APPROVAL SHEET

Title of Thesis: Clinical-Genomic Analysis for Disease Prediction

Name of Candidate: Darshana Dalvi
M.S. in Computer Science, 2011

Thesis and Abstract Approved: Yelena Yesha
Dr. Yelena Yesha
Professor
Department of Computer Science and
Electrical Engineering

Date Approved: 07/08/11

Curriculum Vitae

Name: Darshana Dalvi.

Permanent Address: 4751 Drayton Green, Baltimore, MD - 21227.

Degree and date to be conferred: Masters in Computer Science, July 2011.

Date of Birth: 12th March, 1986.

Place of Birth: Mumbai, India.

Secondary education: Fergusson College, Pune, India.

Collegiate institutions attended:

University of Maryland Baltimore County, M. S. Computer Science, 2011.
Cummins College of Engineering, Pune, B. E. Computer Engineering, 2007.

Major: Computer Science.

Professional Publications:

Image Classification of vascular smooth muscle cells, IHI '10 Proceedings of the 1st ACM International Health Informatics Symposium.

Professional positions held:

Systems Engineer, Siemens Information Systems Limited, Pune, MH, India. (July 2007 – June 2009).
Summer Intern, Akamai Technologies Inc., Cambridge, MA, USA. (June 2010 – Aug 2010).

ABSTRACT

Title of Thesis: Clinical-Genomic Analysis for Disease Prediction

Darshana Dalvi, Master of Science, 2011

Thesis directed by: Dr. Yelena Yesha, Professor,
Department of Computer Science and
Electrical Engineering.

Recent advances in genomic research have generated vast amounts of information that can help identify individuals who differ in their susceptibility to a particular disease or response to a specific treatment. This information may offer solutions for the treatment of complex chronic diseases that are influenced by a wide array of factors. This vast amount of information brings critical challenges in applying advanced technology to synthesize clinical-genomic patient data. Synthesizing this information is necessary to derive the knowledge that would empower physicians to provide personalized care with the best possible therapeutic interventions.

We used statistical methods and data mining approaches to understand clinical-genomic risk factors that differentiate Type II Diabetes cases from healthy controls. We investigated whether inclusion of genomic risk factors in conjunction with clinical information improves classification accuracy. We also demonstrate how a biased and an unbiased method for selection of risk associated single nucleotide polymorphisms (SNPs) effect clustering along with clinical information. We determined the optimal method based on its clustering performance.

Keywords: clinical-genomic risk, SNPs, classification, clustering.

Clinical-Genomic Analysis for Disease Prediction

by

Darshana Dalvi

Thesis submitted to the Faculty of the Graduate School
of the University of Maryland, in partial fulfillment
of the requirements for the degree of
Master of Science
2011

Dedicated to Aai and Pappa

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my graduate advisor Dr. Yelana Yesha. I thank her for her constant support, guidance and her continued belief in me throughout this thesis work. I would like to thank Dr. Michael Grasso for all the valuable ideas, long discussions and guidance which were vital in bringing this work to completion.

Thanks to Dr. Yaacov Yesha and Dr. Milton Halem for graciously agreeing to be on my thesis committee. A special note of thanks to Dr. Anupam Joshi, Dr. Aryya Gangopadhyay and Jessia Nadler for valuable inputs and timely help. I would also like to thank Dr. Eddie Karnieli from Technion – Israel Institute of Technology for his initial guidance in understanding the dataset and designing the problem.

All my friends at MC^2 lab have been extremely supportive in building an atmosphere conducive to research. It has been a great joy working with them throughout the year.

Table of Contents

LIST OF TABLES

LIST OF FIGURES

Chapter 1

INTRODUCTION

In this chapter we present an introduction to personalized medicine. We will discuss the need of clinical-genomic analysis in disease diagnosis and provide a formal thesis definition.

1.1 Personalized Medicine

The Human Genome Project[1]; a 13 year multi-national project has laid the groundwork for our understanding of the roles of genes in normal human development. In addition, importance of realizing the genetic base of a disease has now become evident [1]. The researchers and the medical practitioners are now focusing on studying the genome sequences that are responsible for pathogenesis of common diseases and trying to bridge a gap between clinical and genetic factors [2].

Increasing availability of genetic tests, over 1500 till date and the emergence of privately held organizations such as 23andMe[2] allows individuals to understand their own genomic profile and monitor the changes over the time.

The field that is emerging due to these advances in medicine today is the use of Personalized Medicine in patient care. It involves the systematic use of a patient's

[1] http://www.ornl.gov/sci/techresources/Human_Genome/home.shtml
[2] https://www.23andme.com/

medical and related information to optimize therapeutic care and diagnosis [3]. More specifically, information about a patient’s demographic, clinical, genomic and metabolic characteristics is used to tailor medical care to meet specific needs and provide customized healthcare.

1.2 Motivation: The need of clinical-genomic analysis

A precise prediction of the disease outcome is of paramount importance for accurate diagnosis and treatment. For oligogenic disorders that are determined by only a small set of genes, the assessment of risk factors is fairly straightforward. However, for complex chronic diseases where there are wide array of factors that are responsible for the disease development, the risk assessment is a complicated process. As explained by Khoury & Yang [4], the diversity in the genes due to origins or demographics conditions as well as gene-environment interactions further adds to the complexity of understanding controversial etiology of complex diseases. Genomic studies generate vast amount of data in the form of gene expressions and Single Nucleotide Polymorphisms (SNPs)[3]. A human genome is estimated to contain 10 million SNPs out of which ~ 30,000 SNPs have significant genetic variations [5]. Primarily, these SNPs and their patterns in a human genome have been used to differentiate individuals based on skin color, hair, response to drugs, etc. However a novel approach of making use of SNPs as prognostic information to determine disease susceptibility in individuals is gaining attention in genetic research.

Some of the challenging tasks in this research are described below.

[3] http://www.ncbi.nlm.nih.gov/About/primer/snps.html

- The selection of appropriate clinical or non-genetic risk factors along with genetic risk factors that would help to discriminate between individuals based on susceptibility to the disease is essential.
- Various data mining techniques are used for assessing human genome as a whole such as the sequence comparison, gene expression analysis, gene finding, etc. [6]. However there is a need of new approaches to identify and extract risk associated SNPs out of the gene expressions and analyze their relation with the clinical and environmental risk factors in predisposition to the disease.

We believe that the use of advanced data analysis and data mining techniques will help us to derive clinically significant knowledge for studying development of complex chronic diseases using case-control population datasets. The derived knowledge and population based analysis of the disease would significantly benefit the physicians for accurate diagnosis.

1.3 Thesis Contribution

Our focus is on Type II Diabetes[4], a chronic disease that comprises 90% of the diabetic people in the world. As stated by Thompson [7], Type II Diabetes is also highly prevalent in United States. We intend to contribute to the efforts of diabetes prevention and eradication by deriving useful knowledge from a case-control based study dataset of white American population. We apply several data analysis techniques along with knowledge driven data mining methods to find hidden patterns and trends in the dataset and select the methods that give accurate results.

[4] http://www.diabetes.org/diabetes-basics/type-2/

Initially we perform literature study to identify the Type II Diabetes risk associated SNPs found in the population under consideration as well those which are independent of race or ethnicity. From these, we focus on those SNPs with considerably higher frequency of occurrence in cases than in controls. We define genomic risk of an individual in terms of associations with the selected SNPs.

We then assess correlations amongst clinical as well as genomic risk factors and compare them based on their severity towards the disease. We show how the use of genomic risk factors along with other clinical data would help in predictive analysis of the disease as well as while grouping individuals based on their similar characteristics. We evaluate the accuracy of the results using several measures to show the merit of our approaches.

Chapter 2

BACKGROUND AND RELATED WORK

In this chapter, we will explain few background concepts and genomic details that would help to understand the techniques and algorithms mentioned in the subsequent chapters. We will also talk about some existing researches.

2.1 Single Nucleotide Polymorphisms (SNPs)

Single Nucleotide Polymorphism is a single change that can occur in a person's DNA [8]. A change occurs when a single nucleotide for example A is replaced by one of the other nucleotides C, G or T.

For example, consider a DNA segment AACGTTA when altered gives new segment as AATGTTA, where C in the first segment is replaced by T in the second segment.

2.1.1 Allele

Each member of a pair of chromosomes in a human or any multicellular organism is termed as an Allele[5]. These chromosomes are referred as heterozygous if two different alleles are present for the same trait or homozygous if identical alleles are present for the same trait.

[5] http://www.ncbi.nlm.nih.gov/About/primer/snps.html

Consider following example,

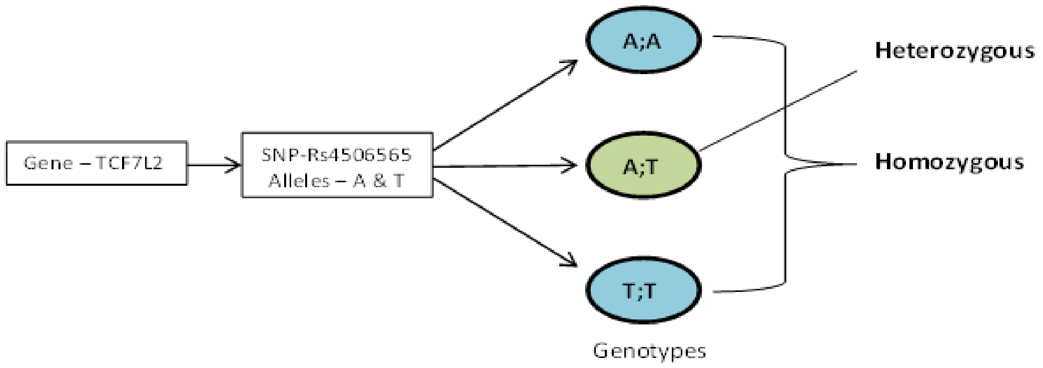


Figure 2.1. Example of different genotypes possible for a SNP in a TCF7L2 gene

2.1.2 SNPs and diseases diagnosis

SNPs are not responsible for causing the disease but can help to determine likelihood of someone will have the disorder. Recent studies [9] suggest that SNPs can lead to effective prognosis of the disease as well as can affect individual's susceptibility to a disease. Genome Wide Association Studies (GWAS)[6], a study conducted at large to find genes who have susceptibility towards a disorder has suggested common variants in certain genes such as TCF7L2, FTO, TSPAN8, CDAKL1, MCR4, etc. that are susceptible to Type II Diabetes. Out of these TCF7L2 gene is reproducibly in existence in various ethnic groups [10]. McCarthy [11] focuses on importance of combinations of SNPs instead of a single SNP in the development of Type II Diabetes. This drives our approach of selecting a set of risk associated SNPs to determine genomic risk of an individual.

[6] http://gwas.nih.gov/

2.1.3 Type II Diabetes associated genes

Some of the Type II Diabetes genes[7] are explained below.

i. **TCF7L2 (Transcription factor 7-like 2):** Variants of this gene have been associated with Type II Diabetes in multiple ethnic groups and are known as strongest genetic risk factors. TCF7L2 polymorphisms are associated with impaired insulin secretion and glucose tolerance.

ii. **FTO:** Variants of this gene have appeared to be associated with fat mass and obesity in large populations which predispose to diabetes through an effect on body mass index.

iii. **MC4R (Melanocortin receptor 4):** Mutations of this gene have been associated with inherited human obesity mostly in an autosomal dominant[8] inheritance pattern.

iv. **CDKAL1 (CDK5 regulatory subunit associated protein 1-like 1):** Variants of these have been reported to cause Type II Diabetes due to impaired pancreatic β-cell function and decreased insulin sensitivity

v. **TSPAN8 (Tetraspanin-8):** This gene plays a key role in cell development, activation, growth and motility. Variants of this gene have been found associated with Type II Diabetes, metabolic syndrome and other disorders such as color cancer, schizophrenia.

[7] http://www.ncbi.nlm.nih.gov/gene

[8] http://www.nlm.nih.gov/medlineplus/ency/article/002049.htm

2.2 Overview of Data Mining

Data mining is a field of computer science that aims to apply different techniques to transform abundant data into intelligence [12].

There are two main approaches in data mining.

2.2.1 Supervised Learning (Classification)

Classification is a supervised learning approach which generates an inferred function using set of input data that should predict a correct output value for any valid input provided.

Bayesian Network [13] is one of the methods of classification which uses directed acyclic model to represent conditional dependencies between a set of random variables. Edges represent conditional dependencies and nodes which are not connected represent conditional independence.

C 4.5 [14] is another method of classification which uses decision tree model that makes use of information gain to effectively split samples into subsets enriched with one of the classes. In this case, leaves represent the classification and the branches represent a set of features which leads to the classification.

2.2.2 Unsupervised Learning (Clustering)

Clustering is an unsupervised learning approach where there are no explicit outputs associated with the inputs. The main aim is to find hidden patterns in the input data that represent statistical information about the input.

