이학박사 학위논문

Predicting disease predisposition patterns of the personal genome based on disease hierarchy

질병 계층 기반 개인 유전체의 질병 위험도 예측

2013년 2월

서울대학교 대학원

협동과정 생물정보학

나 영 지

Abstract

Predicting disease predisposition patterns of the personal genome based on disease hierarchy

Young-Ji, Na

Interdisciplinary Program in Bioinformatics

College of Natural Science

Seoul National University

The advent of next-generation sequencing (NGS) technologies has had a huge impact upon functional genomics. The NGS technologies generate millions of short sequence reads per run, making it possible to sequence entire human genomes in a matter of weeks. These NGS technologies have already been employed to sequence the constitutional genomes of several individuals. Ambitious efforts like the 1000 Genomes Project and the Personal Genomes Project hope to add thousands more. The first five cancer genomes revealed thousands of novel somatic mutations and implicated new genes in tumor development and progression. Current knowledge of the genetic variants that underlie disease susceptibility, treatment response and other

phenotypes will continually improve as these studies expand the catalog of DNA sequence variation in humans.

As the cost of sequencing continues to freefall, the challenge of solving the data analysis and storage problems becomes more pressing. But those issues are nothing compared to the challenge facing the clinical community who are seeking to mine the genome for clinically actionable information. However, present analytical methods are insufficient to make genetic data accessible in a clinical context, and the clinical usefulness of these data for individual patients has not been formally assessed. Here, I focus on evaluating individual predispositions to specific phenotypic traits given their genetic backgrounds.

In this dissertation, I present a computational method for associating variants in the personal genome sequencing data with predispositions to disease. The method works by ranking all variants in the personal genome as potential disease risks, and reporting MeSH terms that are significantly associated with highly ranked genes. To identify genetic variants associated diseases, I obtained high-throughput sequencing data in several cancer types (acute myeloid leukemia, bladder cancer, breast cancer, colon cancer, glioblastoma multiforme, kidney cancer, lung adenocarcinoma, lung squamous cell carcinoma, malignant melanoma, ovarian serous cystadenocarcinoma and prostate cancer) and non-cancer types (Crohn's disease, focal segmental glomerulosclerosis, and retinitis pigmentosa). From disease-gene association in the

OMIM, I reconstructed relations of diseases and genes in the MeSH tree structures in order to consider the human disease hierarchical structure of human disease ontology.

The results showed the distribution of mutual information in the MeSH disease category differs according to the population in the healthy people. It suggests that in order to interpret personal genome properly, we may consider population information together. In addition, MeSH disease terms are more highly ranked in the patients than healthy people. Disease-enrichment analysis showed Cancer, Neurological, Endocrine, and Immunological categories were over-represented in the patients as well as healthy people. Namely, it is possible to speculate systemic response patterns to diseases: Neuro-Endocrine-Immune Circuitry. In conclusion, although this study could not answer accurately the disease risk assessment, this study can provide data analysis scheme for the personal genome sequencing data. The scheme of this method has extendibility in genomic-based knowledge: drug-gene, environmental factor-gene and so on.

Keywords: next-generation sequencing, MeSH tree structure, disease risk, personal genome

Student number: 2004-23352

Contents

List of Figures

List of Tables

1. Introduction

1.1. Background

As a result of continuing advances in high-throughput sequencing technologies, whole-genome sequencing will soon become an affordable approach to identify all sequence variants in an individual human. Recent evidence suggests that each human genome has more than 3 million sequence variants, some common, some infrequent. Despite more comprehensive databases and better methods for the analysis of genetic variants, the problem of genome interpretation is still far from being solved. It is questionable that the scientists and physicians who first started talking about the $1,000 genome in 2001 could have imagined being on the verge of that achievement within the decade. As the cost of sequencing continues to freefall, the challenge of solving the data analysis and storage problems becomes more pressing. But those issues are nothing compared to the challenge facing the clinical community who are seeking to mine the genome for clinically actionable information.

Recently, whole-genome or whole-exome sequencing has been used to identify new disease predisposing variants in various familial disorders, such as familial pancreatic cancer and Miller syndrome. In particular, Ashley and colleagues made an impressive and ambitious effort to use full-genome sequence data in the

clinical setting [1]. They estimated a patient's risk of several common diseases using several pieces of information, including single nucleotide polymorphisms (SNPs) that have been associated with the risk of these diseases. In addition, now I can measure and digitize the entire genetic material of an individual, additional molecular phenotyping will be necessary to track down genetic effects in the genotype–phenotype chain and to discover relevant biomarkers for further personalization of diagnoses and therapeutics [2]. High-throughput experimentation technologies that give us insight into the transcriptomics, proteomics, metabolomics, and other biological aspects of an individual will have to mature further before they can be used in the clinic. Data from studies of disease concordance in monozygotic twins suggest that for many common diseases, such as cancer, a negative test result from whole genome sequencing (WGS) data would not appreciably reduce an individual's risk relative to the baseline population risk and would therefore not enable meaningful medical intervention [3].

The constant improvement in sequencing technologies makes it likely that a patient's whole genome will be incorporated into the clinical profile. When integrated with the electronic health records and analyzed using tools for automatic genome analysis, the full genetic information of patients will provide a strong foundation for the development of drugs and therapeutics tailored to specific genetic profiles. It will also enable hypothesis-free, large-scale population studies with enough power to reliably discern loci of interest with single-nucleotide resolution.

However, the potential utility of genome-wide sequencing for personalized medicine in the general population is unclear. In the following sections, I show the range of human genetic variation and its impact on genomic-guided therapies, describe current examples of translating whole-genome data into personalized diagnostics and therapeutics, and pinpoint challenges faced by the several disciplines involved in whole-genome clinical analysis.

1.2. Bioinformatic approach for interpreting personal genomes

1.2.1. Genetic variation resources

1) Sequence and structural variation databases

Several databases aid in the classification of variants as either known or novel, and rare or common (Table 1.1). The National Center for Biotechnology Information (NCBI) dbSNP database [4] is the largest source of short genetic variation data. dbSNP currently contains over 40 million, both common and rare human SNVs, short indels and microsatellites (Build 134, August 2011). Where available, the database also reports SNV clinical significance. The 1000 Genomes Project Consortium is a major contributor of novel variants to dbSNP, aiming to catalog 95% of human variants with an allele frequency of at least 1% in each of five major

human population groups. The Consortium is expanding on the work of the International HapMap Project [5] to catalog genetic variation shared within and between members of various populations. So far, over 38 million variant sites have been identified within the framework of this effort (Phase 1 Low Coverage Data, May 2011). In addition, the Consortium data includes inferred genotypes for individual samples, useful for future association studies utilizing genotype imputation.

While databases like dbSNP and HapMap and project like 1000 Genomes focus primarily on short-length variants like missense, nonsense and short insertion and deletion mutations (indels), larger-scale structural rearrangements, copy number variants (CNVs) and large indels can also dramatically affect human phenotypes. NCBI's database of genomic structural variation (dbVar) [6] and the collaborative effort Database of Genomic Variants (DGV) [7] are two of the largest repositories for large-scale (typically >1 kb in length) structural variations. DGV only contains entries from healthy human controls, while dbVar contains entries from all species and includes variants with associated phenotypes. The DGV archive (DGVa) [6] is a new database maintained by the European Bioinformatics Institute that also contains structural variants from all species with associated phenotypes when available.

2) Genotype/phenotype annotation databases

The Online Mendelian Inheritance in Man (OMIM) database [8] is a catalog of human genes and diseases. OMIM is manually curated and contains descriptions of over 13 000 genes and almost 7000 phenotypes (September 2011). Over 2600 genes in OMIM contain listings of specific allelic variants associated with disease.

The SwissVar database [9] is another manually curated source of variant–phenotype association data. The database also includes a number of variant features, e.g. physicochemical properties, affected functional features and conservation profiles for amino-acid changing variants in SwissProt proteins. SwissVar currently contains information on over 24 000 deleterious variants linked to over 3300 diseases (September 2011).

The Human Gene Mutation Database (HGMD) [10] is a large collection of variants associated with human inherited diseases. HGMD is available in two versions: a free version for academic/nonprofit users, and a more regularly updated, paid professional version. The free version of HGMD contains associations of approximately 82 000 variants of all kinds to approximately 3000 diseases (September 2011).

NCBI's ClinVar database, currently in development, aims to provide a freely available, comprehensive listing of variants associated with phenotypes along with links to regularly updated evidence for the associations.

3) GWAS and other association study databases

As previously noted, large-scale GWAS have identified thousands of variants associated with disease. The National Human Genome Research Institute Catalog of Published Genome-Wide Association Studies (NHGRI GWAS catalog) (www.genome.gov/gwastudies) conveniently lists significantly associated marker SNPs from these studies in a manually curated, online database. The catalog includes data on study designs, individual SNV P-values, odds ratios and links to the published studies. The Genetic Association Database (GAD) [11] predates the NHGRI GWAS catalog, and contains curated information on both positive and negative variant associations from GWAS and candidate gene association studies primarily from studies of common, complex diseases. In addition to the summary data in the GWAS catalog and GAD and the database of Genotype and Phenotype (dbGaP) [12] provide controlled access to individual-level genotype and phenotype data from many large-scale association studies.

4) Cancer gene and variant databases

Given the significance of somatic mutations in oncogenesis, several large-scale projects sequencing multiple cancer types have emerged including the Cancer Genome Atlas (TCGA) [13] and the Cancer Genome Project (CGP) (http://www.sanger.ac.uk/genetics/CGP). The International Cancer Genome

Consortium (ICGC) [14] was developed to coordinate cancer sequencing projects around the world, including TCGA and the CGP, for over 50 different cancer types and subtypes. Data portals for ICGC and TCGA are available to retrieve open access variant data, and individual level, controlled access genotype data by application. Data from the CGP and curated mutations from the literature for a list of genes previously associated with cancer (the Cancer Gene Census list [15]) are available from COSMIC, the Catalog of Somatic Mutations in Cancer [16]. COSMIC (Release 54) currently contains data on over 177 000 mutations from almost 620 000 tumors. The NCI Cancer Gene Index is another comprehensive source of genes related to cancer, containing gene–disease and gene–drug relationships text-mined and manually validated from over 20 million MEDLINE abstracts.

5) Pharmacogenomic gene and variant databases

Specialized databases also now exist to link genes and genotypes with drug targets and drug response. The Pharmacogenomics Knowledgebase (PharmGKB) [17] contains both manually curated and automatically text-mined associations of human variations to drug response. The database includes information for variants in over 1500 genes related to approximately 375 drugs and almost 300 diseases (September 2011). DrugBank [18] is a more drug-focused resource containing structural,

chemical and pharmacologic properties for over 6800 drugs (September 2011). DrugBank also contains the amino acid target sequences for individual drugs, enabling the identification of variants falling in drug binding sites.

6) Crowdsourcing model for gene and variant annotation

Many annotation databases use automated searches followed by expert human curation to identify and validate variant–disease associations from literature. As the pace of association studies continues to increase, this process will become increasingly unsustainable. To address this problem, several databases have been developed to harness a crowdsourcing model for gene and variant annotation including Gene Wiki [19], WikiGenes [20] and SNPedia (http://www.SNPedia.com). While all of these resources include some information automatically extracted from public sources like PubMed, OMIM and dbSNP, the community contribution and curation could potentially provide more comprehensive and update-to-date information as new studies are published.

7) Epigenome databases

Changes in gene regulation due to epigenetic mechanisms, other than variation in the DNA sequence, can also be disease associated. Local patterns of DNA methylation, chromatin structure and histone modification states, and nonprotein-

coding RNAs (ncRNA), e.g. microRNAs, affect gene expression levels. Thus, genome-wide studies to comprehensively catalog the various structural and functional elements of the genome, as well as studies to map the epigenetic elements affecting gene expression levels, are now being undertaken. These will lead to a better understanding of genome complexity and gene regulation. The ENCyclopedia of DNA Elements Project (ENCODE) [21] includes studies to catalog the full human transcriptome including protein coding, noncoding and pseudogene transcripts, in addition to local chromatin states and methylation patterns. The National Institutes of Health Roadmap Epigenomics Mapping Consortium [22] was recently organized to map DNA methylation, histone modifications, chromatin accessibility and ncRNA transcripts in each human tissue and cell type.

