

53. Klass MR. A method for the isolation of longevity mutants in the nematode Caenorhabditis elegans and initial results. Mech. Ageing Dev. 1983; 22:279–286.

54. Friedman DB, Johnson TE. A mutation in the age-1 gene in Caenorhabditis elegans lengthens life and reduces hermaphrodite fertility. Genetics 1988; 118:75–86.

55. Kenyon C. et al. A C. elegans mutant that lives twice as long as wild type. Nature. 1993; 366(6454):461-4.

56. Kenyon C. The plasticity of aging: insights from long-lived mutants. Cell 2005; 120:449–460.

57. Amrit FR, May RC. Younger for longer: insulin signalling, immunity and ageing. Curr Aging Sci. 2010; 3(3):166-76.

58. Bjedov I, Partridge L. A longer and healthier life with TOR down-regulation: genetics and drugs. Biochem Soc Trans. 2011; 39(2):460-5.

59. Choksi KB, Nuss JE, DeFord JH, Papaconstantinou J. Mitochondrial electron transport chain functions in long-lived Ames dwarf mice. Aging (Albany NY). 2011; 3(8):754-67.

60. Evert J., Lawler E., Bogan H., Perls T. Morbidity profiles of centenarians: survivors, delayers, and escapers. J Gerontol A Biol Sci Med Sci 2003; 58:232–237.

61. Terry D.F., Wilcox M., McCormick M.A., Lawler E., Perls T.T. Cardiovascular advantages among the offspring of centenarians. J Gerontol A Biol Sci Med Sci 2003; 58:M425–M431.

62. Barzilai N., Gabriely I., Gabriely M., Iankowitz N., Sorkin J.D. Offspring of centenarians have a favorable lipid profile. J Am Geriatr Soc 2001; 49:76–79.

63. Evert J., Lawler E., Bogan H., Perls T. Morbidity profiles of centenarians: survivors, delayers, and escapers. J Gerontol A Biol Sci Med Sci 2003; 58:232–237.

64. Barzilai N., Gabriely I., Gabriely M., Iankowitz N., Sorkin J.D. Offspring of centenarians have a favorable lipid profile. J Am Geriatr Soc 2001; 49:76–79.

65. Brand F.N., Kiely D.K., Kannel W.B., Myers R.H. Family patterns of coronary heart disease mortality: the Framingham Longevity Study. J Clin Epidemiol 1992;.45:169–174.

66. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J et al. Initial sequencing and analysis of the human genome. Nature. 2001. 409: (6822)860-921.

67. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG et al. The sequence of the human genome. Science. 2001. 291; (5507)1304-51.

68. International HapMap Consortium. The International HapMap Project. Nature. 2003. 426: (6968)789-96.

69. Collins FS, Guyer MS, Chakravarti A. Variations on a theme: cataloging human DNA sequence variation. Science. 1997; 278:1580–1581.

70. International HapMap 3 Consortium. Altshuler DM, Gibbs RA, Peltonen L, Altshuler DM, Gibbs RA et al. Integrating common and rare genetic variation in diverse human populations. Nature. 2010. 467; (7311)52-8.

71. Policy for sharing of data obtained in NIH supported or conducted genome-wide association studies (GWAS). Federal Register 8/30/07. 2007 [accessed 6/11/2011]. http://grants.nih.gov/grants/guide/notice-files/NOT-OD-07-088.html

72. Klein RJ, Zeiss C, Chew EY, et al. Complement factor H polymorphism in age-related macular degeneration. Science. 2005; 308:385–389.

73. Coon KD, Myers AJ, Craig DW, et al. A high-density whole-genome association study reveals that APOE is the major susceptibility gene for sporadic late-onset Alzheimer's disease. J Clin Psychiatry. 2007; 68(4):613-618.

74. Reiman EM, Webster JA, Myers AJ, et al. GAB2 alleles modify Alzheimer’s risk in APOE epsilon4 carriers. Neuron. 2007; 54(5):713-720.

75. Grupe A, Abraham R, Li Y, et al. Evidence for novel susceptibility genes for lateonset Alzheimer’s disease from a genome-wide association study of putative functional variants. Hum Mol Genet. 2007; 16(8):865-873.

76. Liu F, Arias-Vasquez A, Sleegers K, et al. A genomewide screen for late-onset Alzheimer disease in a genetically isolated Dutch population. Am J Hum Genet. 2007; 81(1):17-31.

77. Potkin SG, Turner JA, Guffanti G, Lakatos A, Fallon JH, Nguyen DD et al. A genome-wide association study of schizophrenia using brain activation as a quantitative phenotype. Schizophr Bull. 2009; 35: (1)96-108.

78. Need AC, Ge D, Weale ME, Maia J, Feng S, Heinzen EL et al. A genome-wide investigation of SNPs and CNVs in schizophrenia. PLoS Genet. 2009; 5: (2) e1000373.

79. Libioulle C, et al. Novel Crohn disease locus identified by genome-wide association maps to a gene desert on 5p13.1 and modulates expression of PTGER4. PLoS Genet. 2007; 3:e58.

80. Rioux JD, et al. Genome-wide association study identifies new susceptibility loci for Crohn disease and implicates autophagy in disease pathogenesis. Nature Genet. 2007; 39:596–604.

81. Hampe J, et al. A genome-wide association scan of nonsynonymous SNPs identifies a susceptibility variant for Crohn disease in ATG16L1. Nature Genet. 2007; 39:207–211.