Partition clustering is a method of clustering based on the concept of division of a set X into non-overlapping and non-empty parts. K-means clustering [15] belongs to this category in which each point is assigned to a cluster whose centroid is nearest to the point.

Spectral clustering is another method based on spectral graph theory [16] which uses graph-based representation for clustering large datasets. Here the basic aim is to analyze spectrum of matrix representing a graph. Graphs are useful when we want to extract pairwise information of data points such as similarity or distances. In a graph, data points represent nodes and each edge between two nodes has a weight associated representing similarity or distance between them. Spectral clustering performs graph partitioning based on eigenvalues and eigenvectors of the similarity matrix which is nothing but the spectrum or global structure of the adjacency matrix.

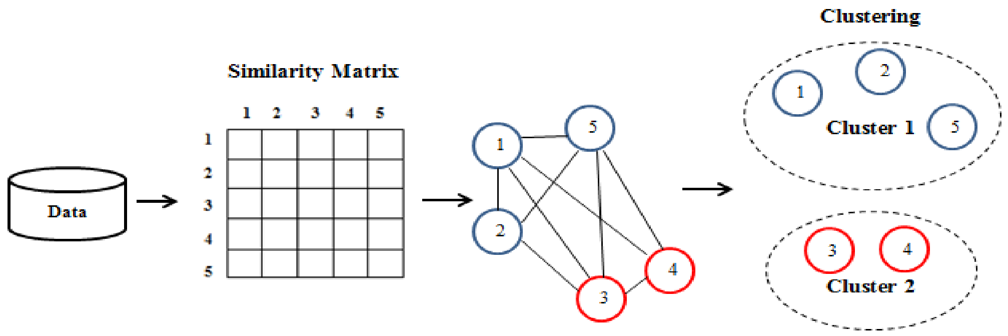


Figure 2.2. Overview of Spectral Graph Clustering

Spectral clustering algorithm used [17]:

I. Preprocessing:

1. Calculate distance matrix of the input dataset (M).

2. Calculate similarity matrix also called as weight matrix (W). Distances and similarities are inverse of each other.

   Therefore, W = exponential (-M).

3. Calculate degree matrix (D). Degree of a vertex is the sum of edge points at that vertex.

$$D(i,j) = \begin{cases} \text{degree}(V_i) & \text{if } i=j \\ 0 & \text{otherwise} \end{cases}$$

4. Calculate unnormalized laplacian matrix (L) = D – W.

II. Decomposition:

5. Compute the eigenvectors of L as V_1…..V_n. Graph laplacian and its eigenvalues, eigenvectors are used to describe properties of a graph.

III. Clustering multiple eigenvectors:

6. Let U be a matrix containing first k eigenvectors of V as U_1…U_k. First k eigenvectors here refer to the eigenvectors corresponding to the k smallest eigenvalues.

7. For i=1…n, let R_i represent i^{th} row of U.

8. Applying k-means:

   Cluster the rows R_i to R_n of U_k into k clusters, C_1…C_k.

This method is useful to obtain well-separated clusters where associations between similar points are amplified and those between dissimilar points are diminished.

2.3 Clinical-Genomic Data Description

The raw data is broadly of two types, clinical and genomic. Our aim in such cases is to do away with individual level information and preserve the disease level context of the data. As an example for Type II Diabetes, we plan to store the clinical data with different attributes available across given patients and combine it with risk associated SNPs from the genotyped data.

2.3.1 Clinical Data

This data is generated when a patient visits any clinic which could be a routine visit to a doctor to Emergency Room visit. All the vital parameters such as age, sex, height, color, temperature and weight etc. are recorded besides the other condition specific parameters. These are certain clinical lab tests conducted on the patient.

2.3.2 Genomic Data

Genomic data is mainly available at the nucleotide level. Based on the study type conducted the goal is to ascertain in the lab the nucleotide (AGCT) makeup of the patient at the given location which is known to have been disease causing.

2.3.3 dbGAP

dbGAP[9] is a database of genotypes and phenotypes that provides unprecedented access to the large-scale genetic and phenotypic datasets required for GWAS [10]designs, including public access to study documents linked to summary data on specific phenotype variables, statistical overviews of the genetic information, position of published associations on the genome, and authorized access to individual-level data [18]. dbGaP accommodates studies of varying design. It

[9] http://www.ncbi.nlm.nih.gov/gap

[10] http://gwas.nih.gov/

contains four basic types of data: (i) study documentation, including study descriptions, protocol documents, and data collection instruments, such as questionnaires (ii) phenotypic data for each variable assessed, both at an individual level and in summary form (iii) genetic data, including study subjects' individual genotypes, pedigree information, fine mapping results and resequencing traces and (iv) statistical results, including association and linkage analyses, when available.

Data access structure:

- Study protocols and summary phenotype and genotype data are available to the public without restrictions on use.
- Access to individual-level data requires preauthorization from sponsoring NIH (National Institutes of Health) programs. Use of the data is limited to the approved research activities, and must follow the basic principles set forth in the NIH policy for GWAS.

2.3.4 GENEVA Diabetes Case-Control Study

The dataset consists of genotype files as well as dietary and lifestyle information of cohorts of nurses and health professionals. The dataset is a part of Gene Environment Association Studies Initiative (GEI)[11] undertaken with the goal of identifying novel genetic factors that contribute to Type II Diabetes.

[11] http://www.genome.gov/19518663

2.4 Related Work

Cancer prognostics and therapeutics are among the first major research contributors in genomic personalized medicines. Researchers at the Unit of Medical Technology and Intelligent Information Systems, Dept. of Computer Science, University of Illinois [19] proposed a framework that integrates Single Nucleotide Polymorphism related to colon cancer with the clinical data and profiles patients with similar clinical genomic characteristics for decision support, diagnosis and treatment of colon cancer.

Interdisciplinary Study Group on Gene Environment Interaction and Breast Cancer in Germany [20] aimed to investigate gene to environment interactions by exploring several ways to find similarity between two objects based on their clinical attributes and gene locii. Such studies opened up an approach of observing correlations amongst genetic variations and the clinical risk factors.

2.4.1 Data Mining in Personalized Medicine

Most of the current data mining methodologies applied in genetic research focus on gene expression profiles of individuals for classification and correlation with the traditional clinical outcomes. This creates a challenge to select appropriate data analysis methods and data mining algorithms due to the complexity and volume of genomic data [21]. Ban and Heo [22] suggests use of Multi Dimensionality Reduction (MDR) techniques to select a subset of risk associated SNPs followed by classification using Support Vector Machine that optimally predicts the risk. Gregory F. Cooper [23] demonstrated use of Bayesian method and several machine learning methods on a genome-wide dataset in predicting outcomes of Alzheimer’s disease. A tree-based classification framework combining gene signatures and clinical factors has been used in breast cancer prediction [24].

Finding patterns in clinical-genomic data has been an open area of research for data mining experts due to heterogeneous nature of the healthcare data. K-modes clustering method proposed at an International Conference on Data Mining [25] gave us an exposure on effective use of SNPs for finding their association with the disease. Several data mining approaches have been applied to assess Type II Diabetes SNP variants along with its clinical risk factors. The Framingham offspring study [26] proved that 40-SNP weighted genomic risk score along with clinical risk factors improved diabetes prediction in younger people < 50 years of age.

Chapter 3

SYSTEM OVERVIEW AND METHODOLOGY

In this chapter, we explain the high level overview and different components that influence the system. We then describe the dataset and different tools we used for data mining purposes.

3.1 System Outline

Following figure shows overview of the system.

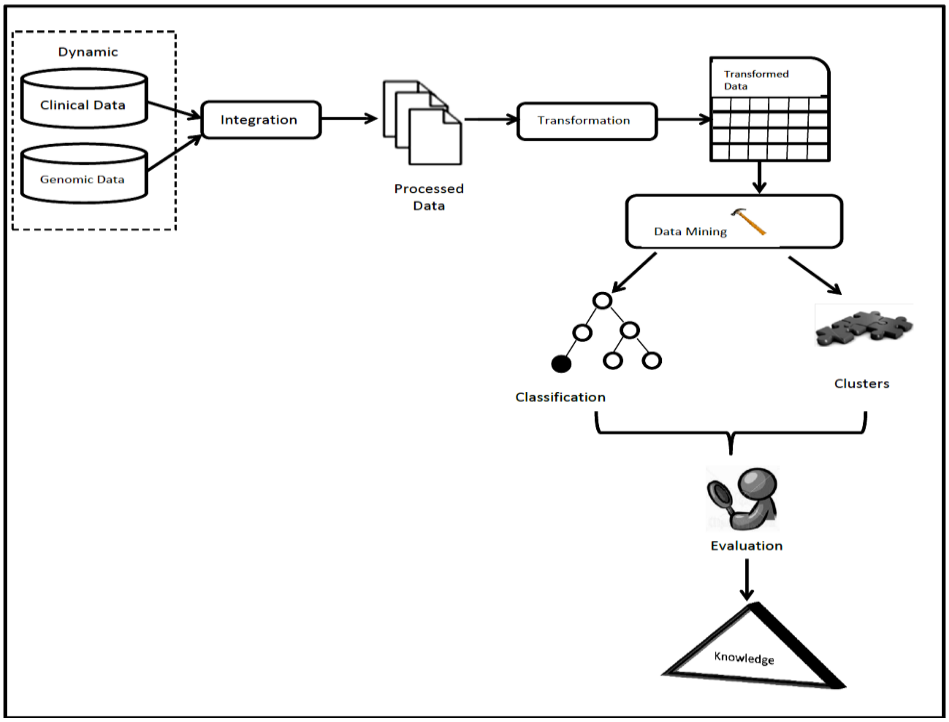


Figure 3.1. Overview of steps followed in knowledge discovery

Description:

I. Integration layer: This is the first step to integrate the heterogeneous clinical and genomic data downloaded from dbGAP. While phenotype[12] data is present in comma separated file format, the genotype information is present in .CHP Affymetrix[13] files generated during genotype analysis.

II. Data Preprocessing: Before the data is sent to data mining engine such as WEKA or MATLAB, it processed and transformed into a suitable format. Data cleaning and normalization also takes place at this step.

III. Data Mining: In this step, selected classification and clustering algorithms are run on the gathered data.

IV. Validation and Knowledge Discovery: The results obtained after executing data mining algorithms are validated with the help of a physician or a doctor which then become a part of the derived knowledge.

3.2 Dataset Description

Following are some statistics regarding phenotype and genotype information present in the dataset.

3.2.1 Phenotype Data

Number of individuals in the dataset based on race:

Race	Female	Male
White	3303	2436
Asian	17	25
African-American	30	25
Other		50
American-Indian	14	

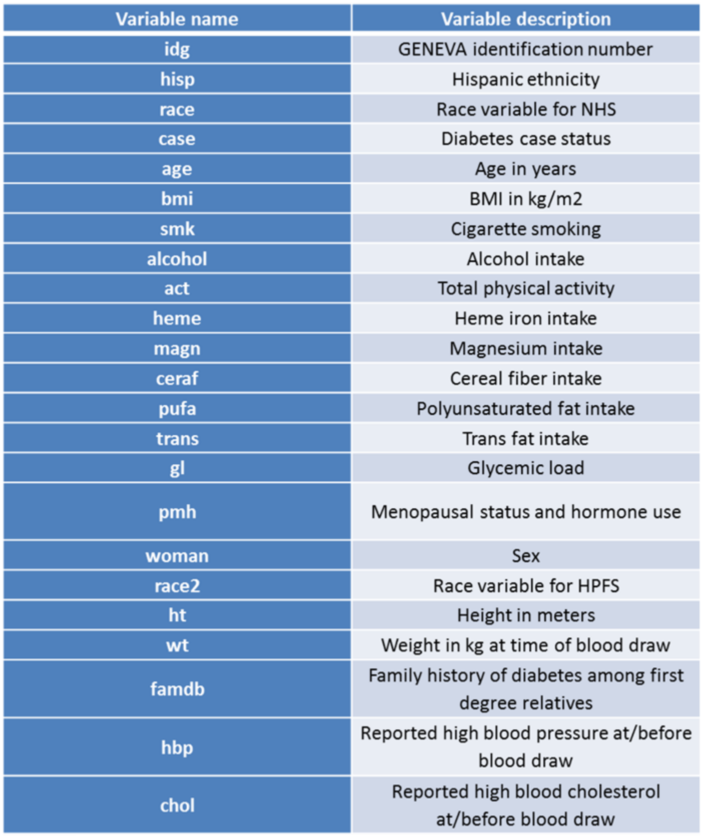
Table3.1. Population division based on race

	Hispanic	Non-Hispanic
Female	37	3327
Male	Unknown	Unknown

Table 3.2. Population division based on hispanicity

[12] http://www.ncbi.nlm.nih.gov/About/primer/genetics_genome.html
[13] http://www.affymetrix.com/support/developer/powertools/changelog/gcos-agcc/chp.html

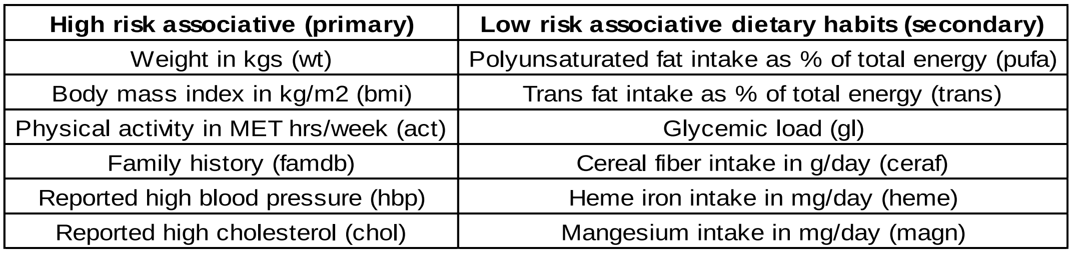
The dataset consists of following set of 23 variables.