Table 1.1. Databases and resources for personal genome interpretation

Category		Database	Description	Ref.
Sequence and structural variation databases	Short variations-SNVs, Short indels	1000 Genomes	Human short variants and inferred genotypes	[23]
		dbSNP	Short variants from all species	[4]
		HapMap	Human short variants and population group haplotypes	[5]
	Structural variations-CNV, large indels	dbVar	Structural variants from all species	[6]
		DGV	Structural variants from healthy human controls	[7]
		DGVa	Structural variants from all species	[24]
Genotype/phenotype annotation databases		HGMD	Human variant-disease associations (inherited diseases)	[10]
		OMIM	Human variant-disease associations (includes extensive gene and phenotype descriptions)	[8]
		SwissVar	Human variant-disease associations (non-synonymous SNVs only)	[9]
GWAS and other association study databases		dbGaP	Controlled access to individual genotype/phenotype data from association studies	[12]
		GAD	Mainly complex disease SNVs from association studies	[11]
		NHGRI GWAS Catalog	Significant SNVs from GWAS	[25]
Cancer gene and variant databases		ICGC	Somatic variants from tumor sequencing projects	[26]
		COSMIC	Somatic variants from tumor sequencing and literature	[16]
		Cancer Gene	Comprehensive list of cancer-related genes	[15]

	Census		
	NCI Cancer Gene Index	Comprehensive list of cancer-related genes, including gene-disease and gene-drug relationships	
	TCGA	Somatic variants from tumor sequencing projects	[13]
Pharmacogenomic gene and variant databases	DrugBank	Drug properties and protein amino acid target sequences	[18]
	PharmGKB	Curated and text-mined variant-drug response associations	[27]
Crowdsourcing model for gene and variant annotation	Gene Wiki	Human gene/protein annotations	[19]
	SNPedia	Human SNV-disease associations	
	WikiGenes	Gene annotations from all species	[20]
Epigenome databases	ENCODE	Full human transcriptome including local chromatin states and methylation patterns	[21]
	NIH Roadmap Epigenome	DNA methylation, histone modifications, chromatin accessibility	[22]

1.2.2. Algorithms for the prediction of variant effects

1) Predicting the effect of nsSNVs

A non-synonymous or missense variant is a single base change in a coding region that causes an amino acid change in the corresponding protein. If a non-synonymous variant alters protein function, the change can have drastic phenotypic consequences. Most alterations are “deleterious” and so are eventually eliminated through purifying selection. Because non-synonymous single nucleotide

polymorphisms (nsSNPs) can affect protein function, they are believed to have the largest impact on human health compared with SNPs in other regions of the genome. It is important to distinguish those nsSNPs that affect protein function from those that are functionally neutral.

A computational method that could predict whether an amino acid substitution (AAS) affects protein function would help prioritize AASs (Table 1.2). The observation that disease-causing mutations are more likely to occur at positions that are conserved throughout evolution, as compared with positions that are not conserved, suggested that prediction could be based on sequence homology. It was also observed that disease-causing AASs had common structural features that distinguished them from neutral substitutions, suggesting that structure could also be used for prediction. In this study, I used prediction of the functional impact of a non-synonymous variant in order to extract "deleterious" variants in the personal genome.

Constraint-based approaches to annotate deleterious genomic positions assume that such positions will have a detectable history of purifying selection. By contrast, trained classifiers generate prediction rules by identifying heuristic combinations of many potentially relevant properties that optimally differentiate a set of true positives and negatives.

Table 1.2. Amino acid substitution prediction methods

Method	Algorithm	Type	Information	Ref.
SIFT	Based on sequence homology, scores are calculated using position-specific scoring matrices with Dirichlet priors.	Constraint-based predictor	Evolutionary, biochemical (indirect)	[28]
PolyPhen	Based on sequence conservation and structure, scores are calculated using position of amino acid substitution, and SWISS-PROT annotation.	Trained classifier	Evolutionary, biochemical and structural	[29]
SNPs3D	Structure-based support vector machine uses 15 structural factors. Sequence conservation-based support vector machine uses 5 sequence conservation features.	Trained classifier	Evolutionary, biochemical and structural	[30]
PMUT	Prediction provided by one of two neural networks. Neural network uses internal databases, secondary structure prediction, and sequence conservation.	Trained classifier	Evolutionary, biochemical and structural	[31]

2) Predicting the impact of genomic variants in noncoding regions

Until recently, the analysis of the effect of genetic variations strongly focused on those altering the protein sequence. The interpretation of genetic variants occurring in noncoding regions is also a challenging task. Although variants in noncoding regions may exhibit weaker effects than nsSNVs, it is evident that they constitute

the majority of human genetic variations [23, 32], and are also likely to be disease-associated; i.e. ~88% of weakly trait-associated variants from GWAS studies are noncoding [5]. Noncoding variants under purifying selection are five times more common than those in coding regions [33], and the detection of numerous regulatory variants with significant effect [10] has recently spurred interest in their computational annotation. Thus, a considerable number of methods are currently available to perform an evolutionary analysis of the nucleotide sequence to determine conserved regions across species. This approach, also applicable to protein sequences, is more complex for noncoding regions where there is no detectable conservation outside vertebrates [34]. The available algorithms for the detection of deleterious noncoding SNVs estimate the rate of evolution at the mutated position or consider a sliding window around the mutation site. Methods like binCons [35] and phastCons [36] implement a context dependent approach or a Hidden Markov model, in contrast to other algorithms such as GERP [37], SCONE [38] and Gumby [39] which calculate a position-specific score. This class of methods was also reviewed in a recent publication [40].

New approaches to predict the effect of mutations in noncoding regions focus specifically on genetic variations in regulatory regions and splicing sites. For example, Is-rSNP [41] uses a transcription factor position weight matrix and novel convolution methods to evaluate the statistical significance of the score. The RAVEN algorithm combines phylogenetic information and transcription factor

binding site prediction to detect variations in candidate cis-regulatory elements [42]. Recently, a new method including features associated with the mutated site and its surrounding region and gene-based features has been used for the identification of functional, regulatory SNVs involved in monogenic and complex diseases [43]. SNVs affecting splicing sites and their surrounding regions can be evaluated using Skippy [44]. In Table 3, I listed a selection of methods for the prediction of the effect of SNVs.

3) Integrated methods for variant annotation

The steps for interpreting the net effects of variants from an individual genome or from a disease association study have previously been performed one at a time: filtering out common polymorphisms, identifying known deleterious mutations, functionally annotating and predicting the effects of novel variants and prioritizing variants for experimental follow-up. A number of integrated tools are now emerging to automate various portions of this pipeline including ANNOVAR [45], the Ensembl Variant Effect Predictor [46], GAMES [47], SeqAnt [48], Sequence Variant Analyzer (SVA) [49] and MutationTaster [50]. Frameworks for storing patient data along with associated analysis tools like i2b2 [51] and caBIG [52], and workflow management systems like Galaxy [53] and Taverna [54] that can be installed and run 'on the cloud', are also now available to automate and

dramatically speed up variant annotation pipelines.

1.3. Issues in assessment of the risk of disease

An important benefit from the study of the genetics of human disease is to predict the risk that individuals may have a particular disease [55-57]. Recently, knowledge of this risk can be provided by the direct-to-consumer (DTC) genomics that enables consumers to interpret large numbers of genetic markers in their genomes [58, 59]. DTC personal genomic tests utilize high-throughput genotyping of 4500,000 bases of an individual's DNA and provide an individual with information about their genetic risk for between 20 to over 40 (depending on the company) complex diseases. There are a range of types of genetic tests that are currently offered DTC, as well as wide variation with respect to the extent to which each company uses the scientific literature to support their decisions concerning which tests they offer.

1.3.1. Type of data for genomic risk profiling

Genotyping for GWAS are currently performed under two genotyping platforms, Affymetrix and Illumina. 23andMe and deCODEme used DNA microarrays from Illumina, with the HumanHap 550+ Genotyping BeadChip and Human1MDuo

DNA Analysis BeadChip, respectively. Navigenics used Affymetrix's 6.0 array.

1) Affymetrix

Affymetrix is currently being used to analyze over 1.8 million SNPs on one chip of the two platforms. Affymetrix is the more cost-effective of the genotyping platforms using ~250 nanograms of DNA per genotype (http://affymetrix.com/index.affx). Affymetrix uses DNA microarrays for high-throughput SNP detection using photolithography to create site-specific primers that are attached to a silicon chip. Primers attach to specifically amplified DNA that is equivalent to a site-specific complement probe. For detection, a reporter (e.g., fluorescent molecule) attaches to either a photolithographic probe or an amplified DNA probe and is chemically released upon hybridization of the two oligo chains. The Affymetrix Genome-Wide Human SNP Array 6.0 allows for the selection of more than 906,600 SNPs from a combination of 482,000 unbiased "historical SNPs" selected by Affymetrix and a combination of 424,000 additional SNPs that consist of tags, mitochondrial SNPs, SNPs in recombination hotspots, SNPs on chromosomes X and Y, and new SNPs added to the dbSNP database. The remaining 946,000 markers consist of copy number variant (CNV) probes.

2) Illumina

The Illumina genotyping platform is currently able to analyze 1.1 million SNPs on an Illumina Human 1M-Duo BeadChip array (http://www.illumina.com). The Illumina system consists of a complex bead array that allows all SNPs to be detected at once. Much like Affymetrix, this system also relies on a fluorescence reporting mechanism, but with a locus specification step at the beginning of the process that creates a specifically addressed oligonucleotide chain that is then amplified by a process that is similar to whole genome amplification. SNPs for the Illumina platform are selected based on being tagSNPs in the populations from the International HapMap Project dataset. Additional SNPs are selected based on SNP coverage reference sequence (RefSeq) genes (within 10 kilo-bases) (http://www.ncbi.nlm.nih.gov/RefSeq/), nonsynonymous SNPs and SNPS found in the major histocompatibility complex (MHC) region. With regard to genomic coverage, Affymetrix and Illumina are comparable.

1.3.2. Measures to predict disease risks

Services report absolute risk, which is the probability that individuals will develop diseases. Absolute risk is derived from two parameters: ‘relative risk’ and ‘average population disease risk’. Relative risk is modeled from an individual’s genetics. Average population disease risk varies depending on how one defines the

population. Risk markers are determined from GWAS, which survey hundreds of thousands or millions of markers across control and disease patients. Yet no disease has an identical set of markers among the DTC companies because each company has its own criteria for accepting a genome-wide association result into its relative risk calculation.

1.4. Objectives

Sequencing an individual's genome has become economically feasible. Sequencing has an advantage over genotyping because it captures the full spectrum of an individual's variation and determines, rather than infers, a higher resolution of variants. In particular, it is possible that a more complete knowledge of disease-associated variants and their epistatic relationships would be able to reliably predict who will and who will not develop disease in the general population. Here, I focus on evaluating individual predisposition patterns to specific phenotypic traits given their genetic backgrounds from personal genome sequencing data (Figure 1.1).

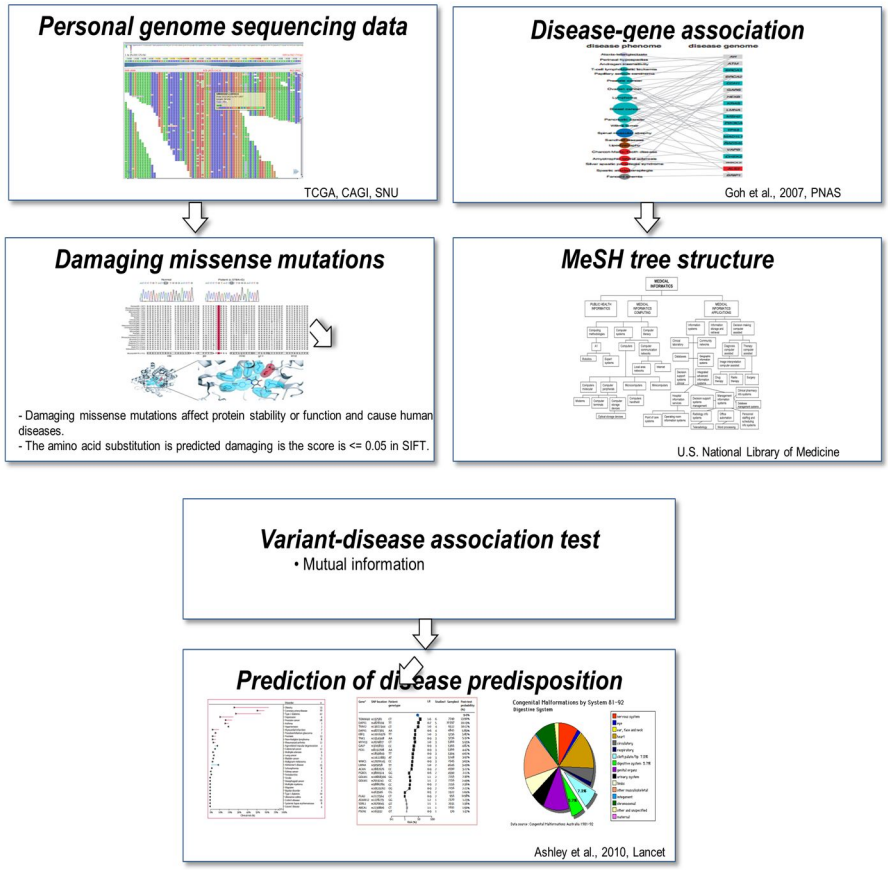

Figure 1.1. The summary of my approach. This approach is variant-disease association test based on disease hierarchy. Damaging missense mutations affect protein stability or function and cause human diseases. The amino acid substitution is predicted damaging is the score is <= 0.05, and tolerated if the score is > 0.05 in SIFT. Due to the damaging effects that mutations can have on genes, organisms have mechanisms such as DNA repair to prevent mutations.

2. Materials and Methods

2.1. Overview of methodology in this approach

The aim of this study is to predict disease predisposition patterns of the personal genome based on disease hierarchy. After obtaining variants from personal genome sequencing data, this method considers both damaging missense mutations and hierarchical disease structure. This method consists of the following steps (Figure 2.1):

Step 1. Predict damaging missense mutations using SIFT from the personal genome sequencing data. SIFT is useful in prioritizing changes that are likely to cause a loss of protein function.