82. Gudmundsson J, et al. Genome-wide association study identifies a second prostate cancer susceptibility variant at 8q24. Nature Genet. 2007; 39:631–637.

83. Gudmundsson J, et al. Two variants on chromosome 17 confer prostate cancer risk, and the one in TCF2 protects against type 2 diabetes. Nature Genet. 2007; 39:977–983.

84. Yeager M, et al. Genome-wide association study of prostate cancer identifies a second risk locus at 8q24. Nature Genet. 2007; 39:645–649.

85. Hunter DJ, et al. A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer. Nature Genet. 2007; 39:870–874.

86. Easton DF, et al. Genome-wide association study identifies novel breast cancer susceptibility loci. Nature. 2007; 447:1087–1093.

87. Stacey SN, et al. Common variants on chromosomes 2q35 and 16q12 confer susceptibility to estrogen receptor-positive breast cancer. Nature Genet. 2007; 39:865–869.

88. Gold B, et al. Variation in factor B (BF) and complement component 2 (C2) genes is associated with age-related macular degeneration. Nature Genet. 2006; 38:458–462.

89. Yates JR, et al. Complement C3 variant and the risk of age-related macular degeneration. N Engl J Med. 2007; 357:553–561.

90. Luke MM, et al. A polymorphism in the protease-like domain of apolipoprotein(a) is associated with severe coronary artery disease. Arterioscler Thromb Vasc Biol. 2007; 27:2030–2036.

91. Cupples LA, Arruda HT, Benjamin EJ, et al. The Framingham Heart Study 100K SNP genome-wide association study resources: overview of 17 phenotype working group reports. BMC Med. Genet. 2007; 8:S1.

92. Gulcher J, Kong A, Stefansson K. The genealogic approach to human genetics of disease. Cancer J. 2001; 7:61–68.

93. Zeggini E, Scott LJ, Saxena R, et al. Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. Nat. Genet. 2008; 40:638–645.

94. Wellcome Trust Case Control Consortium. Overview. http://www.wtccc.org.uk/info/overview.shtml.

95. Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature. 2007; 447:661–678.

96. Manolio TA, Collins FS. The HapMap and genome-wide association studies in diagnosis and therapy. Annu Rev Med. 2009; 60: 443-56.

97. Grant SF, Thorleifsson G, Reynisdottir I, et al. Variant of transcription factor 7-like 2 (TCF7L2) gene confers risk of type 2 diabetes. Nat. Genet. 2006; 38:320–323.

98. Sladek R, Rocheleau G, Rung J, et al. A genome-wide association study identifies novel risk loci for type 2 diabetes. Nature. 2007; 445:881–885.

99. Manolio TA, Brooks LD, Collins FS. A HapMap harvest of insights into the genetics of common disease. J. Clin. Invest. 2008; 118:1590–1605.

100. Maraganore DM, de Andrade M, Lesnick TG, et al. High-resolution whole-genome association study of Parkinson disease. Am. J. Hum. Genet. 2005; 77:685–693.

101. Herbert A, Gerry NP, McQueen MB, et al. A common genetic variant is associated with adult and childhood obesity. Science. 2006; 312:279–283.

102. Myers RH. Considerations for genomewide association studies in Parkinson disease. Am. J. Hum. Genet. 2006; 78:1081–1082.

103. Lyon HN, Emilsson V, Hinney A, et al. The association of a SNP upstream of INSIG2 with body mass index is reproduced in several but not all cohorts. PLoS Genet. 2007; 3:e61.

104. Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS et al.. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc Natl Acad Sci U S A. 2009; 106: (23)9362-7.

105. Cariaso M. and Lennon G. SNPedia. Available at: http://www.SNPedia.com/. Accessed [July, 2011 to October, 2011].

106. Thorrison GA, Lancaster O et al. HGVbaseG2P: a central genetic association database. Nucleic Acids Research, (2009) 37:D797-802.

107. Bertram L, McQueen MB, Mullin K, Blacker D, Tanzi RE. Systematic meta-analyses of Alzheimer disease genetic association studies: the AlzGene database. Nat Genet. 2007; 39(1): 17-23.

108. Allen NC, Bagade S, McQueen MB, Ioannidis JPA, Kavvoura FK, Khoury MJ, Tanzi RE, Bertram L. Systematic Meta-Analyses and Field Synopsis of Genetic Association Studies in Schizophrenia: The SzGene Database Nat Genet. 2008; 40(7): 827-34.

109. Agrawal S, Dimitrova N, Nathan P, Udayakumar K, Lakshmi SS, Sriram S, Manjusha N, Sengupta U. T2D-Db: an integrated platform to study the molecular basis of Type 2 diabetes. BMC Genomics. 2008; 7: 9:320.

110. Goh KI, Cusick ME, Valle D, Childs B, Vidal M, et al. The human disease network. Proceedings of the National Academy of Sciences. 2007; 104: 8685.