Variable name	Variable description
idg	GENEVA identification number
hisp	Hispanic ethnicity
race	Race variable for NHS
case	Diabetes case status
age	Age in years
bmi	BMI in kg/m2
smk	Cigarette smoking
alcohol	Alcohol intake
act	Total physical activity
heme	Heme iron intake
magn	Magnesium intake
ceraf	Cereal fiber intake
pufa	Polyunsaturated fat intake
trans	Trans fat intake
gl	Glycemic load
pmh	Menopausal status and hormone use
woman	Sex
race2	Race variable for HPFS
ht	Height in meters
wt	Weight in kg at time of blood draw
famdb	Family history of diabetes among first degree relatives
hbp	Reported high blood pressure at/before blood draw
chol	Reported high blood cholesterol at/before blood draw

Table 3.3. Dataset Variable description

Out of the available set of variables, we select following set of variables for data mining purposes. We further categorize them depending upon their severity towards diabetes.



High risk associative (primary)	Low risk associative dietary habits (secondary)
Weight in kgs (wt)	Polyunsaturated fat intake as % of total energy (pufa)
Body mass index in kg/m2 (bmi)	Trans fat intake as % of total energy (trans)
Physical activity in MET hrs/week (act)	Glycemic load (gl)
Family history (famdb)	Cereal fiber intake in g/day (ceraf)
Reported high blood pressure (hbp)	Heme iron intake in mg/day (heme)
Reported high cholesterol (chol)	Mangesium intake in mg/day (magn)

Table 3.4. Categorization of clinical risk factors

This way we could differentiate between physical or biological characteristics of an individual from the dietary habits of an individual.

Age-wise distribution of cases and controls: We divide the population into different age groups and observe distribution of cases and controls amongst them.

Age Category	Age in years
1	<45
2	45-50
3	51-55
4	56-60
5	61-65
6	66-70
7	>70

Table 3.5. Different age categories

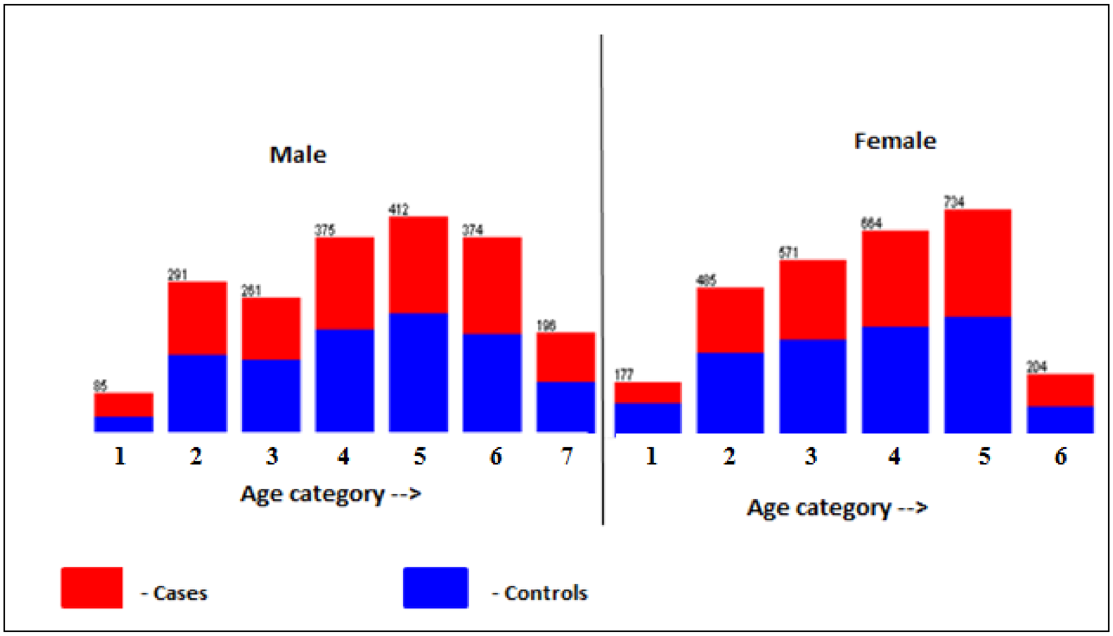


Figure 3.2. Histogram of age cateories vs cases and controls

We observe that since diabetes is a disease that is independent of the age, the distribution of cases and controls is seen overall uniform across different age groups with a marginal increase in cases for age above 55.

Phenotype data vs Risk:

We observe effect of age on some phenotypes such as high blood pressure and cholesterol which are known to vary with the age.

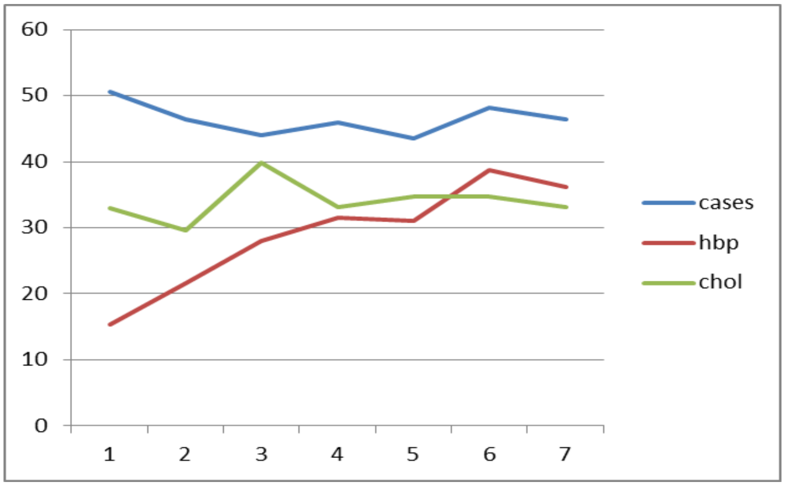


Figure 3.3. Age category vs % of individuals (male dataset)

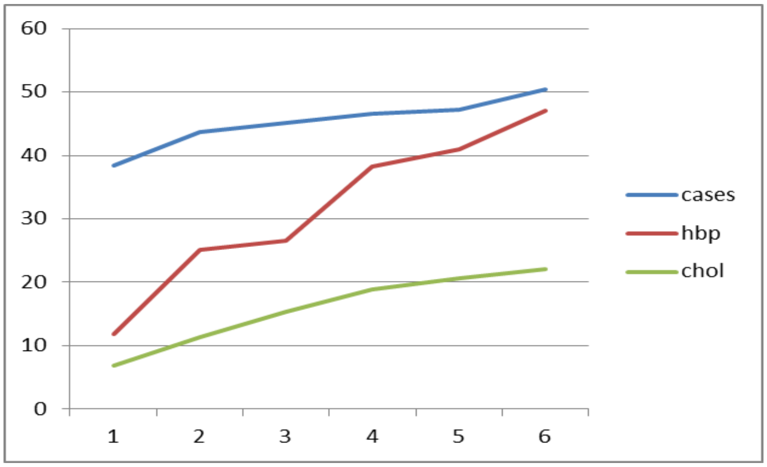


Figure 3.4. Age category vs % of individuals (female dataset)

Here we observe that % of cases doesn't increase much with the increase in age. Whereas in case of the phenotype, % of individuals with high blood pressure tends to increase with the increase in age, more significantly in females. % of individuals with high cholesterol also increases considerably with the increasing age.

3.2.2 Genotype Data

Identification of Type II Diabetes associated SNPs and corresponding risk allele:

To identify which SNPs are associated with Type II Diabetes and which of the alleles has more tendencies towards the risk, we apply following approaches:

1. Using existing researches and literature articles: Over 40 or more SNPs have been reported as associated with Type II Diabetes in the existing research. We used this evidence to identify Type II Diabetes associated SNPs from our dataset of white American population.

2. SNPedia: More Type II Diabetes related SNPs are learnt using SNPedia. For example, for a TCF7L2 gene related SNP - Rs4506565 [A/T], SNPedia gives following information.

Genotypes	Effect
Rs4506565 (A;A)	1.9x increased risk for type-2 diabetes
Rs4506565 (A;T)	1.4x increased risk for type-2 diabetes
Rs4506565 (T;T)	normal

Table 3.6. Genotype vs risk association (Courtesy: SNPedia)

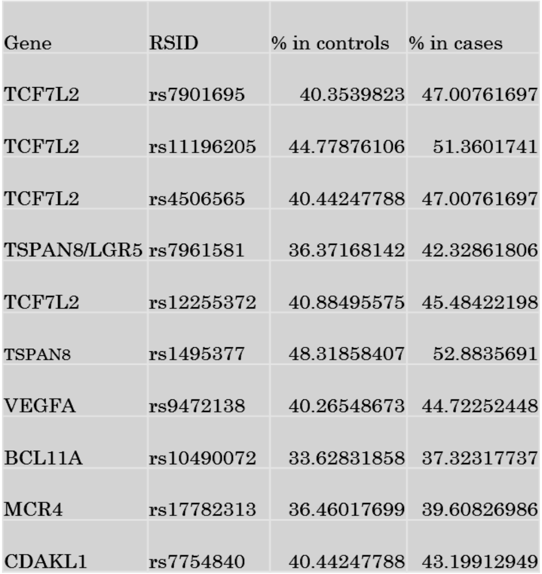
From this, we can infer that ‘A’ allele is at a higher risk than ‘T’.

Using these two approaches, we obtained a set of 20 SNPs that are found prominent in our dataset. Some of the common SNPs observed amongst both male and female population are summarized below.

Gene	**RsId**	**Risk Allele**
TCF7L2	rs12255372	T
TCF7L2	rs4506565	A
TCF7L2	rs7901695	C
TSPAN8	rs7961581	C
TSPAN8	rs1495377	G
MCR4	rs17782313	C
CDAKL1	rs7754840	C
KCNJ11	rs5215	T
FTO	rs9930506	C
FTO	rs8050136	A

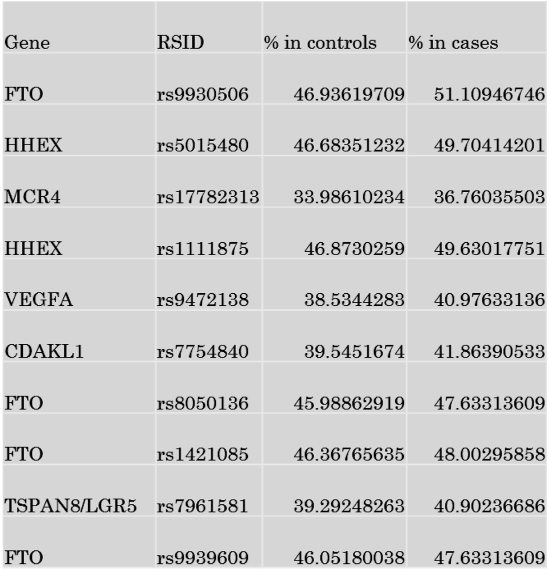
Table 3.7. Type II Diabetes SNPs found common between male and female datasets

Following are sets of Type II Diabetes SNPs which shows higher % of presence of a risk allele in cases than in controls.