Step 2. Extract disease-gene relations from OMIM that provides human disease gene entries associated with at least one mutant allele.

Step 3. Assign disease-gene relations to MeSH tree structure to search for disease predisposition patterns in the disease hierarchy.

Step 4. Apply disease-variant association test for the personal genome sequencing data using mutual information.

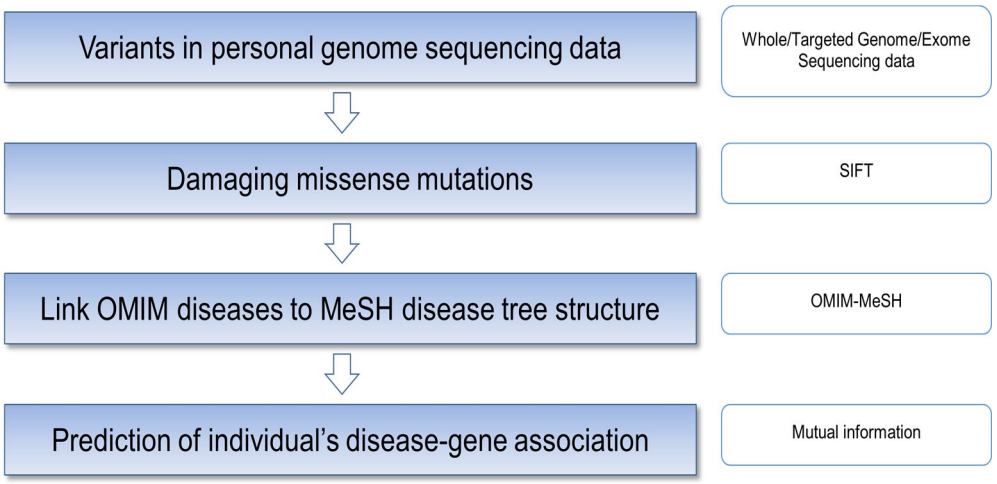

Figure 2.1. Steps for evaluating individual predispositions to phenotypic traits

Comparison of current methods with this study

The goal of population association studies is to identify patterns of polymorphisms that vary systematically between individuals with different disease states and could therefore represent the effects of risk-enhancing or protective alleles. Population association studies compare unrelated individuals, but 'unrelated' actually means that relationships are unknown and presumed to be distant. Therefore, I cannot trace transmissions of phenotype over generations and must rely on correlations of current phenotype with current marker alleles [60].

Broadly speaking, association studies are sufficiently powerful only for common causal variants. The threshold for 'common' depends on sample and effect

sizes as well as marker frequencies, but as a rough guide the minor-allele frequency might need to be above 5% [61, 62]. Arguments for the common-disease common-variant (CDCV) hypothesis essentially rest on the fact that human effective population sizes are small [61, 63].

The most important spurious cause of an association is population structure [64]. This problem arises when cases disproportionately represent a genetic subgroup, so that any SNP with allele proportions that differ between the subgroup and the general population will be associated with case or control status [65].

Linkage studies directly examine the transmission across generations of both disease phenotype and marker alleles within a known pedigree, seeking correlations that suggest that the marker is linked with a causal locus [66]. 'Tagging' refers to methods to select a minimal number of SNPs that retain as much as possible of the genetic variation of the full SNP set [67]. Tagging is only effective in capturing common variants. The desired cause of a significant result from a single-SNP association test is tight linkage between the SNP and a locus that is involved in disease causation [68].

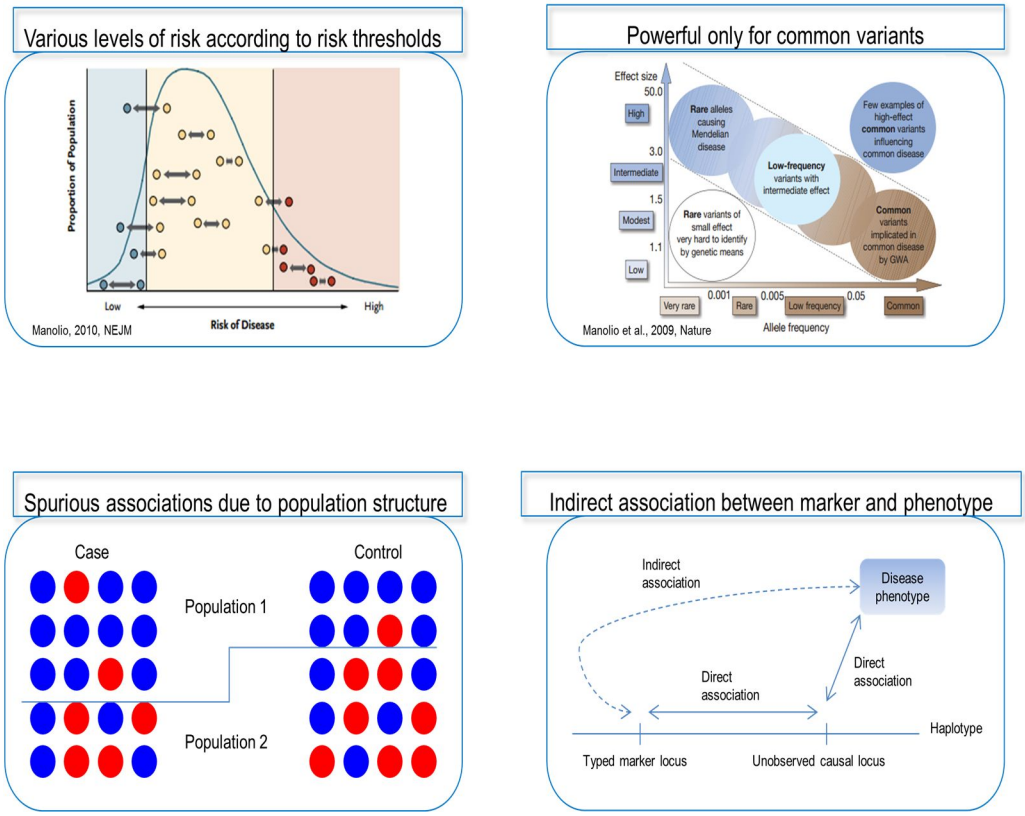

Figure 2.2. Information leaks in population-based association study. Several problems exist such as various levels of risk according to risk thresholds, spurious associations due to population structure and indirect association between marker and phenotype.

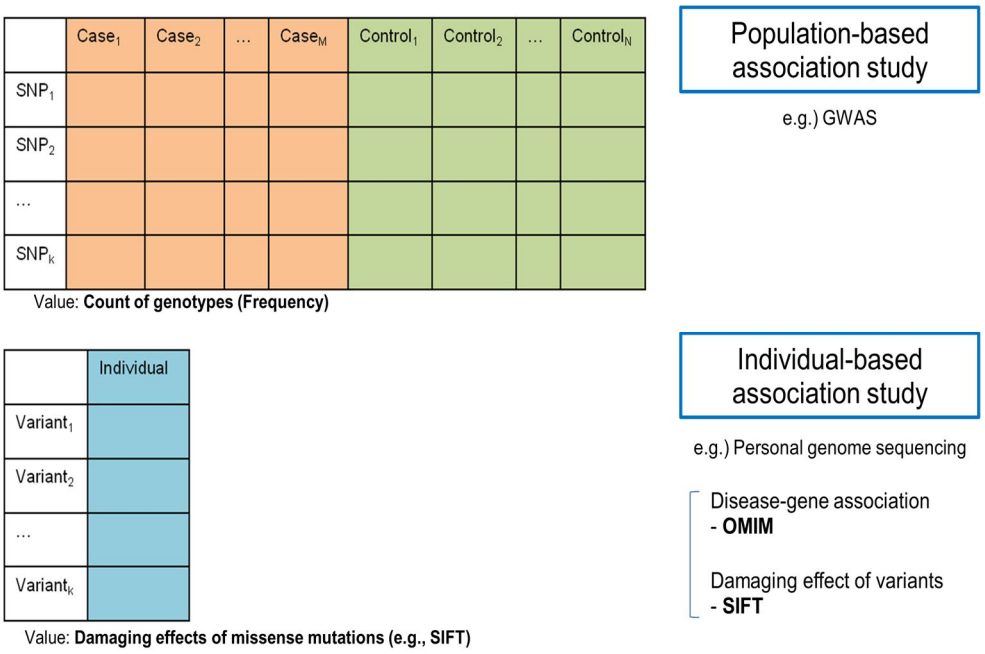

Figure 2.3. In population-based association study, experimental design is case-control phenotype and used values are count of genotypes and statistical tests are chi-square test or Fisher exact test. Whereas, in individual-based association study, damaging effects of missense mutations of personal genome sequencing data are used for disease association test.

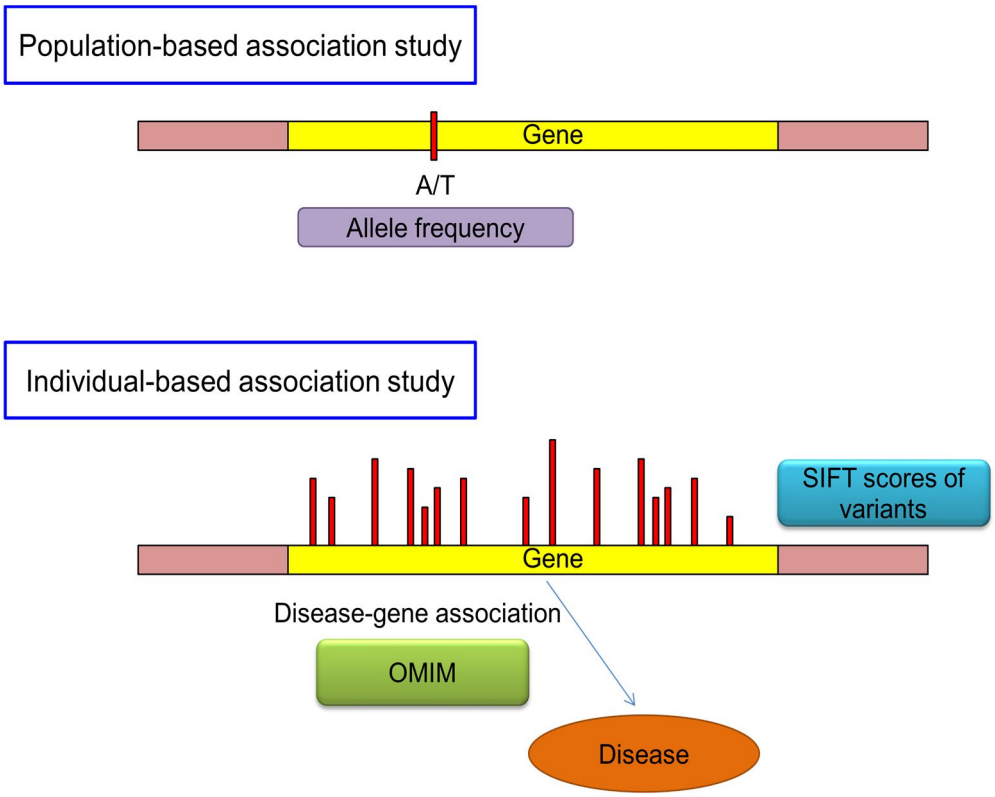

Figure 2.4. Comparison of population-based association study with individual-based association study. Population-based studies use allele frequency, while individual-based studies use SIFT scores of variants and knowledge of disease-gene association.

2.2. Data set

2.2.1. Personal genome sequencing data

I obtained whole/targeted genome/exome sequencing data for several diseases from various sources (Table 2.1). The Cancer Genome Atlas (TCGA) Genome Sequencing Centers (GSCs) performed large-scale DNA sequencing using the latest sequencing technologies [69-72]. Supported by the National Human Genome Research Institute (NHGRI) large-scale sequencing program, the GSCs generated the enormous volume of data required by TCGA, while continually improving existing technologies and methods to expand the frontier of what can be achieved in cancer genome sequencing. All sequencing data are available in the TCGA Data Portal or from the TCGA page at NIH's database of Genotype and Phenotype (dbGaP) [12]. TCGA targeted 601 genes, comprised of 7932 coding exons for the re-sequencing. Putative variants were identified using Polyphred [73], Polyscan [74], SNPdetector [75], and SNPCompare. Somatic mutations include missense and nonsense, splice site which is defined as within 2 bp of the splice junction, silent mutations.

The Critical Assessment of Genome Interpretation (CAGI) is a community experiment to objectively assess computational methods for predicting the phenotypic impacts of genomic variation. In this experiment, modeled on the Critical Assessment of Structure Prediction (CASP), participants will be provided

genetic variants and will make predictions of resulting molecular, cellular, or organismal phenotype. These predictions will be evaluated against experimental characterizations, and independent assessors will perform the evaluations. Community workshops will be held to disseminate results, assess our collective ability to make accurate and meaningful phenotypic predictions, and better understand progress in the field. The dataset includes exome sequence data (in the VCF format) from Crohn's 42 disease patients.