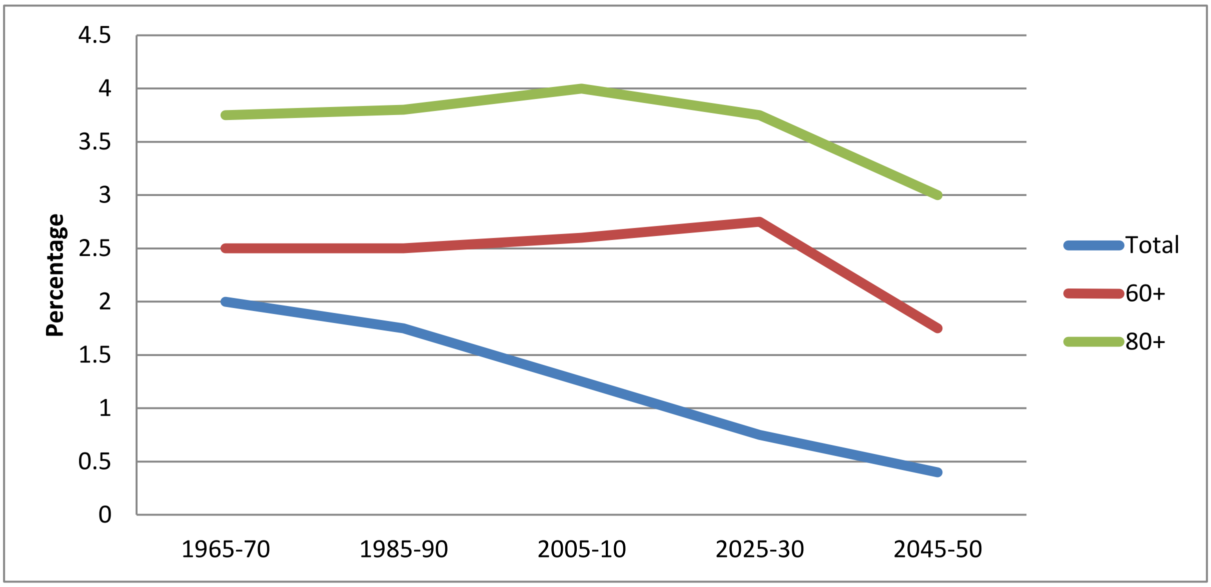
111. Thorleifsson G, Walters GB, Hewitt AW, Masson G, Helgason A, DeWan A et al. Common variants near CAV1 and CAV2 are associated with primary open-angle glaucoma. Nat Genet. 2010; 42: (10)906-9.

112. Thorleifsson G, Magnusson KP, Sulem P, Walters GB, Gudbjartsson DF, Stefansson H et al. Common sequence variants in the LOXL1 gene confer susceptibility to exfoliation glaucoma. Science. 2007; 317: (5843)1397-400.

113. Paynter NP, Chasman DI., Paré G. Association between a literature-based genetic risk score and cardiovascular events in women. JAMA. 2010; 303:631–637.

APPENDIX I

Average annual growth rate of population from 1965-2050

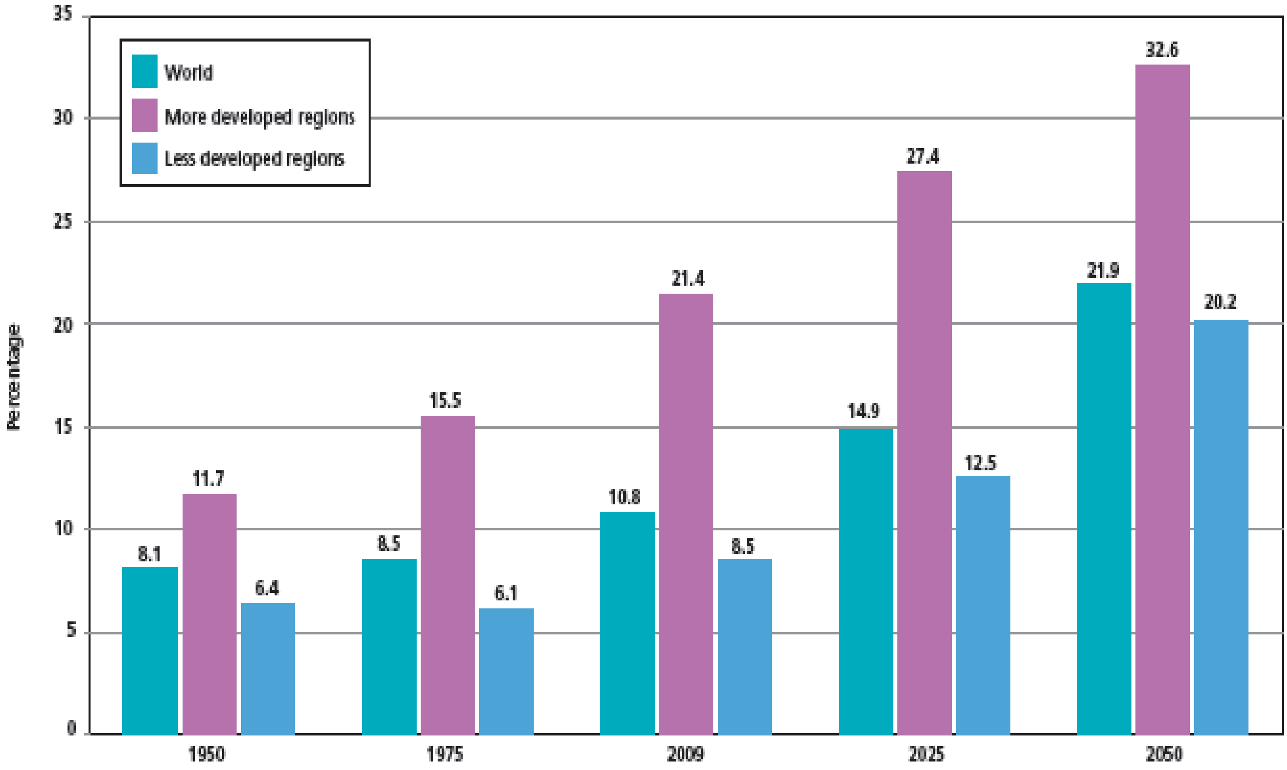


Average annual growth rate of the total population, the population aged 60 or over, and the population aged 80 or over, 1965-2050