Gene	RSID	% in controls	% in cases
TCF7L2	rs7901695	40.3539823	47.00761697
TCF7L2	rs11196205	44.77876106	51.3601741
TCF7L2	rs4506565	40.44247788	47.00761697
TSPAN8/LGR5	rs7961581	36.37168142	42.32861806
TCF7L2	rs12255372	40.88495575	45.48422198
TSPAN8	rs1495377	48.31858407	52.8835691
VEGFA	rs9472138	40.26548673	44.72252448
BCL11A	rs10490072	33.62831858	37.32317737
MCR4	rs17782313	36.46017699	39.60826986
CDAKL1	rs7754840	40.44247788	43.19912949

Male Population



Gene	RSID	% in controls	% in cases
FTO	rs9930506	46.93619709	51.10946746
HHEX	rs5015480	46.68351232	49.70414201
MCR4	rs17782313	33.98610234	36.76035503
HHEX	rs1111875	46.8730259	49.63017751
VEGFA	rs9472138	38.5344283	40.97633136
CDAKL1	rs7754840	39.5451674	41.86390533
FTO	rs8050136	45.98862919	47.63313609
FTO	rs1421085	46.36765635	48.00295858
TSPAN8/LGR5	rs7961581	39.29248263	40.90236686
FTO	rs9939609	46.05180038	47.63313609

Female Population

Table 3.8. Comparison of cases and controls based on% of risk SNPs

Genomic data vs risk:

Presence of a risk allele increases individual's tendency towards the risk. Following graph shows the variation in % of cases with respect to different genetic variants.

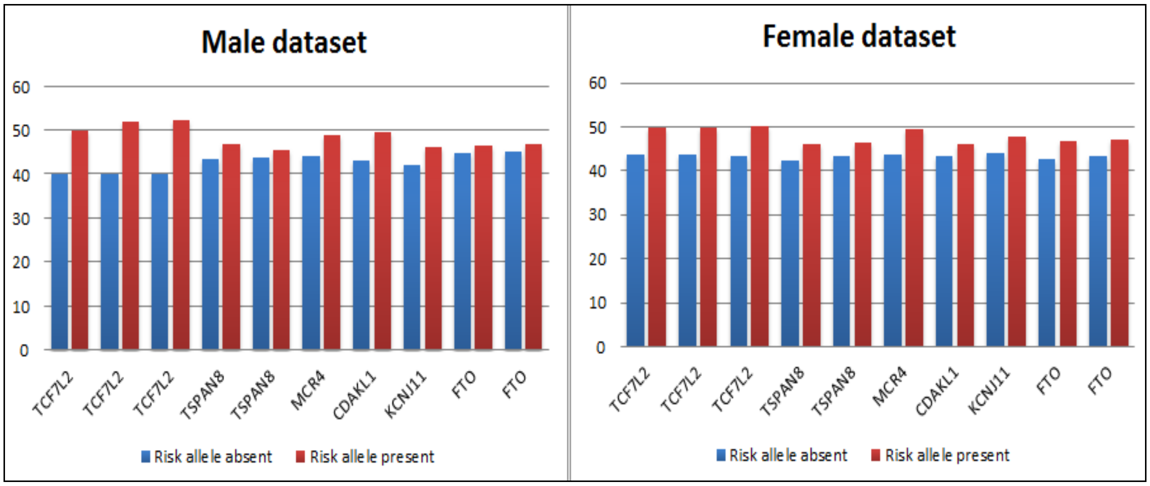
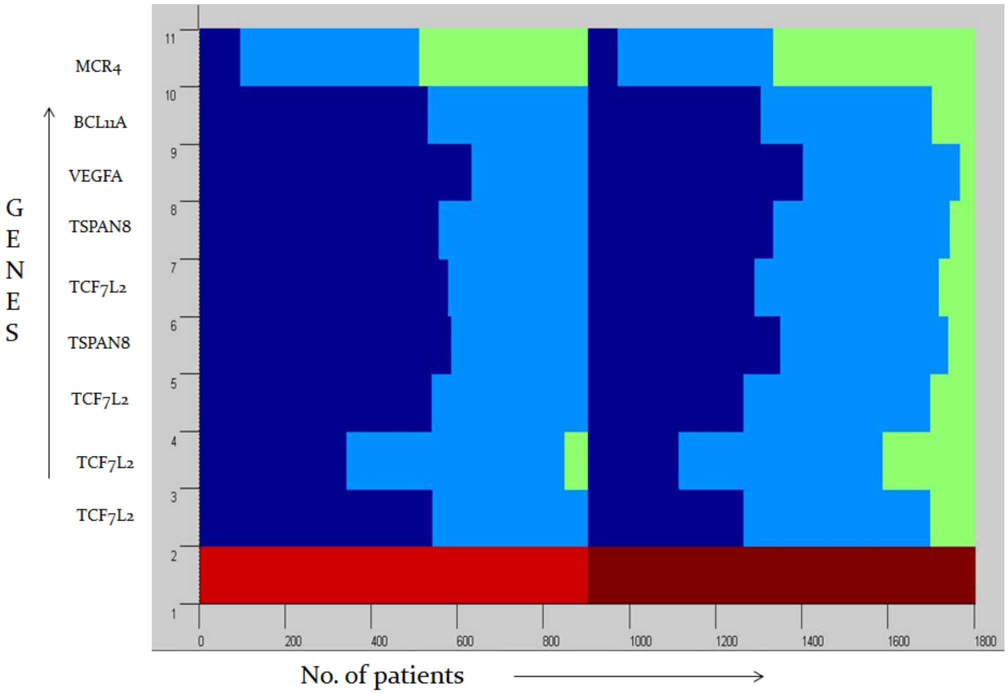


Figure 3.5. Risk associated gene vs % of cases

It was confirmed that the risk allele is present in the majority of the cases. It also confirmed prominence of TCFL2 gene. To further understand risk alleles and their associativity with the cases, we used following methods.

SNP Density Distribution Graph

Consider a graph showing SNPs variations amongst male cases and male controls. The red band at the bottom is 800 controls in the beginning followed by 800 cases. Values on the Y axis after the red band are 10 dominant risk SNPs in males.



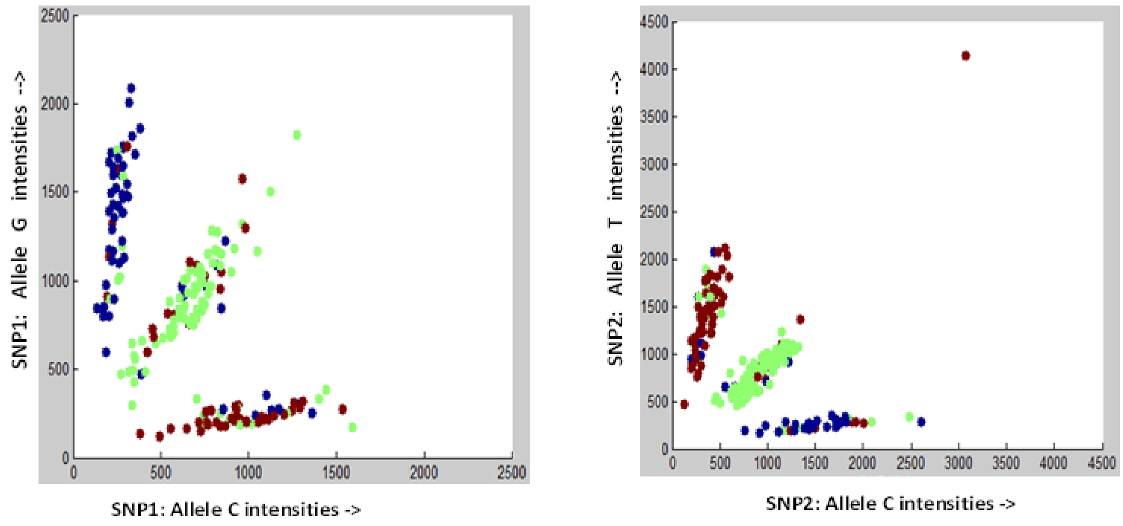
Dark blue band: Absence of risk allele
Light blue band: Presence of risk allele in one chromosome
Green band: Presence of risk alleles at both positions

Figure 3.6. SNP Density distribution graph

Here we observe that the light blue and the green band is spread more in cases than in controls which indicates that Type II Diabetes risk allele association is observed more in cases than in controls. Similarly widespread dark blue band in controls indicate absence of risk allele.

Tri-colored scattered plot of SNP intensities

We compare homozygous vs heterozygous risk SNPs based on their intensities. Considering allele A and B for a SNP, the following scattered plot gives intensity distribution for the genotypes AA, AB and BB. The analysis is performed on 180 male samples using a trial version of Golden Helix SVS software for extracting data from .CEL files.



In the scattered plots,
Red indicates risk allele is present at both positions.
Green Indicates risk allele is present at a single position.
Blue indicates absence of a risk allele.

Figure 3.7. Tri-colored scatter plot showing allele intensities

Using these plots, we can differentiate between a risk (red/green) and non-risk allele (blue). For example, if allele A is the risk allele then red color dominance is observed on the x-axis where as if allele B is the risk allele then red color dominance is observed on the y-axis. This helps us in interpreting the genotyping results.

For example, SNP1 in the left graph is variation ‘rs11196205’ of TCF7L2 gene susceptible to Type II Diabetes. In this case, C is the risk allele, hence red band is observed on the X-axis. SNP2 in the right graph is a variation ‘rs864745’ of JAZF1 gene susceptible to Type II Diabetes. In this case, the red band is leaning towards Y-axis since T is the risk allele. Such scattered plot can be used to observe the intensity variations amongst risk and non-risk alleles which also help to determine genotyping accuracy based on density of scattered points.

3.3 Preprocessing the data

3.3.1 Nature of the data

Some epidemiologic variables from the dataset are quantitative while other clinical risk factors are binary in nature. We select 9 quantitative and 3 binary variables which are also the dominant clinical risk factors.

Quantitative	bmi, act, wt, pufa, trans, magn, ceraf, gl, heme
Binary	famdb, chol, hbp

Table 3.9. Data type of clinical risk factors

Genomic risk factors (SNPs):

The difficulty in analyzing SNP data is the large occurrence of homozygous genotypes. This is handled by assigning each SNP variation into either of the following categories:

- Homozygous with risk allele
- Homozygous without risk allele
- Heterozygous.

3.3.2 Representation of SNP data

We use following approaches for representing genomic characteristics of an individual.

I. **Genomic risk score:** Genomic risk score of an individuals is calculated as the count of risk associated SNPs present in an individual out of the 20 risk associated SNPs selected.

For example, genomic risk score of 10 indicates that a particular individual carries 10 risk associated SNPs out of the selected 20 SNPs.

II. Biased SNP selection (Ternary representation):

A ternary number[14] has a base of 3. Some examples,

Ternary	1	10	12	22	212
Decimal	1	3	5	8	23

Consider a SNP of two alleles A & T, where T is the risk allele. We assign following values:

AA → 0
AT → 1
TT → 2

Before we form a ternary number using all 20 SNPs, we need to determine the order of SNPs such as which SNP would go at MSB position and which one at the LSB position in the ternary number representation. For this, we rank the SNPs based on the information gain [27]. Information gain ranks SNPs based on how well they separate data points with respect to the underlying class label.

Consider 20 SNPs as SNP$_1$, SNP$_2$,……SNP$_{20}$ $_{in}$ aligned in the order of decreasing information gains where SNP$_1$ is the one with the highest information gain with respect to a class indicating a case or a control. WEKA is used to obtain ranks of SNPs based on their information gain. We get the ternary representation as:

2	2	1	1	2	0	0	1	2	2	1	0	0	1	1	2	2	1	0	2

SNP$_1$ . . . . SNP$_{20}$

[14] http://en.wikipedia.org/wiki/Ternary_numeral_system

We convert this ternary number into its equivalent decimal value which would vary between 3,486,784,400 if 20 SNPs have value of 2 and the minimum 0 with complete absence of risk allele. This decimal value represents genomic weight of an individual.

Consider an individual with ternary representation of 20 SNPs as:

22112010210112202200

Then, genomic weight = 3,302,267,040 and genomic risk score = 14.

We include genomic weight or genomic risk score with other phenotypes as a new feature for determining similarity between individuals. Euclidean distance is applied on all features.

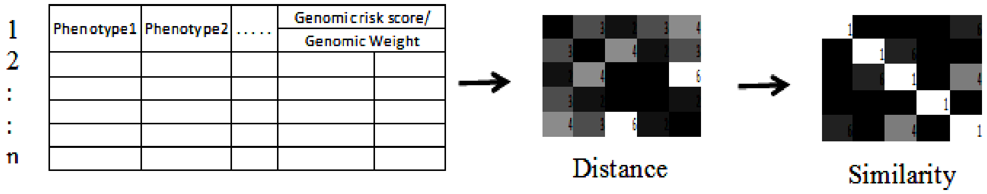


Figure 3.8. Similarity matrix formation using genomic score/weight

III. Unbiased SNP selection (Binary representation):

Again consider a SNP of two alleles A & T, where T is the risk allele. We assign following binary values:

AA → 00
AT → 01
TT → 11

Here our main purpose is to distinguish individuals based on the genomic pattern, hence there is no ordering of the SNPs while forming a binary representation.

A binary pattern of 20 SNPs is gives as:

1111011111000101000011110101111101110000

While using this method to group individuals, distances are calculated separately for phenotype characteristics and genomic characteristics. The phenotype attributes are compared based on Euclidean distances where as SNP data is compared based on Hamming distances.