The catalogue of Somatic Mutations in Cancer (COSMIC) (http://www.sanger.ac.uk/cosmic/) is the largest public resource for information on somatically acquired mutations in human cancer and is available freely without restrictions [76-78]. The variant set obtained from COLO-829BL was subtracted from that of COLO-829 to establish the catalogue of somatic mutations in COLO-829 [79]. They identified 33,345 somatic base substitutions. Using massively parallel sequencing technology, a small-cell lung cancer cell line, NCI-H209 was sequenced. A total of 22,910 somatic substitutions were identified, including 134 in coding exons [80]. Twenty-four breast cancers were investigated by sequencing both ends of ~65,000,000 randomly generated ~500-bp DNA fragments. Paired-end sequence reads of 37 bp were generated on the Illumina Genome Analyser from 500-bp insert genomic libraries and aligned to the human reference genome (NCBI Build 36) using MAQ.

Table 2.1. A description of whole/targeted genome/exome sequencing data

Source	Disease	No. of patients	Data type	Platform
TCGA[1]	Acute myeloid leukemia	50	Targeted exome sequencing	Illumina GA[3]
TCGA	Bladder urothelial carcinoma	28	Targeted exome sequencing	Illumina GA
TCGA	Breast invasive carcinoma	60	Targeted exome sequencing	Illumina GA
TCGA	Cervical squamous cell carcinoma and endocervical adenocarcinoma	36	Targeted exome sequencing	Illumina GA
TCGA	Colon adenocarcinoma	11	Targeted exome sequencing	Illumina GA
TCGA	Cutaneous melanoma	129	Targeted exome sequencing	Illumina GA
TCGA	Glioblastoma multiforme	271	Targeted exome sequencing	Illumina GA
TCGA	Kidney renal clear cell carcinoma	35	Targeted exome sequencing	Illumina GA
TCGA	Lung adenocarcinoma	26	Targeted exome sequencing	Illumina GA
TCGA	Lung squamous cell carcinoma	30	Targeted exome sequencing	Illumina GA
TCGA	Ovarian serous cystadenocarcinoma	331	Targeted exome sequencing	Illumina Hiseq/ ABi SOLiD

TCGA	Prostate adenocarcinoma	82	Targeted exome sequencing	Illumina GA
TCGA	Rectum adenocarcinoma	4	Targeted exome sequencing	Illumina GA / ABi SOLiD
Pleasance et al.	Small Cell Lung Carcinoma	1	Whole genome sequencing	SOLiD
Pleasance et al.	Malignant melanoma	1	Whole genome sequencing	SOLiD
COSMIC[2]	Breast cancer	24	Whole genome sequencing	Illumina GA
CAGI[3]	Crohn's disease	42	Whole exome sequencing	Illumina GA
SNU	Focal segmental glomerulosclerosis	2	Whole exome sequencing	Illumina GA II
SNU	Retinitis pigmentosa	4	Whole exome sequencing	Illumina GA II

[1]TCGA: The Cancer Genome Atlas, [2]COSMIC: Catalogue of Somatic Mutations in Cancer, [3]CAGI: Critical Assessment of Genome Interpretation, [3]GA: Genome Analyzer

1000 Genomes Project

Recent improvements in sequencing technology (next-generation sequencing platforms) have sharply reduced the cost of sequencing. The 1000 Genomes Project is the first project to sequence the genomes of a large number of people, to provide a comprehensive resource on human genetic variation [81, 82].

The goal of the 1000 Genomes Project is to find most genetic variants that have frequencies of at least 1% in the populations studied [83]. This goal can be attained by sequencing many individuals lightly [84]. To sequence a person's genome, many copies of the DNA are broken into short pieces and each piece is

sequenced [85]. The many copies of DNA mean that the DNA pieces are more-or-less randomly distributed across the genome. The pieces are then aligned to the reference sequence and joined together. To find the complete genomic sequence of one person with current sequencing platforms requires sequencing that person's DNA the equivalent of about 28 times. If the amount of sequence done is only an average of once across the genome (1X), then much of the sequence will be missed, because some genomic locations will be covered by several pieces while others will have none. The deeper the sequencing coverage, the more of the genome will be covered at least once. Also, people are diploid; the deeper the sequencing coverage, the more likely that both chromosomes at a location will be included. In addition, deeper coverage is particularly useful for detecting structural variants, and allows sequencing errors to be corrected.

Sequencing is still too expensive to deeply sequence the many samples being studied for this project. However, any particular region of the genome generally contains a limited number of haplotypes. Data can be combined across many samples to allow efficient detection of most of the variants in a region. The Project currently plans to sequence each sample to about 4X coverage; at this depth sequencing cannot provide the complete genotype of each sample, but should allow the detection of most variants with frequencies as low as 1%. Combining the data from 2500 samples should allow highly accurate estimation (imputation) of the variants and genotypes for each sample that were not seen

directly by the light sequencing.

The samples for the 1000 Genomes Project mostly are anonymous and have no associated medical or phenotype data; for some of the populations the collectors have phenotype data but these data are not at Coriell and are not distributed. Although the 1000 Genomes samples have no phenotype data, the genetic variation data produced by the Project will be used by researchers to study many diseases, in sets of disease and control samples that have been carefully phenotyped.

Alignment

Sequence reads were aligned to the human genome reference sequence, The copy of the fasta file that is used for the alignments can be found on our ftp site here It currently represents the full chromosomes of the GRCh37 build of the human reference.

Differing alignment algorithms were used for each sequencing technology. For the Trio and Low Coverage pilots, Illumina data were mapped using the MAQ algorithm [86], SOLiD data were mapped using Corona Lite, and 454 data were mapped using SSAHA2 [87]. For the Exon pilot, Illumina and 454 reads were mapped using MOSAIK.

MAQ: http://maq.sourceforge.net/

Corona Lite: http://solidsoftwaretools.com/gf/project/corona/

SSAHA2: http://www.sanger.ac.uk/resources/software/ssaha2/

MOSAIK: http://code.google.com/p/mosaik-aligner/

Recalibration

All data from Illumina and 454 platforms has been recalibrated using the GATK package [88].

GATK:

http://www.broadinstitute.org/gsa/wiki/index.php/The_Genome_Analysis_Toolkit

SNP calling

Multiple SNP calling procedures have been used by the 1000 Genomes Project. In part this was because different methods were appropriate for different aspects of the Project, in part because different approaches were under active development by individual members of the Project, and in part because I found empirically that, given the state of current methods, the consensus of multiple primary call sets from different methods proved of higher quality than any of the primary call sets themselves.

Low Coverage SNP Calling

Calling was separate for the three analysis panels, CEU, YRI and CHB+JPT. For each, three primary SNP call sets were generated, from the Broad Institute (Broad), the University of Michigan (Michigan) and the Sanger Institute (Sanger). Details of the methods are given in separate publications [89-91]. All three were produced by a two step processes, involving in the first step a set of candidate calls made based on the evidence at each base pair in the reference, independent of neighboring sites. Then in each case a more computationally intensive linkage disequilibrium (LD)/imputation based approach was used to refine the call set and genotypes for all individuals.

Exon SNP calling

The Exon Pilot SNP call set was composed of the intersection of SNP calls made using the GATK Unified Genotyper essentially as described above for low coverage and trio calls, with calls from an alternate pipeline using the MOSAIK read mapper and the GigaBayes SNP caller. Note that unlike the low coverage and trio pilots, the different SNP call sets started with different read alignments. SNP calls were made separately for each of the 7 Exon Project populations.

Table 2.2. 1000 Genome Project Dataset

Category	Population	Abbrev.	No. of samples
American	Colombian in Medellin, Colombia	CLM	60
American	Mexican Ancestry in Los Angeles, CA	MXL	66
American	Puerto Rican in Puerto Rico	PUR	55
East Asian ancestry	Han Chinese in Beijing, China	CHB	97
East Asian ancestry	Han Chinese South	CHS	100
East Asian ancestry	Japanese in Toyko, Japan	JPT	89
European ancestry	Utah residents (CEPH) with Northern and Western European ancestry	CEU	85
European ancestry	Finnish from Finland	FIN	93
European ancestry	British from England and Scotland	GBR	89
European ancestry	Iberian populations in Spain	IBS	14
European ancestry	Toscani in Italia	TSI	98
West African ancestry	African Ancestry in Southwest US	ASW	61
West African ancestry	Luhya in Webuye, Kenya	LWK	97
West African ancestry	Yoruba in Ibadan, Nigeria	YRI	88

2.2.2. Database for predicted functional impact of non-synonymous variants

The SIFT (Sorting Intolerant From Tolerant) algorithm is based on the assumption that amino-acid positions that are important for the correct biological function of the protein are conserved across the protein family and/or across evolutionary history. SIFT uses protein sequence and multiple alignment information of these sequences to estimate 'tolerance indices' that predict tolerated and deleterious (that is, intolerant) substitutions for every position of the query sequence. Substitutions at each position with normalized tolerance indices that are below a chosen cut-off point are predicted to be deleterious. Substitutions that are greater than or equal to the cut-off point are predicted as being tolerated (that is, putatively non-functional). Here, in order to predict the impact of amino acid substitutions, I used SIFT algorithm.

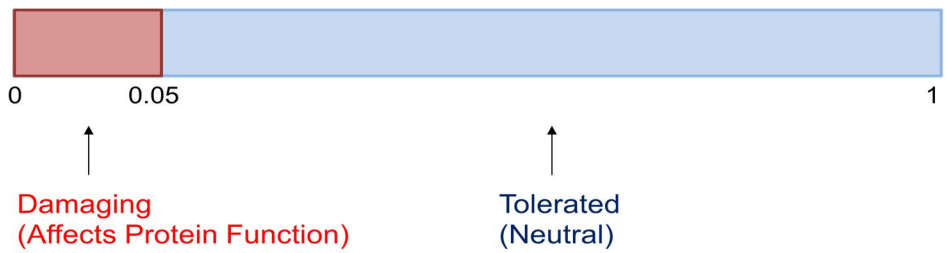

Figure 2.5. Definition of SIFT score. SIFT predicts whether an amino acid substitution affects protein function based on sequence homology and the physical properties of amino acids. SIFT can be applied to naturally occurring nonsynonymous polymorphisms and laboratory-induced missense mutations.

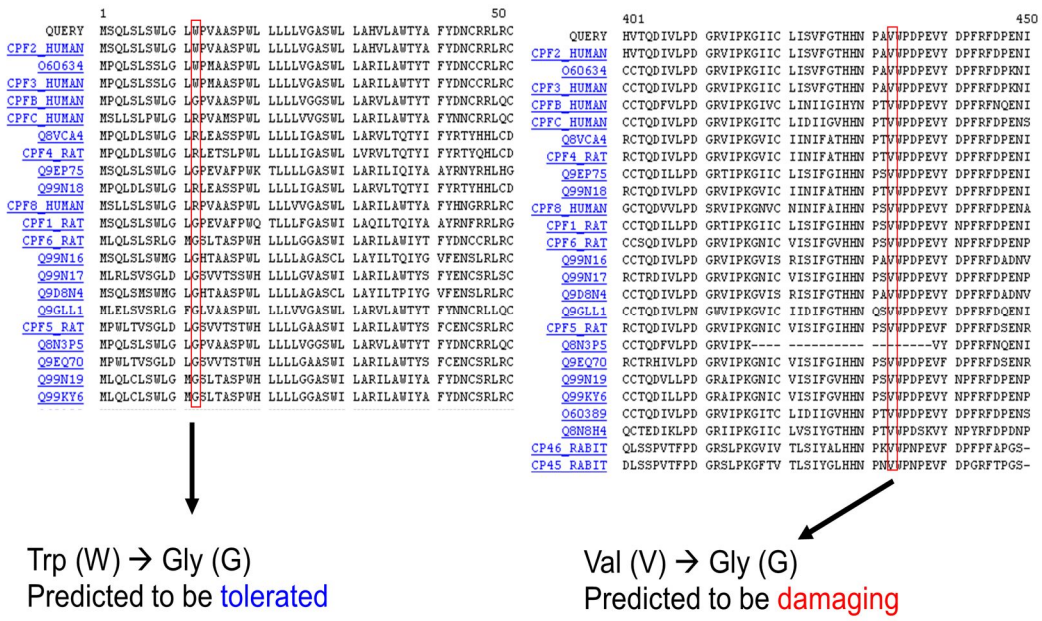

Figure 2.6. Example of SIFT score. SIFT Coding Changes in CYP4F2:cytochrome P450, subfamily IVF, polypeptide 2

2.2.3. Disease-gene association database

It is imperative to use disease-gene association databases in order to query association data in a systematic manner. To collect high-quality data of high significance and the standard in genetic association study data, the list of disorders, disease genes, and associations between them was obtained from the OMIM, a compendium of human disease genes and phenotypes. As of November 2011, this list contained 5,911 disorders and 2,721 disease genes. OMIM initially focused on monogenic disorders and has only recently expanded to include complex traits and the genes mutations of which confer susceptibility for these common disorders, so the current disorder-disease gene associations are biased towards those transmitted in a Mendelian manner. Despite such potential biases and the evident incompleteness of the disease association records, OMIM represents the most up-to-date and reliable repository of known disease genes and the disorders they confer.