Adapted from [8]

APPENDIX II

The population aged 60 or over as a percentage of the total population from 1950-2050



The population aged 60 or over as a percentage of the total population, world and development groups, 1950-2050. In 2050, 22 per cent of the world population is projected to be 60 years or over.

From [17]

APPENDIX III

Age distribution of the growth of Indian population from 1950-2050

Age	1950	1975	2000	2025	2050
Total	357561	620701	1008937	1351801	1572055
0-14	139156	247062	337921	314118	308996
15-59	198307	339150	594165	869174	938743
60-69	13466	25630	47646	100343	170381
70-79	5710	10939	23096	50807	105778
80-89			5735	15564	42194
90-99			371	1762	5822
100+			3	33	142

The numbers give the population in thousands. The number of people over 60 is expected to reach 300 million by 2050

Adapted from [8]

APPENDIX IV

List of diseases included in disease-SNP list

ALS
ALZHEIMER'S DISEASE
AMD
ATRIAL FIBRILLATION
BREAST- OVARIAN CANCER
CARDIAC HYPERTROPHY
COPD
CORNEAL DYSTROPHY
CORONARY ARTERY DISEASE
CROHN'S DISEASE
DEMENTIA
DIABETIC RETINOPATHY
DILATED CARDIOMYOPATHY
GLAUCOMA
GOUT
GRAVES' DISEASE
HYPERCHOLESTEROLEMIA
HYPERLIPIDEMIA
HYPERTENSION
LUNG CANCER
MULTIPLE SCLEROSIS
MYOCARDIAL INFACTION
OBESITY
OSTEOARTHRITIS
OSTEOPOROSIS
PARKINSON'S DISEASE
PRESBYCUSIS
PROSTATE CANCER
PSORIATIC ARTHRITIS
RHEUMATOID ARTHRITIS
SCHIZOPHRENIA
SLE
STROKE
TYPE 2 DIABETES

APPENDIX V

Classification of diseases

CATEGORY	DISEASE			
BONE	OSTEOPOROSIS			
CANCER	BREAST- OVARIAN CANCER	COLORECTAL CANCER	LUNG CANCER	PROSTATE CANCER
CARDIOVASCULAR	CARDIAC HYPERTROPHY	ATRIAL FIBRILLATION	STROKE	DILATED CARDIOMYOPATHY
	CORONARY ARTERY DISEASE	HYPERTENSION	MYOCARDIAL INFARCTION	
CONNECTIVE TISSUE	OSTEOARTHRITIS	RHEUMATOID ARTHRITIS		
DERMATOLOGICAL	PSORIATIC ARTHRITIS			
ENT	PRESBYCUSIS			
GASTROINTESTINAL	CROHN'S DISEASE			
IMMUNOLOGICAL	SLE			
METABOLIC	HYPERCHOLESTEROLEMIA	DIABETIC RETINOPATHY	GOUT	GRAVES DISEASE
	TYPE 2 DIABETES	HYPERLIPIDEMIA		
NEUROLOGICAL	ALZHEIMER DISEASE	ALS	DEMENTIA	PARKINSON DISEASE
	MULTIPLE SCLEROSIS			
NUTRITIONAL	OBESITY			
OPTHALMOLOGICAL	AMD	GLAUCOMA	CORNEAL DYSTROPHY	
PSYCHIATRIC	SCHIZOPHRENIA			
RESPIRATORY	COPD			

APPENDIX VI

Sample of a list record in the disease-SNP list

Class	NEUROLOGICAL
Disease	ALZHEIMER’S DISEASE
rsID	rs429358
PMID	17152785
Population	Meta-analysis
Geographical location	
p-value	5E-34
Odds-ratio (OR)	NR
Gene	APOE
Chromosome	19
Ref_contig	NT_011109.1
Ref_mRNA	NM_000041.2
Ref_protein	NP_000032.1
Type	Missense
Old Nucleotide/Codon	TGC
New Nucleotide/Codon	CGC
Old Amino Acid	C
New Amino Acid	R
Location	Exon
Position	130
Remarks	
Annotator	Vaibhav

APPENDIX VII

Sample SQL queries

1. To count the number of occurrences of gene TP53 in the table *snp*:

```
SELECT COUNT(*) FROM snp WHERE Gene = ‘TP53’;
```

2. For filters:

a) To select rsIDs from table *snp* for gene TP53:

```
SELECT rsID FROM snp WHERE Gene = ‘TP53’;
```

b) To select Record numbers from table *studied_by* where p-value is equal to 10^{-9}:

```
SELECT Record FROM studied_by WHERE ` p-value` = ‘1E-9’;
```

c) To select Record numbers from table *studied_by* where Ethnicity of reference GWAS population is Caucasian:

```
SELECT Record FROM studied_by WHERE ref_No IN (SELECt PMID FROM reference WHERE
Population_ethnicity LIKE '%Caucasian%');
```