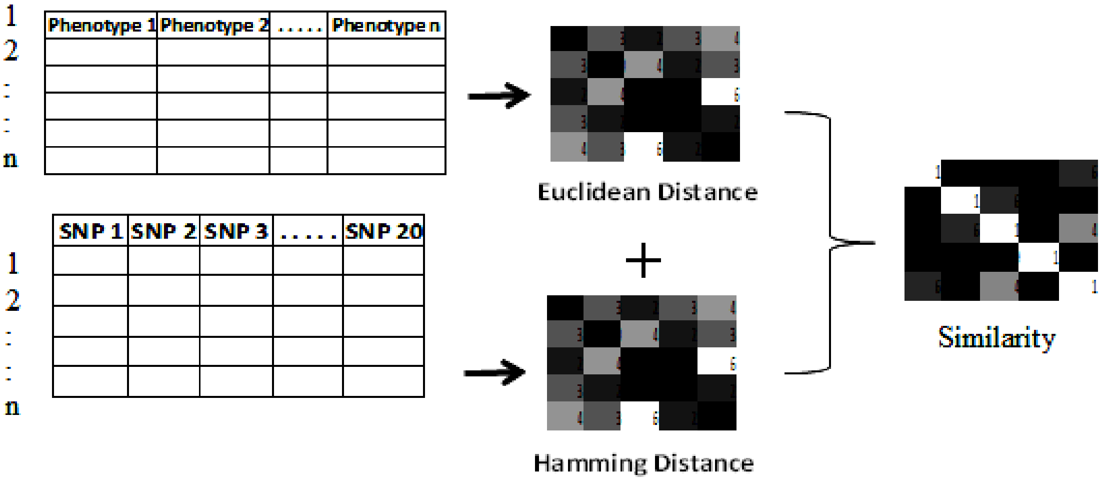


Figure 3.9. Similarity matrix formation using binary representation

3.4 Data normalization

Selection of appropriate measures for data representation before it is used for data mining depends on the nature and the scale of the underlying data set. The clinical-genomic dataset being heterogeneous in nature, we have to decide on a method to find similarity between two the individuals which would account for the diversity in the data. Genomic data has the same structure; however clinical data have different scales due to their different measuring units. The clinical information has to be in the same scale before it is used for fair comparison. In statistics, Normalization is the process which allows comparing data on different scales and brings them on a common scale before it is used for finding distance between two individuals [28].

We perform normalization in following two steps.

I. **Standardization:** Consider an original value as ‘x’ before standardization. Let ‘μ’ be the mean and ‘σ’ be the standard deviation of all values in the set where ‘x’ belongs. Then standardized value of ‘x’ is given by,

The standardized value,

$$s = \frac{x - \mu}{\sigma}$$

II. **Converting to a common range:** After the data is standardized we need to bring them to a common scale to avoid weight biasing while calculating distances between objects. We normalize the data further by calculating numerically equivalent value in the range of -1 and 1. This is nothing but dividing the value by something that is bigger than the value.

Consider d_i as an array containing standardized data from step I.

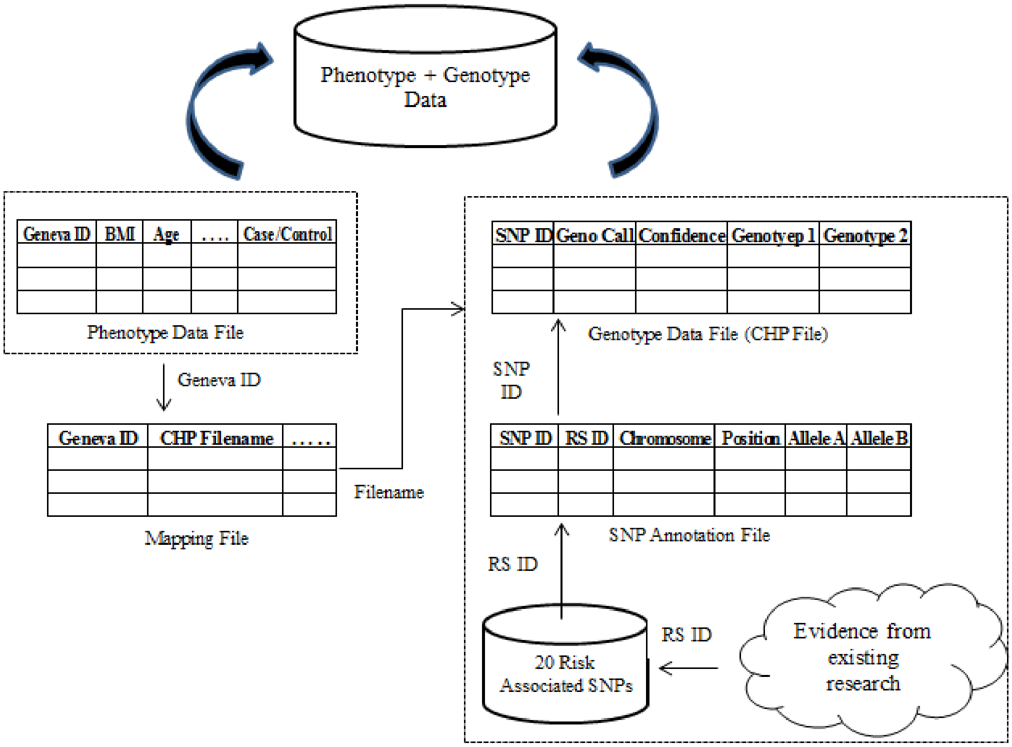
After mapping we get,

$$D_i = \frac{d_i}{\sqrt{\sum_{i=1}^{n} d_i^2}}$$

Thus all the values in the array are now mapped between -1 and 1.

3.5 Clinical-Genomic Data Integration

Following figure explains clinical-genomic data integration process in detail.



Geneva ID	BMI	Age	....	Case/Control

Geneva ID	CHP Filename	.....

SNP ID	Geno Call	Confidence	Genotyep 1	Genotype 2

SNP ID	RS ID	Chromosome	Position	Allele A	Allele B

Figure 3.10. Schematic architecture of clinical-genomic data integration

Data integration steps:

- Each individual in the study is assigned a GENEVA ID which acts a patient identifier. Using this ID, phenotype information about the individual is directly fetched from the file storing phenotype data.
- The mapping file contains .CHP file name corresponding to each GENEVA ID. Each .CHP file stores information of 909623 SNPs including the SNP ID and both the alleles.
- The selected 20 SNPs are extracted from the .CHP file and inserted under corresponding GENEVA ID along with other phenotype data.

3.6 Tools Used

i. WEKA

WEKA is an open source data mining software consisting of collection of machine learning algorithms. It contains tools for data preprocessing, classification, clustering, association and visualization.

ii. MATLAB

It is a computing environment which allows you to perform numeric intensive operations, matrix manipulations, data mining and data plotting, etc. It supports several functions which we used to perform spectral clustering.

iii. Statistical Analysis Tool (Microsoft Excel Add-in)

It is a collection of statistical and engineering macro functions such as Correlation, T-Test, etc. which be performed on the worksheet data directly.

iv. Aspera Client

It is a fully featured desktop client used to initiate and automate high performance transfer of large data over the internet. We used this tool to download datasets from dbGAP repository.

v. MySQL Server

It is an open source database we used for storing integrated clinical-genomic information from the dataset.

vi. Golden Helix Software

It is a high performance analytic tool for managing, analyzing and visualizing complex genomic data. A trial version of this tool was used to retrieve SNP intensity level data from .CEL files.

Chapter 4

EXPERIMENTAL ANALYSIS AND RESULTS

We present the results of the analysis performed on the dataset to understand the contribution of clinical and genomic risk factors in disease development. We present comparison different classification and clustering methods using different input datasets and several validation criteria.

Most of the statistical data analysis is done using Microsoft Excel's library and tools where as data mining is performed using WEKA And MATLAB. We use Windows machine with 3GB RAM for data mining. Linux scripting commands are used to extract data from the Affymetrix CHP files. Relational clinical-genomic knowledge data is maintained using MySQL database server.

4.1 Initial Data Analysis

4.1.1 Correlation Analysis

To determine if two variables are statistically dependent on each other, Spearman Correlation Rank Coefficient is calculated amongst them [29]. It is a rank based statistic method to assess strength of the associations between two variables.

The Spearman rank correlation coefficient is given by,

$$r' \equiv 1 - 6 \sum \frac{d^2}{N\left(N^2 - 1\right)},$$

where *d* is the difference in the statistical ranks of corresponding variables. (Rank here is a specific number indicating order of a value in the list).

Consider two variables X and Y, where we want to find relation of values of Y with respect to the values of X. If the Spearman Correlation Coefficient between X and Y is positive then it indicates that Y tends to increase when X increases. If it's negative then it indicates that Y tends to decrease when X increases. If its zero then it indicates that Y does not tend to either increase or decrease with respect to X.

Correlation between genomic risk score and % of cases:

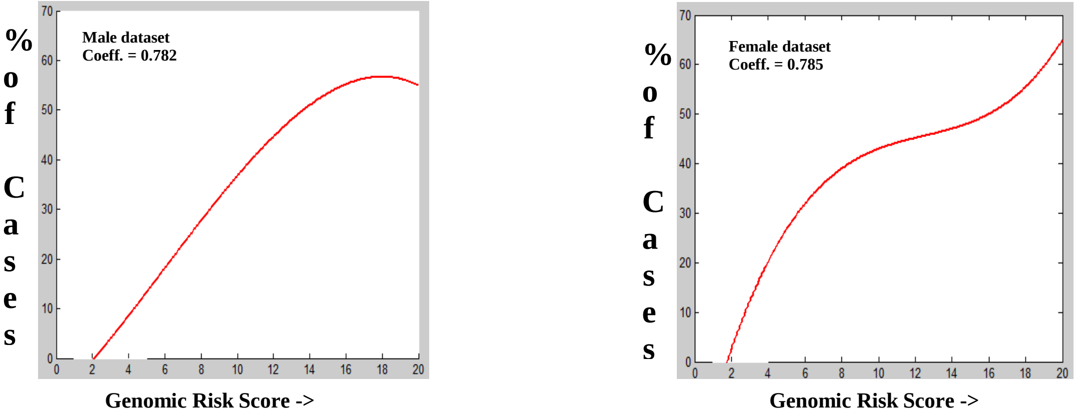
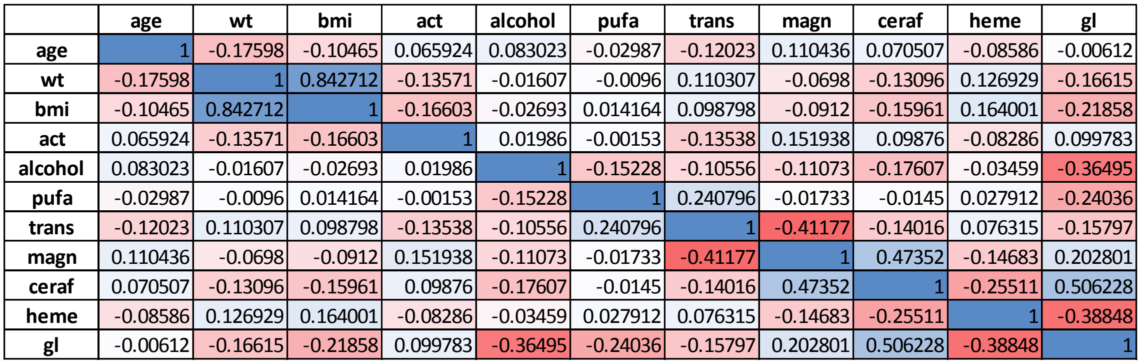


Figure 4.1. % of cases against genomic risk score

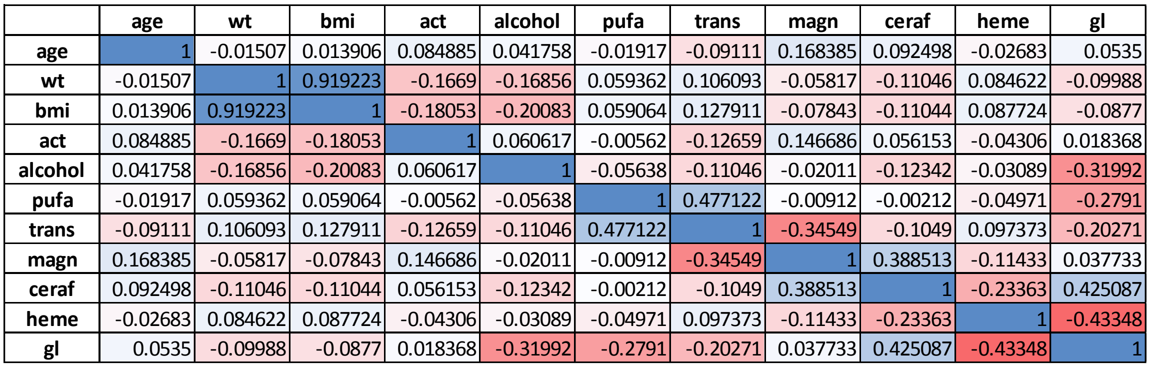
A strong correlation is found between the number of risk alleles and the likelihood of being a diabetic. Cases increase with the increasing number of risk associated alleles.