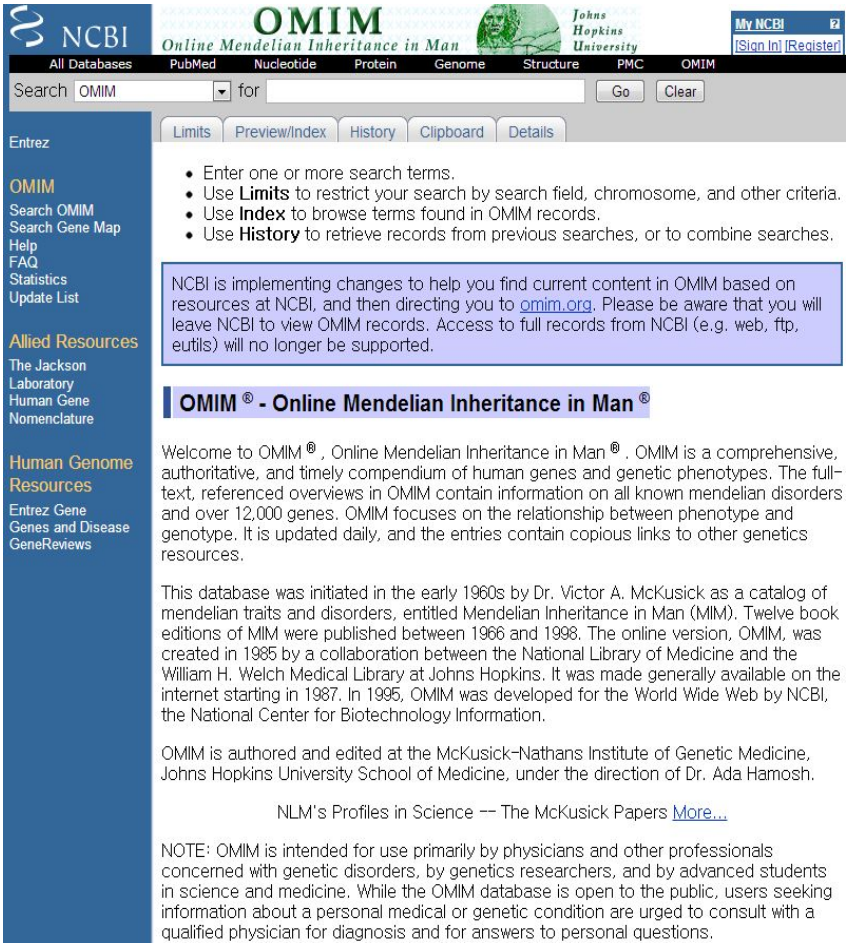

Figure 2.7. Disease Database: Online Mendelian Inheritance in Men (OMIM)[8]

2.2.4. MeSH disease tree structure

The Medical Subject Headings (MeSH) thesaurus is a controlled vocabulary produced by the National Library of Medicine (NLM) and used for indexing, cataloging, and searching for biomedical and health-related information and documents. MeSH descriptors are organized in 16 categories: category A for anatomic terms, category B for organisms, C for diseases, D for drugs and chemicals and so on. Each category is further divided into subcategories. Within each subcategory, descriptors are arrayed hierarchically from most general to most specific in up to twelve hierarchical levels. Each MeSH descriptor showed a structure of a hierarchical directed acyclic graph (DAG). All nodes in the DAG are connected by a direct edge from a more general term, I call it parent, to a more specific term, and I call it child. I downloaded the disease tree file mtree2012.txt from MeSH website (http://www.nlm.nih.gov/mesh/). I then obtained the relationship of various diseases based on the disease DAG from the MeSH descriptor of Category C and F 03.

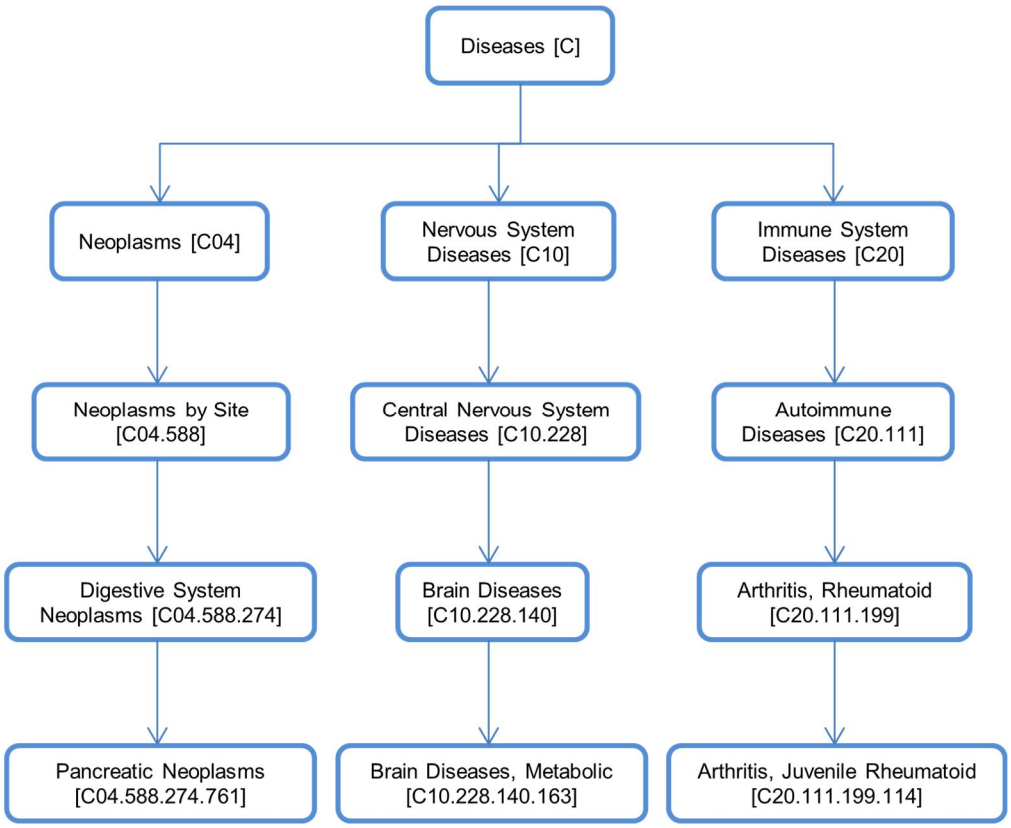

Figure 2.8. Example of MeSH disease tree structure

2.3. Measuring similarity between the personal genome and diseases

2.3.1. Construction of personal genome vectors using damaging missense mutations

SIFT score measures the tolerance of a substitution based on the mutability of the substitution position [28]. To assess the effect of a substitution, SIFT assumes that important positions in a protein sequence have been conserved throughout evolution and therefore substitutions at these positions may affect protein function. Thus, by using sequence homology, SIFT predicts the effects of all possible substitutions at each position in the protein sequence.

Let X_j denote the probability of a perturbed or altered function for a *j*-th gene as an average of SIFT scores for all the variations in the gene (N_j = number of variants in a *j*-th gene) of the estimated probability that affect protein function, *SIFT score* [28].

$$X_j = 1 - \frac{\sum_{i=1}^{N_j} SIFT\ score_i}{N_j}$$

The length of the personal genome vector is the total number of genes in the Online Mendelian Inheritance in Man (OMIM) [8].

2.3.2. Generation of disease vectors using disease-gene associations

The list of disorders, disease genes, and associations between them was obtained from the OMIM, a compendium of human disease genes and phenotypes. As of November 2011, this list contained 5,911 disorders and 2,721 disease genes. To organize the disease features referred to in OMIM, I attempted to use the Medical Subject Headings (MeSH) controlled vocabulary. The extracted disease names from OMIM were mapped to the MeSH terms in two successive term matching steps. First, I looked for exact matches, where all words composing the name had an identical correspondent in a MeSH term and vice versa. The word order and the case were not taken in consideration. When this step failed, I looked for partial matches by at least two words.

The MeSH tree contains a finite set of MeSH codes, M. A specific variant, v $\in$ V, is associated with zero, one or more MeSH codes, *i.e.* forms a set $M_v = \{m: \mathrm{annot}(v, m) \cap m \in M\}$, where the predicate annot() pairs variants with their MeSH codes. The disease vector is a binary vector indicating disease-association or non-disease-association for each gene in the MeSH tree. Let *Y* denote the binary disease-association value. For instance, *Y*=0 for a non-disease-association, *Y*=1 for disease-association for the corresponding gene. The length of the personal genome vector is also the total number of genes in the OMIM.

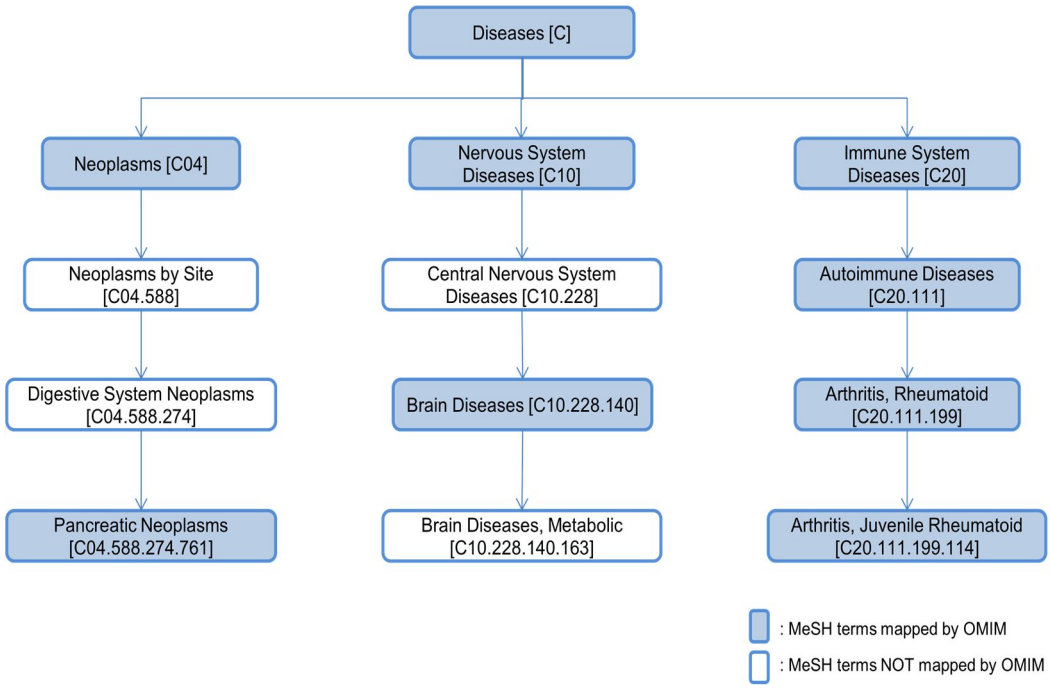

Figure 2.9. The OMIM to MeSH mapping

Tree extension For a given gene, and for each disease node *i* in a MESH tree, let X_i be a binary variable indicating whether the disease is associated with the gene or not. Our goal is to compute

$$Y_i = P(X_i = 1)$$

Let P_i denote the parent node of *i* and let C_{ij} for $j = 1,\ldots,n$ denote each child node of *i*. I define the prior distribution for X_i as follows:

$$P(X_i|X_{Pi} = 1) = p^{I(X_i=1)}(1-p)^{I(X_i=0)}$$

$$P(X_i|X_{Pi} = 0) = I(X_i = 0)$$

where I(·) represents the binary indicator function and p is a parameter reflecting the prior belief about the parent-conditional disease-gene association $P(X_i = 1 \mid X_{Pi} = 1)$.

Then the conditional distribution of X_i given its neighbors is defined as follows:

Case 1: $X_{P_i} = 0$

$$P\left(X_i|X_{Pi} = 0, X_{Ci_1}, \ldots, X_{Ci_k}\right) = I(X_i = 0)$$

Therefore, $Y_i = P(X_i=1|X_{pi}=0, X_{Ci1}, \ldots, X_{Cin}) = 0$

Case 2: $X_{P_i} = 1$

$$P\left(X_i | X_{Pi} = 1, X_{Ci_1}, \ldots, X_{Ci_k}\right) \propto P(X_i | X_{Pi} = 1) P\left(X_{Ci_1}, \ldots, X_{Ci_k} | X_{Pi} = 1, X_i\right)$$

$$P\left(X_i | X_{Pi} = 1, X_{Ci_1}, \ldots, X_{Ci_k}\right) \propto p^{I(X_i=1)}(1-p)^{I(X_i=0)} \prod_j P\left(X_{Ci_j} \middle| X_i\right)$$

$$P\left(X_i = 1 | X_{Pi} = 1, X_{Ci_1}, \ldots, X_{Ci_k}\right)$$

$$\propto p \prod_j P\left(X_{Ci_j} \middle| X_i = 1\right) \propto p \prod_j p^{I(X_{Ci_j}=1)}(1-p)^{I(X_{Ci_j}=0)}$$

$$P\left(X_i = 1 | X_{Pi} = 0, X_{Ci_1}, \ldots, X_{Ci_k}\right) \propto p^{1+n(X_{Ci_j}=1)}(1-p)^{n(X_{Ci_j}=0)}$$

$$P\left(X_i = 0 | X_{Pi} = 1, X_{Ci_1}, \ldots, X_{Ci_k}\right)$$

$$\propto (1-p) \prod_j P\left(X_{Ci_j} \middle| X_i = 0\right) \propto (1-p) \prod_j I(X_{Ci_j} = 0)$$