3. To convert rsID rs699 to their corresponding Record numbers:

```
SELECT Record FROM studied_by WHERE snp_No = ‘rs699’;
```

4. For attributes:

a) To select disease for Record number 14:

```
SELECT Disease Name FROM studied_by WHERE Record = 14;
```

b) To select Gene for Record number 14:

```
SELECT Gene FROM snp WHERE rsID IN (SELECT snp_No FROM studied_by WHERE Record =
14);
```

c) Select reference mRNA for Record number 14:

```
SELECT Ref_mRNA FROM transcripts WHERE sid IN (SELECT rsID FROM snp WHERE rsID IN
(SELECT snp_No FROM studied_by WHERE Record = 14));
```

APPENDIX VIII

Retrieving information from a database using Perl

```
my $database = "dbaard";

my $host = "localhost";

my $port = "3306";

my $user = "root";

my $pw = "";

my $dsn = "dbi:mysql:$database:$host:$port";

our $connxn = DBI->connect($dsn, $user, $pw) ;

$sth = $connxn->prepare("select Record from studied_by where snp_No = ?");

foreach my $el (@rsids) {

        $sth->execute($el);

        $sth->bind_columns(undef, \$rec);

        while($sth->fetch()) {

                push(@record, $rec);

        }

}
```

This Perl code is used to get Record numbers corresponding to rsIDs in the table *studied_by*. The rsIDs are present in the array *rsids*. The query contains a placeholder '?' which is replaced by the rsID as the program iterates over the array in the foreach loop. The program demonstrates preparation and execution of query as well as fetching of results.

APPENDIX IX

Perl CGI code for connecting the interface to the database on the XAMPP server

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

use strict;

use warnings;

use DBI;

use CGI;

use Carp;

use HTML::Template;


my $database = "dbaard";

my $host = "localhost";

my $port = "3306";

my $user = "root";

my $pw = "";

my $dsn = "dbi:mysql:$database:$host:$port";

our $connxn = DBI->connect($dsn, $user, $pw) or die $!;


our $query = CGI->new;

our @queries = ();

our @rows = ();


print $query->header('text/html'),

$query->start_html(-title => 'dbAARD results');
```

```
our %FORM = get_checked();

my ($records, $rsids) = get_records_snps();

my @snps = construct_arr($rsids);

my @recs = construct_arr($records);

my @recs2 = rsid2rec(@snps);

my @final = final_records(\@recs, \@recs2);

gen_queries();

gen_results(@final);

create_table();


print $query->end_html;


sub get_checked {

        my %FORM = $query->Vars;

        return %FORM;

}


sub get_records_snps {

        my %records; my %rsids;


        if ( $FORM{Diseases_f} ) {

                my @dis = check_groups("Disease_filter", "Disease", "studied_by");

                $records{dis} = \@dis;

        }


        if( $FORM{Genes_f} ){

                my $gene_text = $FORM{Genes_text};
                my @genes = gene_snp_filter($gene_text, "Gene");
                $rsids{genes} = \@genes;
        }

        if( $FORM{SNP_f} ){
                my $snp_text = $FORM{SNP_text};
                my @snps = gene_snp_filter($snp_text, "rsID");
                $rsids{snps} = \@snps;
        }

        if( $FORM{p_value_f} ){
                my $rec;
                my @arr = ();
                my $pval = $FORM{pval_exp};
                my $sth = $connxn->prepare("select distinct `p-value` from studied_by");
                $sth->execute();
                $sth->bind_columns(undef, \$rec);
                while($sth->fetch()) {
                        if ($rec =~ /\d+$/) {
                                if ($& > $pval) { push(@arr, $rec); }
                        }
                }
                my @pvals = retrieve_records(@arr, "`p-value`", "studied_by");
                $records{pvals} = \@pvals;
        }
```

```
        if ( $FORM{Type_f} ) {
                my @types = check_groups("Type_opt", "Type", "snp");
                $rsids{types} = \@types;
        }

        if ( $FORM{Location_f} ) {
                my @locs = check_groups("Loc_opt", "Location", "snp");
                $rsids{locs} = \@locs;
        }

        if ( $FORM{Ethnicity_f} ) {
                my @arr = $query->param("Ethn_opt");
                my %arr_hash = map { $_ => 1 } @arr;
                if(!exists($arr_hash{"All"})) {
                        my @ethns = check_groups("Ethn_opt", "Population_Ethnicity",
"reference");
                        $records{ethns} = \@ethns;
                }
        }
        return (\%records, \%rsids);
}

sub gene_snp_filter {
        (my $text, my $col) = @_;
        my @arr = split('\n', $text);
        @arr = filters(@arr, $col);
        my @ret = retrieve_records(@arr, $col, "snp");
        return @ret;
}

sub check_groups {
        (my $name, my $col, my $table) = @_;
        my @arr = $query->param($name);
        print $query->p("$_") for @arr;
        my @ret = retrieve_records(@arr, $col, $table);
        return @ret;
}