Correlation amongst phenotypes:



	age	wt	bmi	act	alcohol	pufa	trans	magn	ceraf	heme	gl
age	1	-0.17598	-0.10465	0.065924	0.083023	-0.02987	-0.12023	0.110436	0.070507	-0.08586	-0.00612
wt	-0.17598	1	0.842712	-0.13571	-0.01607	-0.0096	0.110307	-0.0698	-0.13096	0.126929	-0.16615
bmi	-0.10465	0.842712	1	-0.16603	-0.02693	0.014164	0.098798	-0.0912	-0.15961	0.164001	-0.21858
act	0.065924	-0.13571	-0.16603	1	0.01986	-0.00153	-0.13538	0.151938	0.09876	-0.08286	0.099783
alcohol	0.083023	-0.01607	-0.02693	0.01986	1	-0.15228	-0.10556	-0.11073	-0.17607	-0.03459	-0.36495
pufa	-0.02987	-0.0096	0.014164	-0.00153	-0.15228	1	0.240796	-0.01733	-0.0145	0.027912	-0.24036
trans	-0.12023	0.110307	0.098798	-0.13538	-0.10556	0.240796	1	-0.41177	-0.14016	0.076315	-0.15797
magn	0.110436	-0.0698	-0.0912	0.151938	-0.11073	-0.01733	-0.41177	1	0.47352	-0.14683	0.202801
ceraf	0.070507	-0.13096	-0.15961	0.09876	-0.17607	-0.0145	-0.14016	0.47352	1	-0.25511	0.506228
heme	-0.08586	0.126929	0.164001	-0.08286	-0.03459	0.027912	0.076315	-0.14683	-0.25511	1	-0.38848
gl	-0.00612	-0.16615	-0.21858	0.099783	-0.36495	-0.24036	-0.15797	0.202801	0.506228	-0.38848	1

Figure 4.2. Spearman coefficients amongst phenotypes of male dataset



	age	wt	bmi	act	alcohol	pufa	trans	magn	ceraf	heme	gl
age	1	-0.01507	0.013906	0.084885	0.041758	-0.01917	-0.09111	0.168385	0.092498	-0.02683	0.0535
wt	-0.01507	1	0.919223	-0.1669	-0.16856	0.059362	0.106093	-0.05817	-0.11046	0.084622	-0.09988
bmi	0.013906	0.919223	1	-0.18053	-0.20083	0.059064	0.127911	-0.07843	-0.11044	0.087724	-0.0877
act	0.084885	-0.1669	-0.18053	1	0.060617	-0.00562	-0.12659	0.146686	0.056153	-0.04306	0.018368
alcohol	0.041758	-0.16856	-0.20083	0.060617	1	-0.05638	-0.11046	-0.02011	-0.12342	-0.03089	-0.31992
pufa	-0.01917	0.059362	0.059064	-0.00562	-0.05638	1	0.477122	-0.00912	-0.00212	-0.04971	-0.2791
trans	-0.09111	0.106093	0.127911	-0.12659	-0.11046	0.477122	1	-0.34549	-0.1049	0.097373	-0.20271
magn	0.168385	-0.05817	-0.07843	0.146686	-0.02011	-0.00912	-0.34549	1	0.388513	-0.11433	0.037733
ceraf	0.092498	-0.11046	-0.11044	0.056153	-0.12342	-0.00212	-0.1049	0.388513	1	-0.23363	0.425087
heme	-0.02683	0.084622	0.087724	-0.04306	-0.03089	-0.04971	0.097373	-0.11433	-0.23363	1	-0.43348
gl	0.0535	-0.09988	-0.0877	0.018368	-0.31992	-0.2791	-0.20271	0.037733	0.425087	-0.43348	1

Figure 4.3. Spearman coefficients amongst phenotypes of female dataset

Here, values in blue color indicate positive correlation where as those in red indicates negative correlation. All the diagonal values are to be ignored. The strength of the correlation is based on the magnitude of value. Coefficient value < 0.5 represents a mild correlation whereas value > 0.5 represents a strong correlation. Following are the significant correlations observed both in male and female datasets.

Correlation type	Variable X	Variable Y	Correlation Strength
Positive	Weight	BMI	Strong
	Magn	Ceraf	Mild
	Gl	Ceraf	Mild
Negative	Alcohol	Gl	Mild
	Heme	Gl	Mild
	Trans	Magn	Mild

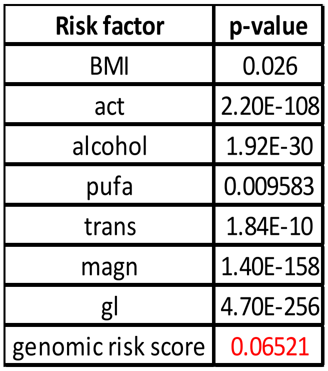
Table 4.1. Spearman coefficients amongst clinical risk factors

Thus we observe a marginal correlation amongst dietary habits.

4.1.2 Student t-test

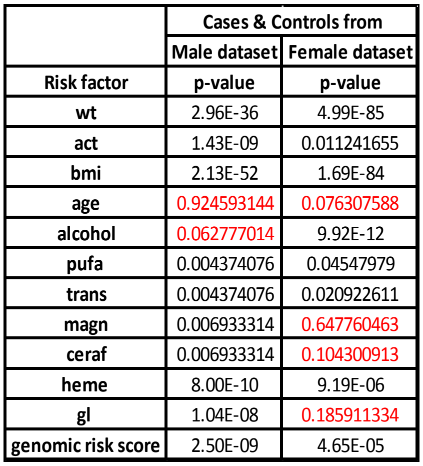
This is one of the most commonly used methods to compare two datasets that are collected independently of each other [30]. It determines probability based on a hypothesis which states that the two datasets are either same or different with respect to a variable or attribute using differences in the means of the two samples. A probability of 0.05 or less indicates that the two datasets can be distinguished using the corresponding variable for which the probability is observed.

We used this method to determine which risk factors best distinguishes male and female population also to determine which risk factors best distinguish cases and controls. The following data is from a male and a female dataset each of 1800 records with control and cases present in equal proportions.



Risk factor	p-value
BMI	0.026
act	2.20E-108
alcohol	1.92E-30
pufa	0.009583
trans	1.84E-10
magn	1.40E-158
gl	4.70E-256
genomic risk score	0.06521

(1)



	Cases & Controls from	
	Male dataset	Female dataset
Risk factor	p-value	p-value
wt	2.96E-36	4.99E-85
act	1.43E-09	0.011241655
bmi	2.13E-52	1.69E-84
age	0.924593144	0.076307588
alcohol	0.062777014	9.92E-12
pufa	0.004374076	0.04547979
trans	0.004374076	0.020922611
magn	0.006933314	0.647760463
ceraf	0.006933314	0.104300913
heme	8.00E-10	9.19E-06
gl	1.04E-08	0.185911334
genomic risk score	2.50E-09	4.65E-05

(2)

Table 4.2 t-test results (1) male vs female (2) cases vs controls
(Here p-values > 0.05 are marked in red.)

This tells us that males and female can be easily distinguished based on their phenotype characteristics; however genomic characteristics are not gender biased, hence does not help in distinguishing the two datasets. Genomic risk score does not differentiate the gender unlike prominent clinical risk factors. However it certainly proved to be a distinguishing factor between cases and controls in both males and females. Above results also confirm that diabetes is independent of the age.

4.2 Supervised Learning (Classification)

We performed classification to determine analyses that could predict diabetes diagnosis in this dataset. Depending upon the nature of the both the male and female dataset, we selected Bayesian Network and Decision Tree for performing the classification.

For males, we used a training dataset of 1600 records with 800 cases and 800 controls. The test dataset is of 200 records. For females, we used a training dataset of 2400 records with 1200 cases and 1200 controls. Similar to the males, the test dataset is of 200 records. Analyses were performed using two separate datasets, one with only phenotypic data and the other with both phenotypic and genotypic data to understand whether inclusion of genetic information improves classification accuracy. To compare performance of the prediction model obtained using the two classification algorithms, we used cross validation method and the ROC area.

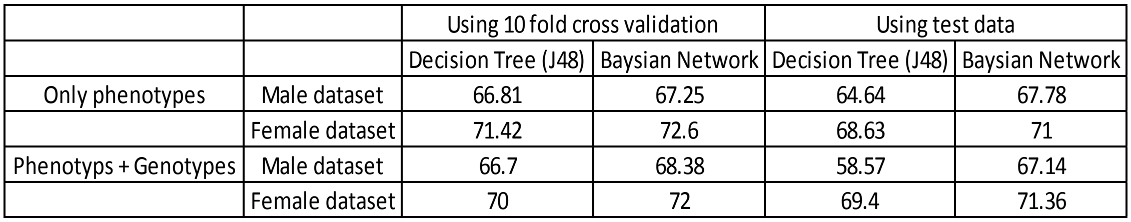
Cross validation: It is a technique to estimate how the prediction model will perform in practice when applied to an independent dataset. Cross-validation consists of several rounds where in each round, analysis is performed on one subset called as training set and the performance is evaluated on another subset called as test set. Multiple sets of such rounds reduce variability in the performance.

We are using 10-fold cross validation, where the original input dataset is divided into 10 partitions. At each of the 10 rounds, one non-repeated partition is used as a test set and the remaining 9 partitions are used as training sets. The results from all 10 rounds are then averaged to give single estimate.

ROC Area: ROC curve is a representation of the trade-off between true positive rate and false positive rates, false positive rate on the X-axis and true positive rate on the Y-axis.

The area under the ROC curve measures ability of the prediction model to accurately classify cases and controls. It is a measure how well a parameter can distinguish between two diagnostic groups.

For example, consider a healthy patient with a score of S_1 and a diseased patient with a score of S_2, then area under the ROC curve is an estimate of $P[S_2>S_1]$ where the larger the value indicates a tendency towards diabetes.



		Using 10 fold cross validation		Using test data	
		Decision Tree (J48)	Baysian Network	Decision Tree (J48)	Baysian Network
Only phenotypes	Male dataset	66.81	67.25	64.64	67.78
	Female dataset	71.42	72.6	68.63	71
Phenotyps + Genotypes	Male dataset	66.7	68.38	58.57	67.14
	Female dataset	70	72	69.4	71.36

Table 4.3 Comparison of classification accuracy

Bayesian Network gives slighly better accuracy as compared to J48 (C 4.5 implementation). To further determine whether genomic data affects the the accuracy of classification, we calculate the ROC area. The following graph shows change in the ROC area when different risk factors are subsequently added to the input. The primary phenotypes are added in the ascending order of their information gain rank.

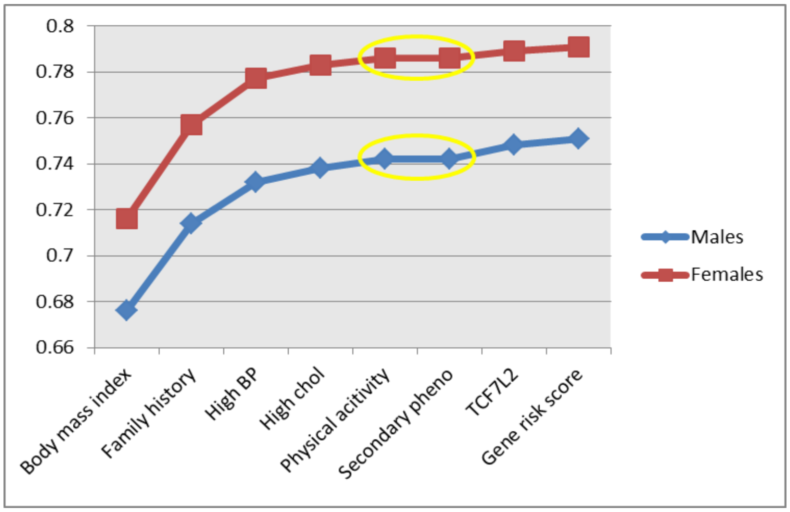
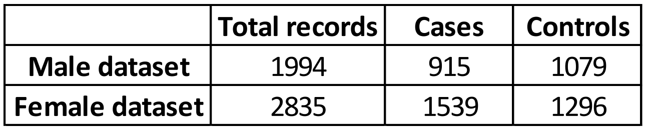


Figure 4.4. ROC Area against risk factors

It was observed that when TCF7L2 gene information or the genomic risk score is included, the ROC area increases by 5-6%. Secondary phenotypes such as fat intake, cereal fiber intake, magnesium intake, glycemic load, etc. do not produce any change in the ROC area.