Therefore, if $X_{Ci_j} = 1$ for some *j*, then $P\left(X_i = 0 | X_{Pi} = 1, X_{Ci_1}, \ldots, X_{Ci_k}\right) = 0$, and hence $Y_i = P\left(X_i = 1 | X_{Pi} = 1 | X_{Pi} X_{Ci_1}, \ldots, X_{Ci_k}\right) = 0$

If $X_{Ci_j} = 0$ for all *j*, then $n = n(X_{Ci_j} = 0)$, and thus

$$Y_i = P\left(X_i = 1 | X_{Pi} = 1, X_{Ci_1} = 0, \ldots, X_{Ci_k} = 0\right)$$

Y_i

$$= \frac{P(X_i = 1 | X_{Pi} = 1, X_{Ci_1} = 0, \dots, X_{Ci_k} = 0)}{P(X_i = 1 | X_{Pi} = 1, X_{Ci_1} = 0, \dots, X_{Ci_k} = 0) + P(X_i = 0 | X_{Pi} = 1, X_{Ci_1} = 0, \dots, X_{Ci_k} = 0)}$$

$$Y_i = \frac{p^{1+n(X_{Ci_j}=1)}(1-p)^{n(X_{Ci_j}=0)}}{p^{1+n(X_{Ci_j}=1)}(1-p)^{n(X_{Ci_j}=0)} + (1-p)\prod_j I(X_{Ci_j} = 0)}$$

$$Y_i = \frac{p(1-p)^n}{p(1-p)^n + (1-p)} = \frac{p(1-p)^{n-1}}{p(1-p)^{n-1} + 1} \text{ (if } n \geq 1) \text{or } \frac{p}{1-p+p} = p(\text{if } n = 0)$$

Note that $Y_i \to 0$ as $n \to \infty$ and this makes intuitive sense because as the more number of children nodes has zero value, the less likely the node has value 1. Now, I assume the joint distribution of variables X_{Pi} , X_{cij} is of the following form:

$$P(X_{Pi}, X_{Ci_1}, \dots, X_{Ci_n}) \propto P(X_{Pi})P(X_{Ci_1}), \dots, P(X_{Ci_n})J(X_{Pi}, X_{Ci_1}, \dots, X_{Ci_n})$$

where $J(X_{Pi}, X_{Ci_1}, \dots, X_{Ci_n})$ is an indicator function whether the values are consistent with the hierarchical relation between ancestor and descendant nodes. The joint probability $P(X_{Pi}, X_{Ci_1}, \dots, X_{Ci_n})$ is non-zero only when $X_{Pi} = 1$ or $X_{Pi} = X_{Ci_1} = \dots = X_{Ci_n} = 0$.

Therefore,

$$P\left(X_{Pi}, X_{Ci_1}, \ldots, X_{Ci_n}\right) = \frac{P(X_{Pi})P\left(X_{Ci_1}\right), \ldots, P\left(X_{Ci_n}\right)J(X_{Pi}, X_{Ci_1}, \ldots, X_{Ci_n})}{P(X_{Pi} = 1) + (1 - P(X_{Pi} = 1))\prod_j(1 - P(X_{Ci_j} = 1))}$$

and hence

$$P\left(X_{Pi} = 1, X_{Ci_1} = 0, \ldots, X_{Ci_n} = 0\right) = \frac{P(X_{Pi} = 1)\prod_j(1 - P(X_{Ci_j} = 1))}{P(X_{Pi} = 1) + (1 - P(X_{Pi} = 1))\prod_j(1 - P(X_{Ci_j} = 1))}$$

Finally, I can compute the marginal probability of P(Xi = 1) as follows:

$$P(X_i = 1) = P\left(X_i = 1 | X_{Pi} = 1, X_{Ci_1} = 0, \ldots, X_{Ci_n} = 0\right)P\left(X_{Pi} = 1, X_{Ci_1} = 0, \ldots, X_{Ci_n} = 0\right) + \sum_{X_{Ci_1}, \ldots, X_{Ci_n}} P\left(X_i = 1 | X_{Pi} = 1, X_{Ci_j} = 1 \text{ for at least one j}\right) P\left(X_{Pi} = 1, X_{Ci_j} = 1 \text{ for at least one j}\right)$$

$$P(X_i = 1) = P\left(X_i = 1 | X_{Pi} = 1, X_{Ci_1} = 0, \ldots, X_{Ci_n} = 0\right)P\left(X_{Pi} = 1, X_{Ci_1} = 0, \ldots, X_{Ci_n} = 0\right) + \sum_{X_{Ci_1}, \ldots, X_{Ci_n}} P\left(X_{Pi} = 1, X_{Ci_j} = 1 \text{ for at least one j}\right)$$

$$P(X_i = 1) = P\left(X_i = 1 | X_{Pi} = 1, X_{Ci_1} = 0, \dots, X_{Ci_n} = 0\right) P\left(X_{Pi} = 1, X_{Ci_1} = 0, \dots, X_{Ci_n} = 0\right) + P(X_{Pi} = 1) - P(X_{Pi} = 1, X_{Ci_j} = 0 \text{ for at least one j})$$

$$P(X_i = 1) = P(X_{Pi} = 1) - \left(1 - \frac{p(1-p)^n}{p(1-p)^n + (1-p)}\right) P\left(X_{Pi} = 1, X_{Ci_j} = 0 \text{ for all j}\right)$$

$$P(X_{Pi} = 1) - \frac{(1-p)}{p(1-p)^n + (1-p)} \frac{P(X_{Pi} = 1) \prod_j (1 - P(X_{Ci_j} = 1))}{P(X_{Pi} = 1) + (1 - P(X_{Pi} = 1)) \prod_j (1 - P(X_{Ci_j} = 1))}$$

In summary,

1. When Y_{Pi} and Y_{Cij} for j = 1,..., n are available, I have

$$Y_i = Y_{Pi} - \frac{(1-p)}{p(1-p)^n + (1-p)} \times \frac{Y_{Pi} \prod_j (1 - Y_{Ci_j})}{Y_{Pi} + (1 - Y_{Pi}) \prod_j (1 - Y_{Ci_j})}$$

2. If X_{Pi} is not available and only X_{cij} are available, I can derive the followings:

$$Y_i = 1 - \frac{(1-p1)}{p1(1-p)^n + (1-p1)} \times \prod_j \left(1 - Y_{Ci_j}\right)$$

Here, p1 is another parameter specifying the prior probability for $X_i = 1$ and *n* is the number of children. Note that if $Y_{Cij} = 1$ for some j, this naturally induces $Y_i = 1$.

3. If X_{Pi} is available only, then

$$Y_i = pY_{Pi}$$

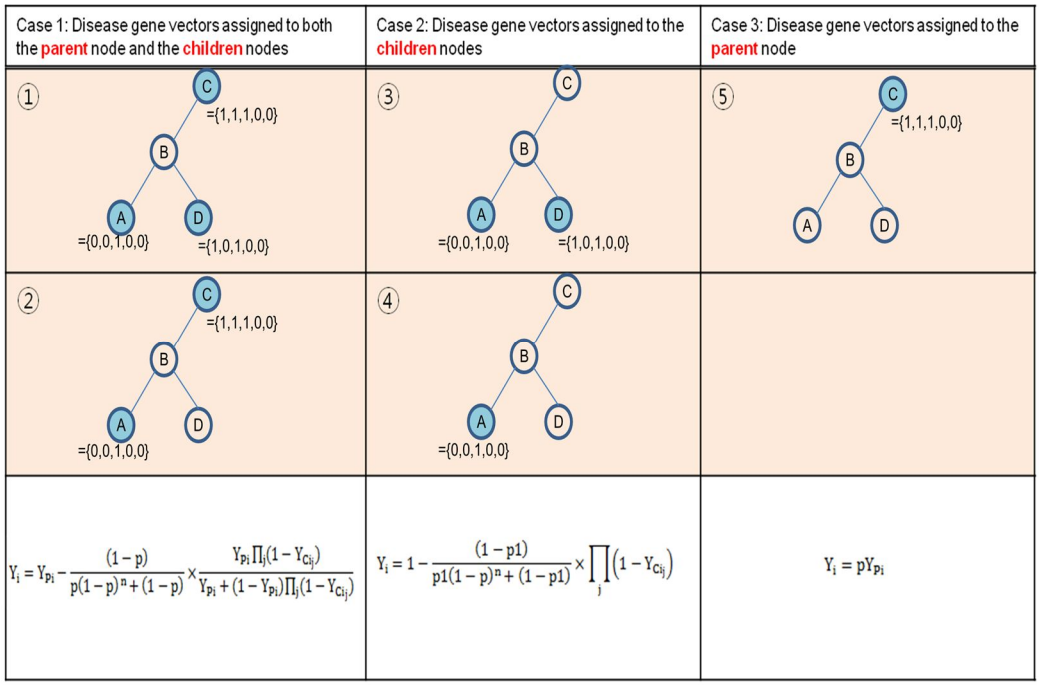

Figure 2.10. Extension of disease-gene association based on tree triad

2.3.3. Measuring similarity between the personal genome and diseases

In order to calculate similarity between the personal genome and diseases, I used the mutual information. Formally, the mutual information of the personal genome vector X and the disease vector Y can be defined as:

$$I(X;Y) = \sum_{y \in Y} \sum_{x \in X} p(x,y) \log_2 \left(\frac{p(x,y)}{p(x)p(y)} \right)$$

where p(x,y) is the joint probability distribution function of X and Y, p(x) is the marginal probability distribution function of X, and p(y) is the marginal probability function of Y. The higher mutual information between the personal genome vectors and the disease vectors means that the personal genome is more similar to the disease. A mutual information at zero means that the joint distribution of personal genome-disease association holds no more information than the personal genome-disease relationship considered separately.

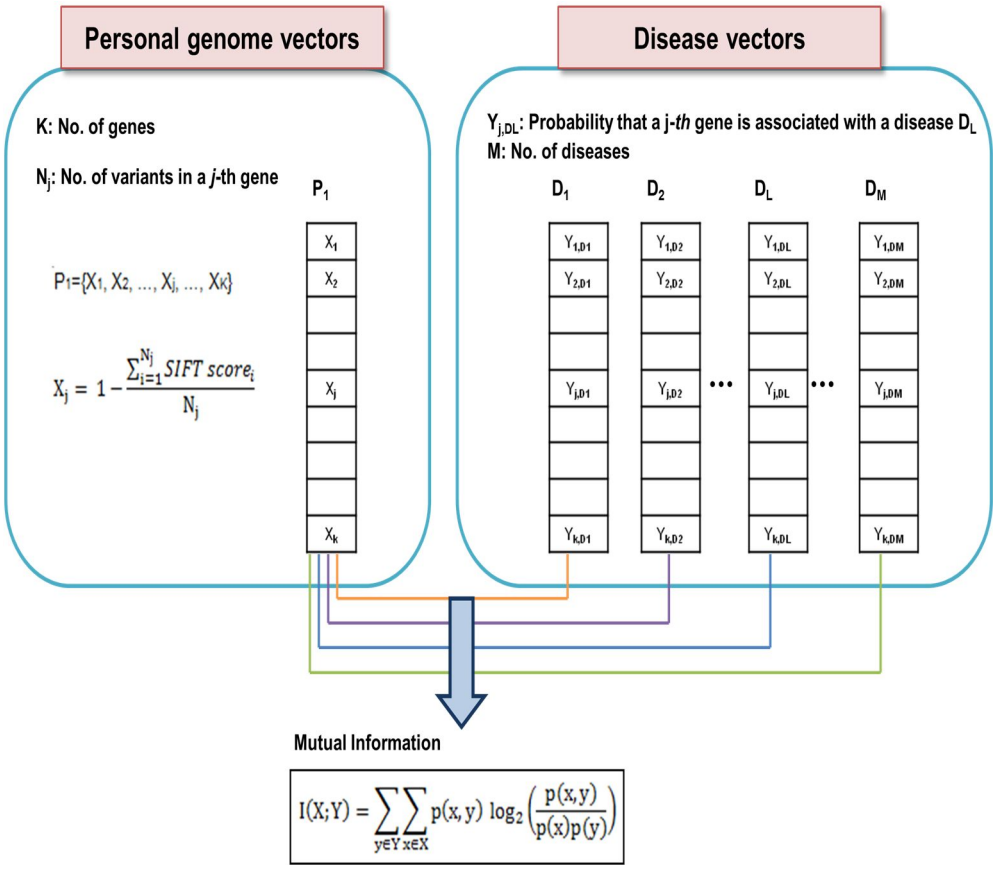

Figure 2.11. Calculating mutual information between the personal genome and diseases