sub filters {
        my @arr = @_[0..@_-2];
        my $col = $_[-1];
        my @arr1 = ();
        my $rec;
        my $i = 0;
        my $sth = $connxn->prepare("select COUNT(*) from snp where $col = ?");
        while (defined $arr[$i]) {
                $arr[$i] =~ s/\s+$//;
                my $el = uc $arr[$i];
                $sth->execute($el);
                $sth->bind_columns(undef, \$rec);
                while($sth->fetch()) {
                        if ($rec == 0) {
                                push(@arr1, $arr[$i]);
                                splice(@arr, $i, 1);
                $i--;
            }
        }
        $i++;
    }
    my $inval = join(" ", @arr1);
    if ($inval ne "" ) {
        print $query->p("The following $col were not found: $inval");
    }

    my %seen;
    my @unique = grep  ! $seen{$_}++ , @arr;
    return @unique;
}

sub retrieve_records {
    my @arr = @_[0..@_-3];
    my $col = $_[-2];
    my $table = $_[-1];
    my @record = ();
    my $rec;
    my $sth;
    if ($table eq "snp") {
        $sth = $connxn->prepare("select rsID from $table where $col = ?");
    }
    else {
        if ($table eq "studied_by") {
            $sth = $connxn->prepare("select Record from $table where $col = ?");
        }
    }
    foreach my $el (@arr) {
        if ($table eq "reference") {
            $sth = $connxn->prepare("select Record from studied_by where
ref_No IN (select PMID from $table where $col like '%$el%')");
            $sth->execute();
        }
        else {
            $sth->execute($el);
        }
        $sth->bind_columns(undef, \$rec);
        while($sth->fetch()) {
            push(@record, $rec);
        }
    }
    return @record;
}


sub construct_arr {
    my %records = %{shift()};
    my %seen;
    my @final;
    while (my ($key, $vals) = each %records) {
        $seen{$_}{$key} = 1 for @$vals;
    }
```

```
	push(@final, $_) for grep { keys %{$seen{$_}} == keys %records } keys %seen;
	return @final;
}

sub rsid2rec {
	my @rsids = @_;
	my @record = ();
	my $sth;
	my $rec;
	$sth = $connxn->prepare("select Record from studied_by where snp_No = ?");
	foreach my $el (@rsids) {
		$sth->execute($el);
		$sth->bind_columns(undef, \$rec);
		while($sth->fetch()) {
			push(@record, $rec);
		}
	}
	return @record;
}

sub final_records {
	(my $recs, my $recs2) = @_;
	 my @intersection;
	 if (!@$recs) {
		 @intersection = @$recs2;
	 }
	 elsif (!@$recs2) {
```

```
                @intersection = @$recs;
        }
        else {
                my %count = ();
                foreach my $element (@$recs, @$recs2) {
                        $count{$element}++;
                }
                foreach my $element (keys %count) {
                        push(@intersection, $element) unless ($count{$element} <= 1);
                }
        }
        return @intersection;
}

sub gen_queries {
        my @rows1 = ();

        if( $FORM{Disease_att} ){
                direct("Disease");
                push(@rows1, "Disease Name");
        }

        if( $FORM{Disease_class_att} ){
                indirect("Class", "disease", "Disease_Name", "Disease");
                push(@rows1, "Disease Class");
        }
```

```
if( $FORM{rsID_att} ){
        indirect("rsID", "snp", "rsID", "snp_No");
        push(@rows1, "rsID");
}
if( $FORM{PMID_att} ){
        direct("ref_No");
        push(@rows1, "PMID");
}
if( $FORM{Ethnicity_att} ){
        indirect("Population_Ethnicity", "reference", "PMID", "ref_No");
        push(@rows1, "Ethnicity");
}
if( $FORM{Geography_att} ){
        indirect("Geographic_location", "reference", "PMID", "ref_No");
        push(@rows1, "Geographic location");
}
if( $FORM{p_value_att} ){
        direct("`p-value`");
        push(@rows1, "p-value");
}
if( $FORM{Odds_ratio_att} ){
        direct("`Odds-ratio`");
        push(@rows1, "Odds-ratio");
}
if( $FORM{Gene_att} ){
        indirect("Gene", "snp", "rsID", "snp_No");
        push(@rows1, "Gene");
```

```
}
if( $FORM{Chromosome_att} ){
        indirect("Chr", "snp", "rsID", "snp_No");
        push(@rows1, "Chromosome");
}
if( $FORM{Ref_contig_att} ){
        indirect2("contig", "contigs", "snpid");
        push(@rows1, "Ref_contig");
}
if( $FORM{Ref_transcript_att} ){
        indirect2("Ref_mRNA", "transcripts", "sid");
        push(@rows1, "Ref_transcript");
}
if( $FORM{Ref_protein_att} ){
        indirect2("Ref_protein", "transcripts", "sid");
        push(@rows1, "Ref_protein");
}
if( $FORM{SNP_Type_att} ){
        indirect("Type", "snp", "rsID", "snp_No");
        push(@rows1, "SNP Type");
}
if( $FORM{Location_SNP_att} ){
        indirect("Location", "snp", "rsID", "snp_No");
        push(@rows1, "SNP Location");
}
if( $FORM{Codon_change_att} ){
        indirect2("Old_Nucleotide_Codon", "changes", "snp");
```