4.3 Unsupervised Learning (Clustering)

Clustering methods are selected that best differentiate between cases and controls, and identify other groups present in the dataset. The following dataset was selected for clustering purposes.



	Total records	Cases	Controls
Male dataset	1994	915	1079
Female dataset	2835	1539	1296

Clusters are selected that have good validity index and are best representative of meaningful groups in the dataset.

4.3.1 Cluster validation and accuracy

Cluster validation using silhouette: The silhouette validation technique [31] calculates the silhouette width for each sample, cluster and the whole dataset. It is a combined measure of:

- Intra-cluster distance a(i): average dissimilarity of sample i to all the other samples from the same cluster (how close the samples are within the same cluster).
- Inter-cluster distance b(i): minimum of average dissimilarity of sample i to all the samples in the other cluster (how far the samples are between different clusters).

The silhouette S(i) is given by,

$$S(i) = \frac{(b(i) - a(i))}{\max\{a(i), b(i)\}} \qquad - 1 <= S(i) <=1$$

S(i) close to 1 => the sample is well clustered to the appropriate cluster.

S(i) close to 0 => the sample could belong to another cluster.

The average of all silhouette values of sample of a particular cluster determines the validity of the cluster.

Clustering accuracy: The centroid of a cluster is its center of gravity. Tightness of a cluster is determined by average distance of every sample in the cluster with respect to its centroid.

Consider two centroids C1 and C2 of two clusters obtained. We focus on distance of every sample from the centroid of the cluster it belongs as well as from the centroid of other clusters. Consider the figure below.

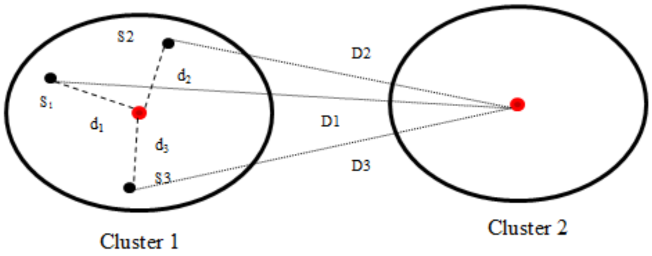


Figure 4.5. Sample to centroid distances

Sample S1, S2 and S3 are at distances d1, d2 and d3 respectively from the centroid of the clusters they belong to i.e. Cluster 1 and are at distances D1, D2 and D3 respectively from the centroid of neighboring Cluster 2.

We calculate,

i. Average of distances of samples from the centroid of its own cluster,

   $D = avg(d_1+d_2+\ldots d_n)/n$

ii. Average of distances of sample from the centroid of another cluster,

   $D' = avg(D_1+D_2+\ldots..D_n)/n$

4.3.2 Selection of number of clusters

We use silhouette values to determine number of clusters that possess maximum validity index.

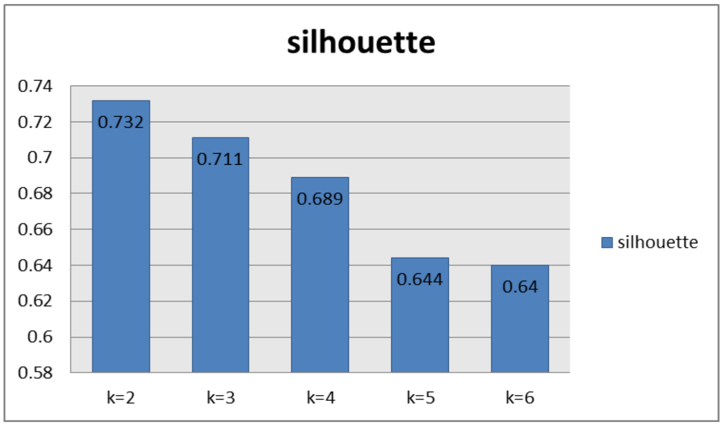
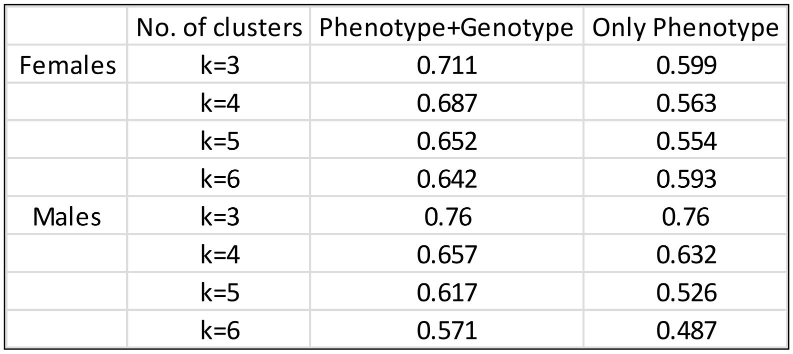


Figure 4.6. Silhouette indices vs number of clusters

From this it was observed that for k>4, the silhouette index drops below 0.68. Since our aim is to identify different groups within the dataset, k>2 and k<=4 is the preferred choice. Considering good silhouette value as well as clinical significance of the resulting groups, k=4 is the selected choice for number of clusters.

4.3.3 Input data selection

Significance of genomic data: To assess whether inclusion of genomic data affects clustering performance, we compare validity of the clusters using silhouette indices. We select primary clinical risk factors as the phenotypes. We take average of the silhouettes obtained using both biased and unbiased SNP selection.



	No. of clusters	Phenotype+Genotype	Only Phenotype
Females	k=3	0.711	0.599
	k=4	0.687	0.563
	k=5	0.652	0.554
	k=6	0.642	0.593
Males	k=3	0.76	0.76
	k=4	0.657	0.632
	k=5	0.617	0.526
	k=6	0.571	0.487

Table 4.4. Silhouette indices comparison

Silhouette index considerably improves (by ~ 0.1) when genetic information is used along with phenotypes for clustering.

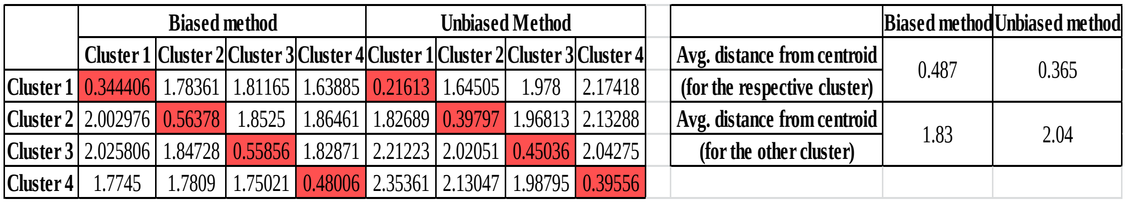
SNP data representation for clustering:

When using risks associated SNPs along with phenotypes for clustering, there are two ways to represent the genomic details (explained earlier in 3.3.2).

- Biased SNP selection (ternary representation)
- Unbiased SNP selection (binary representation)

Clustering on male and female datasets using both these methods produced similar silhouette indices in the range of ~ 0.65 to ~ 0.68 for k=4. To further compare these two methods, we compared the clusters based on goodness with respect to the centroid as explained in 3.1.

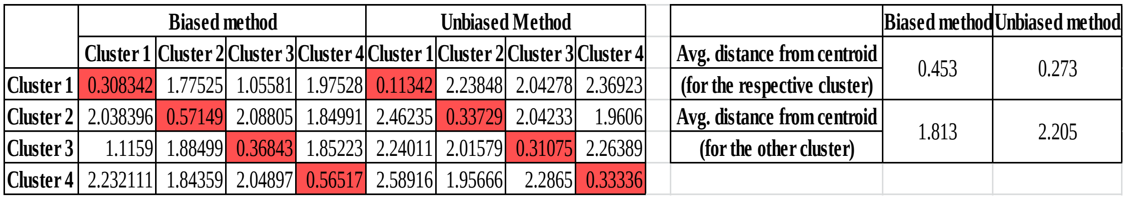
Male dataset:



	Biased method				Unbiased Method			
	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Cluster 1	0.344406	1.78361	1.81165	1.63885	0.21613	1.64505	1.978	2.17418
Cluster 2	2.002976	0.56378	1.8525	1.86461	1.82689	0.39797	1.96813	2.13288
Cluster 3	2.025806	1.84728	0.55856	1.82871	2.21223	2.02051	0.45036	2.04275
Cluster 4	1.7745	1.7809	1.75021	0.48006	2.35361	2.13047	1.98795	0.39556

	Biased method	Unbiased method
Avg. distance from centroid (for the respective cluster)	0.487	0.365
Avg. distance from centroid (for the other cluster)	1.83	2.04

Female dataset:



	Biased method				Unbiased Method			
	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Cluster 1	0.308342	1.77525	1.05581	1.97528	0.11342	2.23848	2.04278	2.36923
Cluster 2	2.038396	0.57149	2.08805	1.84991	2.46235	0.33729	2.04233	1.9606
Cluster 3	1.1159	1.88499	0.36843	1.85223	2.24011	2.01579	0.31075	2.26389
Cluster 4	2.232111	1.84359	2.04897	0.56517	2.58916	1.95666	2.2865	0.33336

	Biased method	Unbiased method
Avg. distance from centroid (for the respective cluster)	0.453	0.273
Avg. distance from centroid (for the other cluster)	1.813	2.205

Figure 4.7. Sample to centroid distances in the resulting clusters

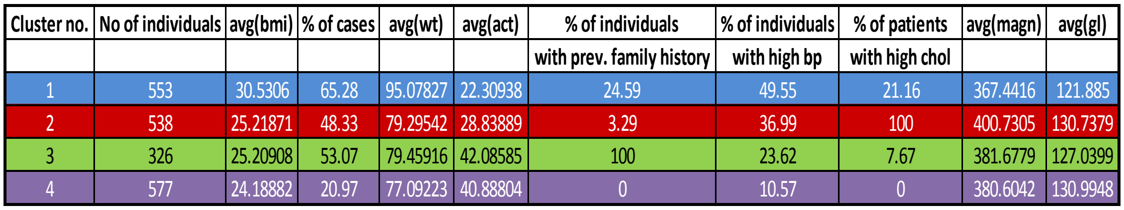
The colored values indicate average distance of samples from centroid of the same cluster it belongs. The sample-to-centroid distance (with respect to its own cluster) is greater using biased SNP selection as compared to the unbiased SNP selection. Thus unbiased SNP selection reduces the intra-cluster distances. The sample-to-centroid distance (with respect to other cluster) is greater using unbiased SNP selection as compared to the biased SNP selection. Thus unbiased SNP selection increases the inter-cluster distances. From this we infer that clustering accuracy improved due to unbiased selection of SNPs.

4.3.4 Clustering results

Previous analysis indicated following specifications for obtaining valid and meaningful clusters:

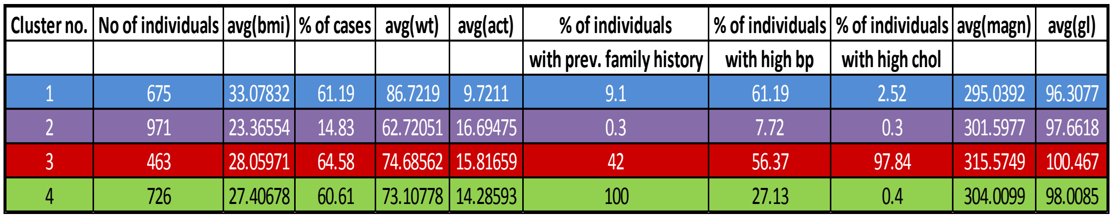
- Number of clusters = 4
- Inclusion of both clinical and genetic information.
- Unbiased SNP selection.

Using these specifications, we produced following clusters.



Cluster no.	No of individuals	avg(bmi)	% of cases	avg(wt)	avg(act)	% of individuals	% of individuals	% of patients	avg(magn)	avg(gl)
						with prev. family history	with high bp	with high chol		
1	553	30.5306	65.28	95.07827	22.30938	24.59	49.55	21.16	367.4416	121.885
2	538	25.21871	48.33	79.29542	28.83889	3.29	36.99	100	400.7305	130.7379
3	326	25.20908	53.07	79.45916	42.08585	100	23.62	7.67	381.6779	127.0399
4	577	24.18882	20.97	77.09223	40.88804	0	10.57	0	380.6042	130.9948

Figure 4.8. Clusters obtained using male dataset



Cluster no.	No of individuals	avg(bmi)	% of cases	avg(wt)	avg(act)	% of individuals	% of individuals	% of individuals	avg(magn)	avg(gl)
						with prev. family history	with high bp	with high chol		
1	675	33.07832	61.19	86.7219	9.7211	9.1	61.19	2.52	295.0392	96.3077
2	971	23.36554	14.83	62.72051	16.69475	0.3	7.72	0.3	301.5977	97.6618
3	463	28.05971	64.58	74.68562	15.81659	42	56.37	97.84	315.5749	100.467
4	726	27.40678	60.61	73.10778	14.28593	100	27.13	0.4	304.0099	98.0085

Figure 4.9. Clusters obtained using female dataset

We then assess the genomic risk associated with each cluster based on % of individuals carrying the risk allele.