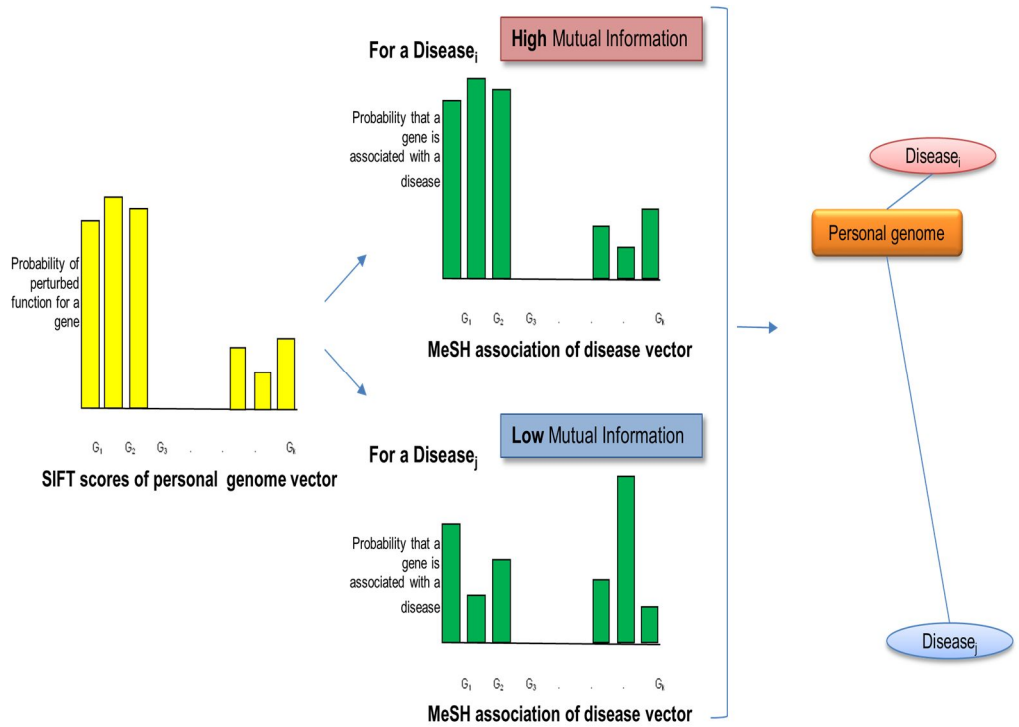

Figure 2.12. Discovering association patterns based on mutual information

2.3.4. Ranking diseases based on MeSH tree structure

The MeSH tree structure is a hierarchy. MeSH vocabulary includes four types of terms: Headings, Subheadings, Supplementary Concept Records, and Publication Types. A MeSH heading represents a concept found in the biomedical literature. Headings are positioned in the tree according to their relationship to other headings. Look at the heading "Eye" in the Body Regions branch under "Anatomy." Notice that the headings, "Eyebrows" and "Eyelids" are indented under "Eye" because these specific parts of the eye are narrower in scope. Also, see that the more specific heading, "Eyelashes" is indented quite logically under "Eyelids."

So, I assume that similarity of a disease in the MeSH tree with the personal genome includes similarity of diseases corresponding children nodes in the MeSH tree of the disease considering hierarchical structure of the MeSH tree. Ranking diseases based on MeSH tree structure consists of three-steps. First, I measure similarity between the personal genome and each disease in the MeSH tree using mutual information. Second, at each disease, I calculate average mutual information including each disease and diseases corresponding children nodes in the MeSH tree of the disease. Let *Children(Di)* denote diseases corresponding children nodes in the MeSH tree of the disease D_i, i.e.

$Children(D_i) = \{D_{i,1}, D_{i,2}, \ldots, D_{i,k_i}\}$, k_i = No. of children nodes in the MeSH tree of the disease D_i. Using children nodes in the MeSH tree of the disease,

MI_{D_i} as mutual information of the disease D_i is averaged to

$$MI_{D_i} = \frac{\sum_{j=1}^{k_i} MI_{D_{i,j}}}{k_i}$$

Finally, I rank diseases using average mutual information based on hierarchical relationships in the MeSH Tree Structures.

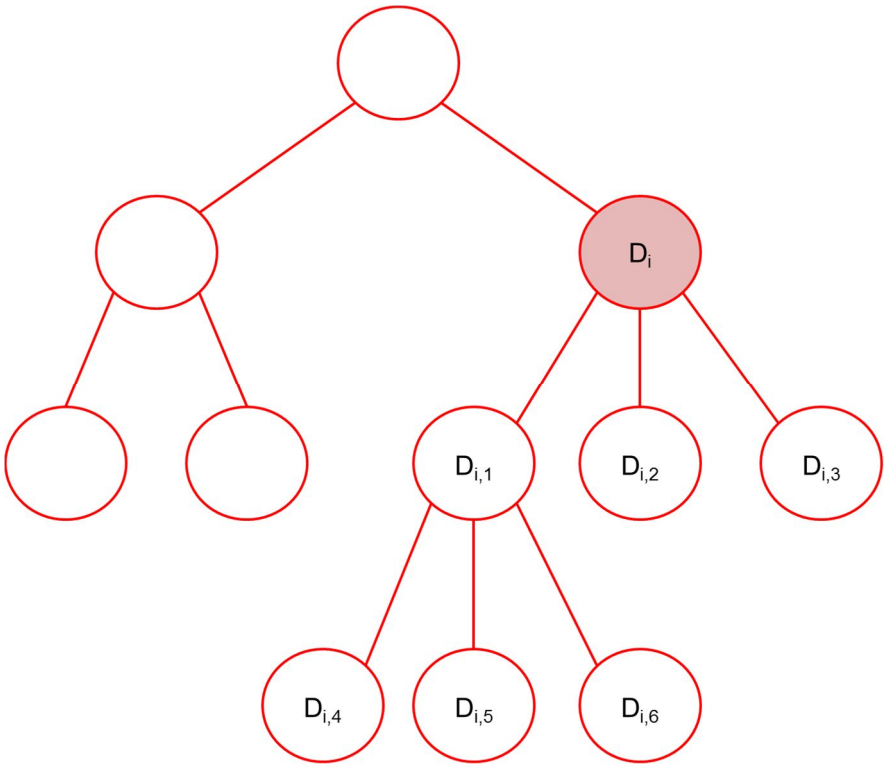

Figure 2.13. Ranking diseases based on MeSH tree structure

2.3.5. Disease enrichment analysis

The disease ontology term for disease enrichment analysis was used Diseasome. The significance of each disease association with the personal genome was evaluated using Fisher's exact test. Some notation need to be defined:

G: Total gene number in the Diseasome

C: Gene number in the personal genome

T: Number of genes which enriched by disease A over all the genes

t : Number of genes which enriched by disease A over the genes in the personal genome

With the above notation, the p-value of a specific disease enriched in the personal genome can be calculated by:

$$p = \sum_{x=t}^{\min(T,C)} \frac{\binom{T}{x}\binom{G-T}{C-x}}{\binom{G}{C}}$$

Table 2.3. Disease category of "diseasome" in the Goh's paper [92]

Category	No. of genes in the category	No. of the category-specific genes
Bone	43	23
Cancer	209	167
Cardiovascular	96	60
Connective	52	26
Dermatological	83	67
Developmental	53	31
Ear,Nose,Throat	44	33
Endocrine	96	63
Gastrointestinal	34	24
Hematological	149	112
Immunological	119	87
Metabolic	288	238
Muscular	70	48
Neurological	258	207
Nutritional	23	19
Ophthalmological	120	100
Psychiatric	30	19
Renal	59	43
Respiratory	34	17
Skeletal	57	38

3. Results

3.1. Reconstruction of MeSH tree by mapping OMIM disease annotation

To generate OMIM-MeSH associations, I allow searching for terms of each OMIM entry in MeSH Headings because MeSH headings are not assigned to OMIM entries directly. Among total number of 4668 MeSH Headings, 25.54% have OMIM associations (Table 3.1, Figure 3.1). Because each MeSH term has one or more descriptor code numbers (MeSH Tree Numbers), 32.36% of MeSH codes have OMIM associations.

Table 3.1. Statistics of OMIM-MeSH association entry

	Disorders in OMIM Morbid Map	MeSH Headings	MeSH Codes
Total Count	5,911	4,668	11,407
Mapped Count	4,385	1,192	3,691
Percentage of mapping	74.18%	25.54%	32.36%

The 2,924 OMIM disease genes are mapped into MeSH Headings (Figure 3.2). *TP53* gene belongs to 21 MeSH Headings (Table 3.2). The tumor suppressor gene TP53 (OMIM no. 191117) encodes a transcription factor which is activated in response to several forms of cellular stress and exerts multiple, anti-proliferative functions [93]. Somatic TP53 gene alterations are frequent in most human cancers and germline TP53 mutations predispose to a wide spectrum of early-onset cancers (Li-Fraumeni (LFS) and Li-Fraumeni-like syndromes (LFL)) [94].

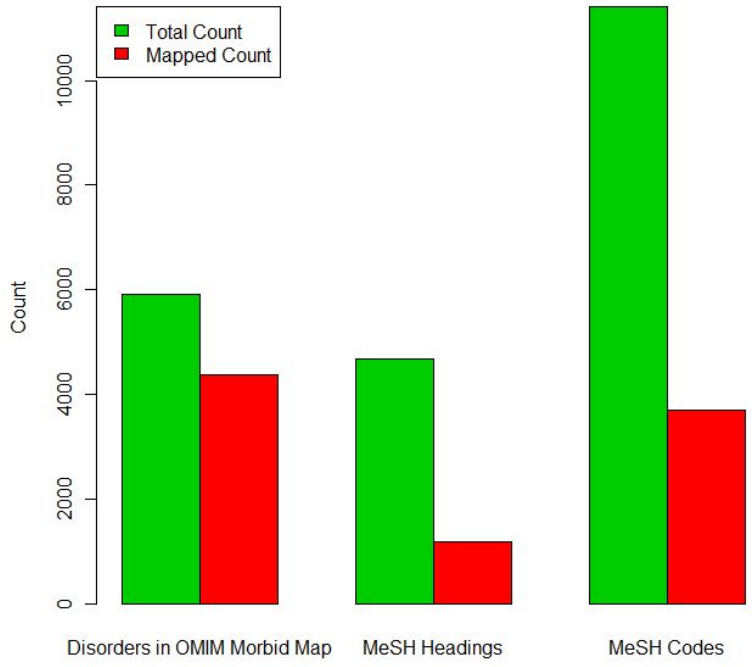

Figure 3.1. Barplot of OMIM-MeSH association entry

In contrast to other tumor suppressor genes that are mainly altered by truncating mutations, the majority of TP53 mutations are missense substitutions (75%). Other alterations include frame shift insertions and deletions (9%), nonsense mutations (7%), silent mutations (5%) and other infrequent alterations [95]. Since initial studies by Soussi and co-workers [96], it is recognized that different forms of mutant p53 proteins may have different functional and biological effects. About 30% of TP53 missense mutations found in cancer correspond to nucleotide substitutions at highly mutable CpG di-nucleotides, at codons encoding residues that play essential structural and chemical roles in the contact between the p53 protein and specific DNA sequences that constitute the p53 response elements (p53-RE) [94]. These mutations result in a very significant loss of DNA binding activity and transactivation capacity [96].

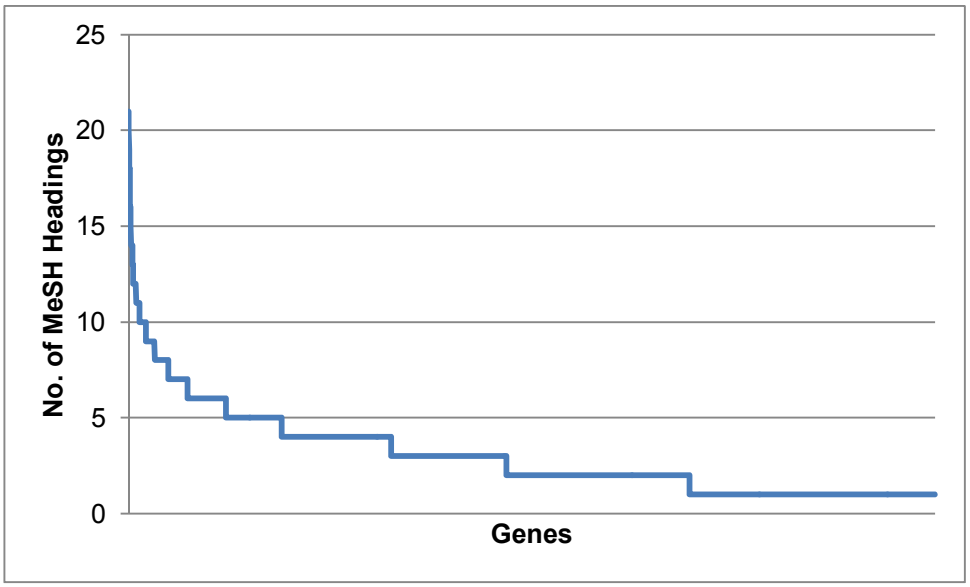

Figure 3.2. Distribution of genes which have MeSH Headings

Table 3.2. Top 9 genes which have multiple MeSH Headings (No. of MeSH Headings >= 15)

Gene	No. of MeSH Headings
TP53	21
BRCA2	20
PSEN1	19
ATM	18
APC	18
PIK3CA	16
SDHD	16
MAPT	16
CHEK2	15

The 1,192 MeSH Headings which have OMIM associations are mapped into OMIM disease genes (Figure 3.1). “Syndrome” MeSH Heading belongs to 605 genes (Table 3.3). MeSH Heading “Syndrome” [C23.550.288.500] belongs to MeSH Category “Pathological Conditions, Signs and Symptoms”.