```
                indirect2("New_Nucleotide_Codon", "changes", "snp");
                push(@rows1, "Old Nucleotide/Codon");
                push(@rows1, "New Nucleotide/Codon");
        }
        if( $FORM{AA_change_att} ){
                indirect2("Old_Amino_Acid", "changes", "snp");
                indirect2("New_Amino_Acid", "changes", "snp");
                push(@rows1, "Old Amino Acid");
                push(@rows1, "New Amino Acid");
        }
        if( $FORM{Position_att} ){
                indirect("Position", "changes", "snp", "rsID");
                push(@rows1, "SNP Position");
        }
        push(@rows, { CELLS => [ map { CELL => $_ }, @rows1 ] });
}

sub direct {
        my $att = shift;
        my $query1 = "SELECT $att FROM studied_by WHERE Record = ?";
        push(@queries, $query1);
}

sub indirect {
        (my $att1, my $table1, my $common, my $att2) = @_;
        my $query1 = "SELECT $att1 FROM $table1 WHERE $common IN (SELECT $att2
FROM studied_by WHERE Record = ?)";
```

```
	push(@queries, $query1);
}

sub indirect2 {
	(my $att1, my $table1, my $common) = @_;
	my $query1 = "SELECT $att1 FROM $table1 WHERE $common IN (SELECT rsID FROM snp WHERE rsID IN (SELECT snp_No FROM studied_by WHERE Record = ?))";
	push(@queries, $query1);
}

sub gen_results {
	my @final = @_;
	my %temp = ();
	my @rows1; my @rows2; my @rows3;
	my $rec; my $rec1;
	my $str;
	my $i = 0; my $j = 0;

	foreach my $el (@final) {
		@rows1 = ();
		foreach my $query1 (@queries) {
			@rows3 = ();
			my $query_handle = $connxn->prepare($query1);
			$query_handle->execute($el);
			$query_handle->bind_columns(undef, \$rec);
			if ($query1 =~ /[OldNewPositionRef]/) {
				while($query_handle->fetch()) {
```

```
                    push(@rows3, $rec);
                }
                $rec1 = join('<br>', @rows3);
                push(@rows1, $rec1);
            }
                else { while($query_handle->fetch()) { push(@rows1, $rec); } }
        }
        push(@{ $rows2[$j] }, @rows1);
        $j++;
    }
    my @sorted = sort { $a->[0] cmp $b->[0] } @rows2;
    for my $el1 (@sorted) {
        $str = join(',', @$el1);
        if (!exists($temp{$str})) {
            push(@rows, { CELLS => [ map { CELL => $_ }, @$el1 ] });
            $temp{$str} = 1;
            $i++;
        }
    }
    print $query->p("$i results found");
}
sub create_table {
    my $tmpl = HTML::Template->new(filehandle => \*DATA);
    $tmpl->param(ROWS => \@rows);
    print $tmpl->output;
}
```

```
__DATA__

<table border="1">

<TMPL_LOOP ROWS>

<tr>

<TMPL_LOOP CELLS>

<td><TMPL_VAR CELL></td>

</TMPL_LOOP>

</tr>

</TMPL_LOOP>

</table>
```

APPENDIX X

The SQL create script exported from the MySQL workbench after creation of the database.

```
SET @OLD_UNIQUE_CHECKS=@@UNIQUE_CHECKS, UNIQUE_CHECKS=0;

SET @OLD_FOREIGN_KEY_CHECKS=@@FOREIGN_KEY_CHECKS,
FOREIGN_KEY_CHECKS=0;

SET @OLD_SQL_MODE=@@SQL_MODE, SQL_MODE='TRADITIONAL';


CREATE SCHEMA IF NOT EXISTS `dbaard` DEFAULT CHARACTER SET latin1 ;

USE `dbaard` ;


-- -----------------------------------------------------

-- Table `dbaard`.`snp`

-- -----------------------------------------------------

CREATE  TABLE IF NOT EXISTS `dbaard`.`snp` (

  `rsID` VARCHAR(11) NOT NULL ,

  `Gene` VARCHAR(20) NOT NULL DEFAULT 'N/A' ,

  `Type` VARCHAR(10) NOT NULL DEFAULT 'N/A' ,

  `Location` VARCHAR(10) NOT NULL DEFAULT 'N/A' ,

  `Chr` VARCHAR(2) NOT NULL DEFAULT 'NA' ,

  PRIMARY KEY (`rsID`, `Type`) ,

  INDEX `rsID` (`rsID` ASC) )

ENGINE = InnoDB

DEFAULT CHARACTER SET = latin1;
```

```
-- -----------------------------------------------------
-- Table `dbaard`.`changes`
-- -----------------------------------------------------
CREATE  TABLE IF NOT EXISTS `dbaard`.`changes` (
  `snp` VARCHAR(11) NOT NULL DEFAULT 'N/A' ,
  `Change#` INT(11) NOT NULL ,
  `Old_Nucleotide_Codon` VARCHAR(3) NOT NULL DEFAULT 'N/A' ,
  `New_Nucleotide_Codon` VARCHAR(3) NOT NULL DEFAULT 'N/A' ,
  `Old_Amino_Acid` CHAR(1) NULL DEFAULT '?' ,
  `New_Amino_Acid` CHAR(1) NULL DEFAULT '?' ,
  `Position` INT(11) NULL DEFAULT '0' ,
  PRIMARY KEY (`snp`, `Change#`) ,
  CONSTRAINT `snp#`
    FOREIGN KEY (`snp` )
    REFERENCES `dbaard`.`snp` (`rsID` )
    ON DELETE CASCADE
    ON UPDATE CASCADE)
ENGINE = InnoDB
DEFAULT CHARACTER SET = latin1;