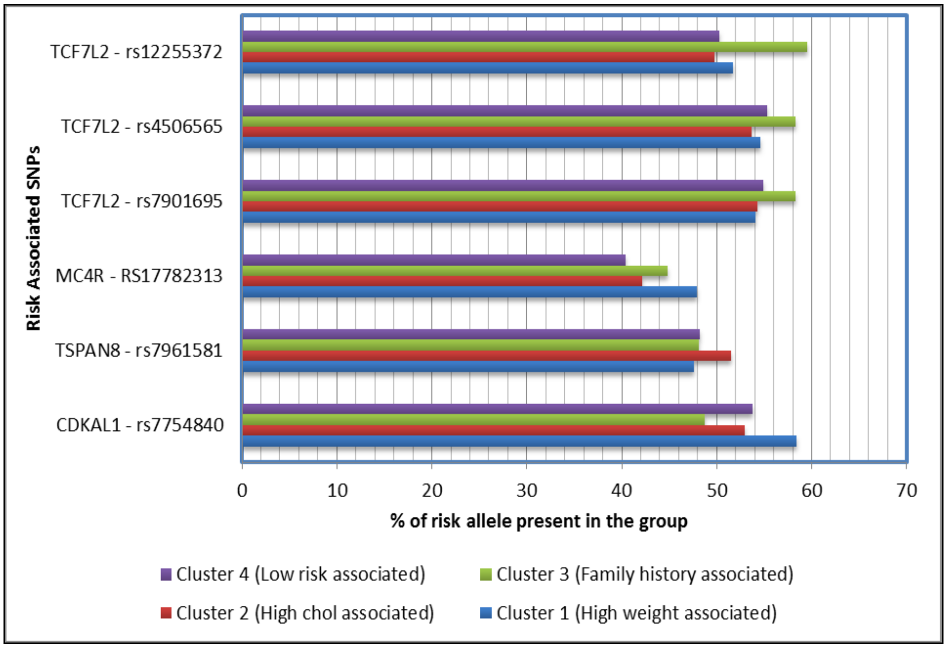


Figure 4.10. Risk associated SNPs characterizing different clusters of the male dataset

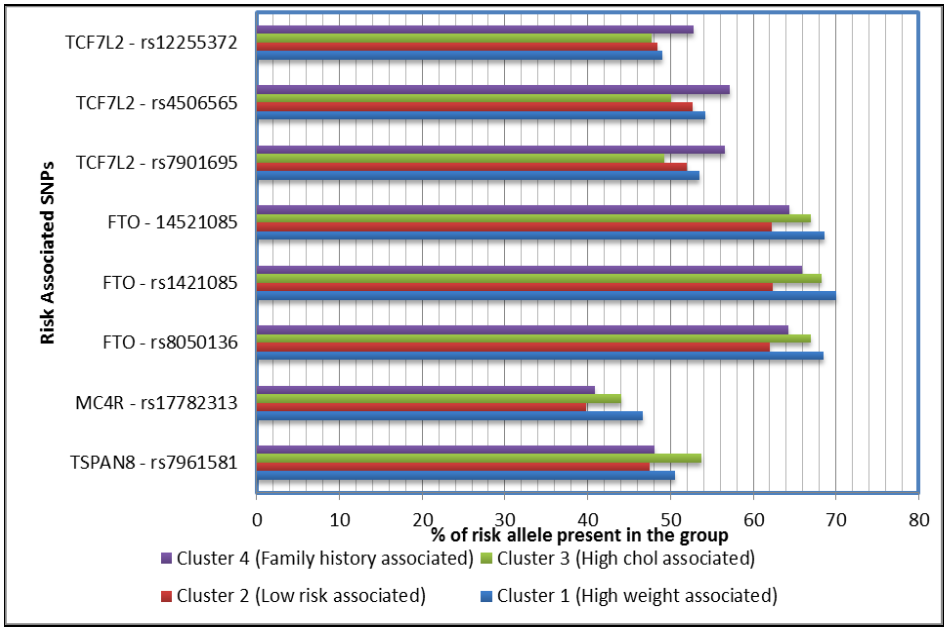


Figure 4.11. Risk associated SNPs characterizing different clusters of the female dataset

Clustering resulted in groups of individuals based on their clinical and genomic risk similarity. Details of the clusters as follows.

- **High risk cluster:**
  - high in weight
  - enrichment of obesity genes; high % of individuals with obesity gene variants – MC4R in case of males and FTO in case of female
  - cases present > 60%.
- **Low risk cluster:**
  - low values for clinical and genomic risk factors
  - cases present <=20%.
- **Intermediate risk cluster 1:**
  - high cholesterol associativity
  - high % of individuals with TSPAN8 variant
  - cases present ~50-60%.
- **Intermediate risk cluster 2:**
  - previous family history associativity
  - enrichment of TCF7L2 gene; high % of individuals with TCF7L2 variant
  - cases present ~50-60%.

We also performed using different combinations of input parameters on male and female datasets separately.

Input dataset	Observations
Secondary phenotypes (dietary habits) + genotype	Silhouette < 0.6 for k> 3. Obtained clusters are poor representative of the knowledge.
Only phenotypes	For k>2, silhouette drops considerably below 0.6. For k=2, two basic clusters indicative of cases group and controls group can be obtained.
Only genotypes	Very poor validity index. Max. Silhouette of 0.45. Use of only genotypes adds confusion to the clustering. Cannot identify similarity between individuals based only on genotypes.

Chapter 5

CONCLUSION AND FUTURE WORK

5.1 Conclusion

We proposed and described analytical methods to study clinical-genomic risk factors of Type II Diabetes in a white American population. Use of risk associated SNPs gave us an exposure to the genetic determinants in diabetes. Data analysis using student t-tests indicated genomic risk score as potential differentiating factor amongst cases and controls along with other prominent clinical risk factors. This was also confirmed by a strong positive correlation observed between the genomic risk score and the occurrence of cases.

In case of predictive analysis, Bayesian Network proved to be a suitable classifier with the classification accuracy of ~71%. Inclusion of genotype data to the phenotype showed marginal improvement in the classification accuracy. We observed effect on ROC area with respect to different risk factors. Type II Diabetes associated genes such as TCF7L2 over dietary habits showed considerable increase in the ROC area proving significance of genetic risk in differentiating cases and controls. In case of clustering, we analyzed effect of variations in the input risk factors, SNP representations, and the number of clusters on the clustering performance. We compared resulting clusters based on these parameters and their clinical significance. Use of genomic data with unbiased SNP selection improved the clustering validity as well as accuracy.

From the results it was evident that genomic risk plays a role in determining Type II Diabetes risk. The clusters yield interesting hints for potentially relevant combinations of clinical-risk factors that would certainly benefit researcher for analyzing case-control groups, gain more insight into the genetic behaviors and generate biological hypotheses. Thus the overall goal of assessing genetic risk along with clinical risk in the development of Type II Diabetes is achieved.

5.2 Future Work

While performing clustering, it was observed that some steps of the spectral clustering steps such as calculating degree matrix from the similarity matrix, eigenvectors from the laplacian matrix were running slow due to large sizes of matrices. In the future, we plan to design a faster implementation of matrix computations and scale the infrastructure to accommodate larger datasets.

The interpretation of genomic risk from the clusters is at present limited to prominent genes. We would like to now focus on a larger set of risk associated SNPs and try to find if clusters can be produced that are suggestive of association between genes and environmental factors. We also plan to extend our study to ethnically diverse populations susceptible to Type II Diabetes that would help us investigate clinical and genomic risk factors amongst different groups.

Chapter 6

REFERENCES

[1] Guttmacher, A. E., & Collins, F. S. (2009). Realizing the Promise of Genomics in Biomedical Research. JAMA.

[2] Ommen, G. B. (2002). The Human Genome Project and the future of diagnostics, treatment and prevention. Journal Of Inherited Metabolic Disease, 183-188.

[3] Ginsburg, G. S., & McCarthy, J. J. (2001, December). Personalized Medicine: revolutionizing drug discovery and patient care. TRENDS in Biotechnology, pp. 491-495.

[4] Khoury, M. J., & Yang, Q. (Vol. 9, No. 3 (May, 1998)). The Future of Genetic Studies of Complex Human Diseases: An Epidemiologic Perspective. Epidemiology, 350-354.

[5] Haynie, H. W. (2008). Discoveries in Genetic Variation Creating Possibilities for Predictive Medicine. Retrieved from Memphis Medical News.

[6] H. Abarbanel, C. Callan, W. Dally, F. Dyson, T. Hwa, S. Koonin, H. Levine, O. Rothaus, R. Schwitters, C. Stubbs, P. Weinberger. (2000). Data Mining and the Human Genome. The MITRE Corporation.

[7] Thompson, D. (2008, December 30). HealthDay. Retrieved from

Washingtonpost:http://www.washingtonpost.com/wp-dyn/content/article/2008/12/30/AR2008123001799.html

[8] W. Gregory Feero, A. E. (2010). Genomic Medicine — An Updated Primer. The New England Journal of Medicine.

[9] Linda Liu, D. N. (2009). : Defining SNP signatures for prediction of onset in complex diseases. Standford University.

[10] Stéphane Cauchi, Y. E. (2007). TCF7L2 is reproducibly associated with type 2 diabetes in various ethnic groups: a global meta-analysis. Springer.

[11] McCarthy, M. I. (2011, JAN 11). From hype to hope? A journey through the genetics of Type 2 diabetes. DIABETIC Medicine.

[12] Witten, I. H., & Frank, E. (2005). Data Mining: Practical Machine Learning Tools and Techniques. Elsevier.

[13] Neapolitan, R. (2004). Learning Bayesian Networks. Upper Saddle River, NJ: Prentice Hall.

[14] Quinlan, J. R. (1993). C4.5: Programs for Machine Learning. Springer, 235-240.

[15] MacQueen, J. B. (1967). Some Methods for classification and Analysis of Multivariate Observations. Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability (pp. 281-297). University of California Press.

[16] Cvetković, D. M., Doob, M., & Sachs, H. (1980). Spectra of Graph: theory and application. Academic Press.

[17] Malik, J. S. (2000). Normalized Cuts and Image Segmentation. IEEE.

[18] Sherry, S. T. (2007). The NCBI dbGaP database of genotypes and phenotypes. Nature Genetics, 1181-1186.

[19] Fotiadis, D. I., Goletsis, Y., Exarchos, T., Giannakeas, N., & Rigas, G. (n.d.). Inferencing in In Silico Oncology: Exploiting Expressions, Biomarkers and Clinical Data for Clinical Decision Support. University of Illinois.

[20] Selinski, S., & Statistik, F. (n.d.). Similarity Measures for Clustering SNP and Epidemiological Data. Interdisciplinary Study Group on Gene Environment Interaction and Breast Cancer in Germany.

[21] West, M., Ginsburg, G. S., Huang, A. T., & Nevins, J. R. (2006). Embracing the complexity of genomic data for personalized medicine. Genomic Research, 559-566.

[22] Ban, H.-J., Heo, J. Y., Oh, K.-S., & Park, K.-J. (2010). Identification of Type 2 Diabetes-associated combination of SNPs using Support Vector Machine. BMC Genetics.

[23] Gregory F. Cooper, P. H.-Y. (n.d.). An Efficient Bayesian Method for Predicting Clinical Outcomes from Genome-Wide Data. University of Pittsburgh.

[24] Nevins, J. R., Huang, E. S., Dressman, H., & Pittman, J. (2003). Towards integrated clinico-genomic models for personalized medicine: combining gene expression signatures and clinical factors in breast cancer outcomes prediction. Human Molecular Genetics, 153-157.

[25] Ng, M. K., Li, M. J., Ao, S. I., Sham, P. C., Cheung, Y.-m., & Huang, J. Z. Clustering of SNP Data with Application to Genomics. ICDMW. Sixth IEEE

International Conference on Data Mining.

[26] Miguel-Yanes, J. M., Shrader, P., Pencina, M. J., Fox, C. S., Manning, A. K., Grant, R. W., et al. (2010, October). Genetic risk reclassification for type 2 diabetes by age below or above 50 years using 40 type 2 diabetes risk single nucleotide polymorphisms. Diabetes Care.

[27] Guyon, I., & Elisseeff, A. (2003). An Introduction to Variable and Feature Selection. Journal of Machine Learning Research, 1157-1182.

[28] Sterne, J., & Kirkwood, B. R. (2003). Essential medical statistics. Blackwell Publishing.

[29] Higgins, J. (2005). The Radical Statistician.

[30] McDonald, J. (2009). Handbook of Biological Statistics. Sparky House Publishing.

[31] P.J. Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. 1987. Journal of Computational and Applied Mathematics. 20. 53-65.