-- -----------------------------------------------------
-- Table `dbaard`.`contigs`
-- -----------------------------------------------------
CREATE  TABLE IF NOT EXISTS `dbaard`.`contigs` (
  `snpid` VARCHAR(11) NOT NULL ,
  `contig` VARCHAR(12) NOT NULL ,
  PRIMARY KEY (`snpid`, `contig`) ,
  INDEX `snpid` (`snpid` ASC) ,
  CONSTRAINT `snpid`
    FOREIGN KEY (`snpid` )
    REFERENCES `dbaard`.`snp` (`rsID` )
    ON DELETE CASCADE
    ON UPDATE CASCADE)
ENGINE = InnoDB
DEFAULT CHARACTER SET = latin1;



-- -----------------------------------------------------
-- Table `dbaard`.`studied_by`
-- -----------------------------------------------------
CREATE  TABLE IF NOT EXISTS `dbaard`.`studied_by` (
  `Record` INT(11) NOT NULL AUTO_INCREMENT ,
  `snp_No` VARCHAR(11) NOT NULL ,
  `ref_No` INT(11) NOT NULL DEFAULT '0' ,
  `Disease` VARCHAR(30) NOT NULL DEFAULT 'N/A' ,
  `p-value` VARCHAR(9) NULL DEFAULT 'N/A' ,
  `Odds-ratio` VARCHAR(4) NULL DEFAULT 'N/A' ,
  `Annotator` VARCHAR(7) NOT NULL DEFAULT 'N/A' ,
  `Remarks` TEXT NULL DEFAULT NULL ,
  PRIMARY KEY (`Record`, `snp_No`) ,
  INDEX `snp_No` (`snp_No` ASC) ,
  INDEX `ref_No` (`ref_No` ASC) ,
  INDEX `Dis` (`Disease` ASC) ,
  CONSTRAINT `snp_No`
    FOREIGN KEY (`snp_No` )
    REFERENCES `dbaard`.`snp` (`rsID` )
    ON DELETE CASCADE
    ON UPDATE CASCADE)
ENGINE = InnoDB
AUTO_INCREMENT = 1733
DEFAULT CHARACTER SET = latin1;



-- -----------------------------------------------------
-- Table `dbaard`.`disease`
-- -----------------------------------------------------
CREATE  TABLE IF NOT EXISTS `dbaard`.`disease` (
  `Disease_Name` VARCHAR(30) NOT NULL ,
  `Class` VARCHAR(25) NOT NULL ,
  PRIMARY KEY (`Disease_Name`) ,
  CONSTRAINT `Disease`
    FOREIGN KEY (`Disease_Name` )
    REFERENCES `dbaard`.`studied_by` (`Disease` )
    ON DELETE CASCADE
    ON UPDATE CASCADE)
ENGINE = InnoDB
DEFAULT CHARACTER SET = latin1;
```

```
-- -----------------------------------------------------
-- Table `dbaard`.`reference`
-- -----------------------------------------------------
CREATE  TABLE IF NOT EXISTS `dbaard`.`reference` (
  `PMID` INT(11) NOT NULL DEFAULT '0' ,
  `Population_Ethnicity` VARCHAR(200) NOT NULL DEFAULT 'N/A' ,
  `Geographic_location` VARCHAR(200) NULL DEFAULT 'N/A' ,
  PRIMARY KEY (`PMID`) ,
  CONSTRAINT `PMID`
    FOREIGN KEY (`PMID` )
    REFERENCES `dbaard`.`studied_by` (`ref_No` )
    ON DELETE CASCADE
    ON UPDATE CASCADE)
ENGINE = InnoDB
DEFAULT CHARACTER SET = latin1;



-- -----------------------------------------------------
-- Table `dbaard`.`transcripts`
-- -----------------------------------------------------
CREATE  TABLE IF NOT EXISTS `dbaard`.`transcripts` (
  `sid` VARCHAR(11) NOT NULL ,
  `Variant_No` INT(11) NOT NULL ,
  `Ref_mRNA` VARCHAR(15) NOT NULL ,
  `Ref_protein` VARCHAR(15) NULL DEFAULT 'N/A' ,
  PRIMARY KEY (`sid`, `Variant_No`) ,
  INDEX `sid` (`sid` ASC) ,
  CONSTRAINT `sid`
    FOREIGN KEY (`sid` )
    REFERENCES `dbaard`.`snp` (`rsID` )
    ON DELETE CASCADE
    ON UPDATE CASCADE)
ENGINE = InnoDB
DEFAULT CHARACTER SET = latin1;



SET SQL_MODE=@OLD_SQL_MODE;
SET FOREIGN_KEY_CHECKS=@OLD_FOREIGN_KEY_CHECKS;
SET UNIQUE_CHECKS=@OLD_UNIQUE_CHECKS;
```

APPENDIX XI

Coloring key used for nodes in the networks

Color	Class of Disease
🟪	Bone
🟥	Cancer
🟪	Cardiovascular
🟧	Connective tissue
🟥	Dermatological
🟩	ENT
🟫	Gastrointestinal
🟪	Hematological
🟨	Immunological
🟦	Metabolic
🟥	Muscular
🟦	Neurological
🟩	Nutritional
🟦	Ophthalmological
🟨	Psychiatric
🟥	Respiratory
🟩	Syndromes
🟧	Mixed