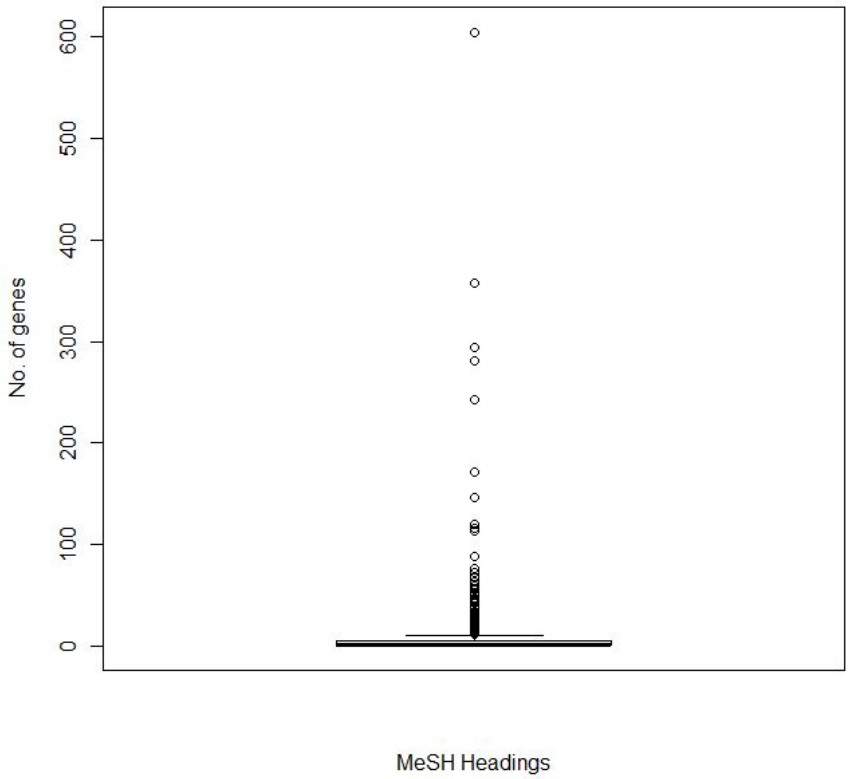

Figure 3.3. Boxplot of MeSH Headings which have genes

The average of OMIM disease genes which have MeSH Headings is 7.414 and the median is 2 (Figure 3.3). The main heading-topical subheading combination is a pre-coordination of terms, reducing the problem of term permutation, which looms large in most manual retrieval systems.

Table 3.3. Top 11 MeSH Headings which have multiple genes (No. of OMIM disease genes > 100)

MeSH Headings	No. of genes
Syndrome	605
Tics	358
Multiple Endocrine Neoplasia Type 1	294
Disease	281
Multiple Endocrine Neoplasia Type 2b	243
Multiple Endocrine Neoplasia Type 2a	171
AIDS-Related Complex	147
Neoplasms	147
Deafness	120
Intellectual Disability	116
Noma	114

Next, I considered distribution of SIFT scores of variants according to the disease groups. SIFT is a sequence homology-based tool that sorts intolerant from tolerant amino acid substitutions and predicts whether an amino acid substitution in a protein will have a phenotypic effect. SIFT is based on the premise that protein evolution is correlated with protein function. Positions important for function

should be conserved in an alignment of the protein family, whereas unimportant positions should appear diverse in an alignment. Positions with normalized probabilities less than 0.05 are predicted to be deleterious, those greater than or equal to 0.05 are predicted to be tolerated. The variants in the GBM patients are more deleterious (Figure 3.4). On the other hand, the variants in the Breast cancer patients are less deleterious (Figure 3.4).

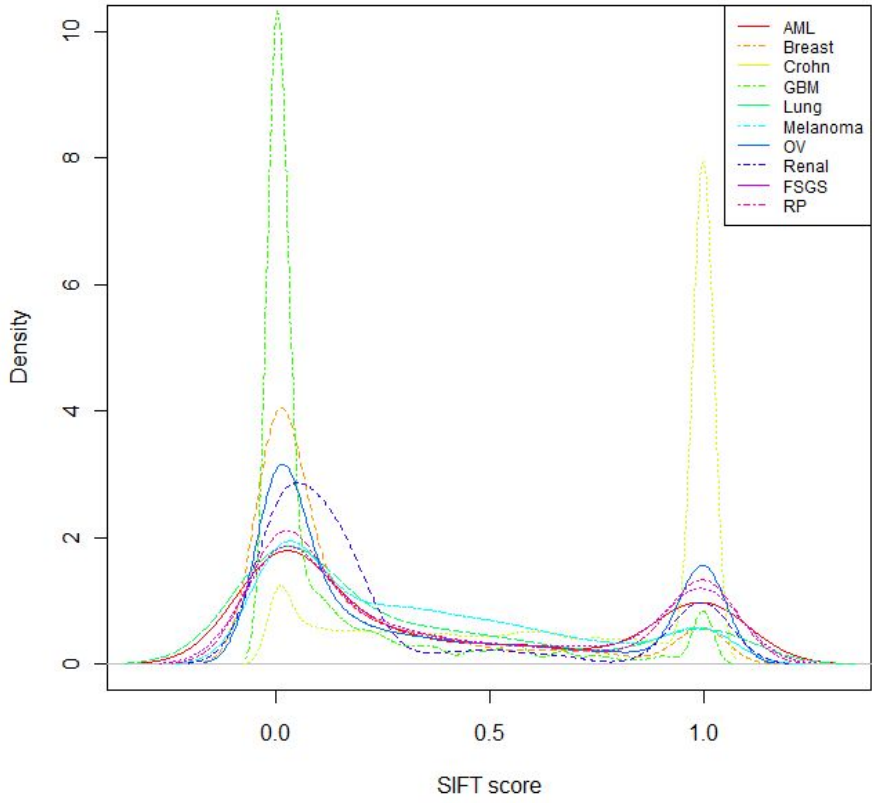

Figure 3.4. Distributions of SIFT scores of disease variants

As can be seen from Figure 3.4 and Table 3.4, SIFT scores of most GBM variants lie toward 0, whereas SIFT scores of most Crohn's disease variants lie toward 1. On the whole, disease variants in the cancer disease group are more damaging than those in the non-cancer disease group (Table 3.4). Specifically, the number of variants in the ovarian cancer patients is greater than that in other cancer patients (Table 3.4). However, average of SIFT scores in the number of variants in the ovarian cancer patients are less damaging than that in other cancer patients. On the other hand, average of SIFT scores in the number of variants in the GBM patients are more damaging than that in other cancer patients. The number of variants in this result depends on the type of sequencing technology platform, the methods, parameters and thresholds to detect variants. Although this result may provide the clue to relationship between the number of variants and degree of damaging effect on gene function according to the type of cancer, it is cautious about interpreting the result because of the type of sequencing technology platform, the methods, parameters and thresholds to detect variants.

Table 3.4. Distribution of SIFT scores between cancer and non-cancer

Category	Cancer diseases							Non-cancer diseases		
Diseases	AML	Breast	GBM	Lung	Melanoma	OV	Renal	Crohn	FSGS	RP
No. of variants	494	436	954	174	847	11567	52	449967	1212	2109
Median of SIFT scores	0.17	0.04	0.01	0.09	0.25	0.13	0.13	0.94	0.23	0.19
Average of SIFT scores	0.38	0.20	0.14	0.29	0.34	0.38	0.27	0.69	0.41	0.39

Table 3.5. Genes which belong to multiple MeSH categories

No. of Code	No. of genes
1~5	2457
6~10	453
< 10	11
Total	2924

Among 2924 OMIM genes, 2457 genes (about 84 percentages) belong to at most five MeSH codes (Table 3.5). Eleven genes belong to over 10 MeSH codes. The reason which genes belong to multiple categories is that each MeSH code can have multiple parents in a structure of directed acyclic graph (DAG).

Table 3.6. Top genes which belong to multiple MeSH categories (> 10)

Gene	No. of Codes
KRAS	14
CDH1	12
NRAS	12
BRAF	12
BRCA2	12
PSEN1	11
TNF	11
PIK3CA	11
HRAS	11
CHEK2	11
TP53	11
CASP8	11
CFH	11
MTHFR	11

Top genes which belong to multiple MeSH categories are almost cancer-related genes. KRAS, BRACA2 and TP53 and so on are well known for oncogenic pathway in various cancers. Considering the fact that genes mapped in to MeSH category are related cancer, it is cautious to interpret results of post hoc analysis.

I compared the MeSH codes of top genes which belong to multiple MeSH categories. The detailed tables are described in Appendix (Table S.1-5). MeSH categories such as C04: Neoplasms, C06: Digestive System Diseases, C16: Congenital, Hereditary, and Neonatal Diseases and Abnormalities, C20: Immune System Diseases, and C23: Pathological Conditions, Signs and Symptoms are included into KRAS (No. of MeSH Codes=14), CDH1 (No. of MeSH Codes=12), NRAS (No. of MeSH Codes=12), BRAF (No. of MeSH Codes=12), and BRCA2 (No. of MeSH Codes=12). In KEGG pathway [97], KRAS are included in 43 pathways. Eleven pathways among them correspond to cancer-related pathways: Acute myeloid leukemia, Cronic myeloid leukemia, Bladder cancer, Colorectal cancer, Endometrial cancer, Glioma, Hepatocellular carcinoma, Non-small cell lung cancer, Pancreatic cancer, Prostate cancer, Renal carcinoma, Thyroid cancer. In addition, 12 pathways among them correspond to signaling-related pathways: MAPK signaling pathway, ErbB signaling pathway, Chemokine pathway, PI3K-Akt signaling pathway, VEGF signaling pathway, T cell receptor signaling pathway, B

cell signaling pathway, Fcepsilon RI singnaling pathway, Neurotrophin signaling pathway, Insulin signaling pathway, GnRH signaling pathway. It is natural that KRAS belongs to multiple MeSH categories because KRAS is associated with several cancer-related pathways and various signaling pathways. The case of CDH1, CDH1 is related to just three pathways: Cell cycle, Cell adhesion molecules (CAMs), Ubiquitin mediated proteolysis. CDH1 may well also belong to multiple MeSH categories because CDH1 is the core of cell cycle progression pathway.

I found diseases which are mapped into KRAS gene. KRAS gene belongs to Nervous system diseases, Immune system diseases as well as Neoplasm. KRAS acts as a molecular on/off switch. Once it is turned on it recruits and activates proteins necessary for the propagation of growth factor and other receptors' signal, such as c-Raf and PI 3-kinase. KRAS binds to GTP in the active state and possesses an intrinsic enzymatic activity which cleaves the terminal phosphate of the nucleotide converting it to GDP. Upon conversion of GTP to GDP, KRAS is turned off. The rate of conversion is usually slow but can be sped up dramatically by an accessory protein of the GTPase-activating protein (GAP) class, for example RasGAP. In turn KRAS can bind to proteins of the Guanine Nucleotide Exchange Factor (GEF) class, for example SOS1, which forces the release of bound nucleotide. Subsequently, KRAS binds GTP present in the cytosol and the GEF is released from ras-GTP. This proto-oncogene is a Kirsten ras oncogene homolog from the mammalian ras gene family. A single amino acid substitution, and in particular a single nucleotide

substitution, is responsible for an activating mutation. The transforming protein that results is implicated in various malignancies, including lung adenocarcinoma, mucinous adenoma, ductal carcinoma of the pancreas and colorectal carcinoma. Several germline KRAS mutations have been found to be associated with Noonan syndrome [98] and cardio-facio-cutaneous syndrome [99]. Somatic KRAS mutations are found at high rates in leukemias, colon cancer [100], pancreatic cancer [101] and lung cancer [102].

Cadherin-1 also known as CAM 120/80 or epithelial cadherin (E-cadherin) or uvomorulin is a protein that in humans is encoded by the CDH1 gene. CDH1 has also been designated as CD324 (cluster of differentiation 324). It is a tumor suppressor gene [103, 104]. Loss of E-cadherin function or expression has been implicated in cancer progression and metastasis. E-cadherin down regulation decreases the strength of cellular adhesion within a tissue, resulting in an increase in cellular motility. This in turn may allow cancer cells to cross the basement membrane and invade surrounding tissues. E-cadherin is also used by pathologists to diagnose different kinds of breast cancer.

NRAS is an N-ras oncogene encoding a membrane protein that shuttles between the Golgi apparatus and the plasma membrane. This shuttling is regulated through palmitoylation and depalmitoylation by the ZDHHC9-GOLGA7 complex. The encoded protein, which has intrinsic GTPase activity, is activated by a guanine

nucleotide-exchange factor and inactivated by a GTPase activating protein. Mutations in this gene have been associated with somatic rectal cancer, follicular thyroid cancer, autoimmune lymphoproliferative syndrome, Noonan syndrome, and juvenile myelomonocytic leukemia.

BRAF is a human gene that makes a protein called B-Raf. The gene is also referred to as proto-oncogene B-Raf and v-Raf murine sarcoma viral oncogene homolog B1, while the protein is more formally known as serine/threonine-protein kinase B-Raf. Mutations in the BRAF gene can cause disease in two ways. First, mutations can be inherited and cause birth defects. Second, mutations can appear later in life and cause cancer, as an oncogene. Inherited mutations in this gene cause cardiofaciocutaneous syndrome, a disease characterized by heart defects, mental retardation and a distinctive facial appearance [105]. Acquired mutations in this gene have been found in cancers, including non-Hodgkin lymphoma, colorectal cancer, malignant melanoma, papillary thyroid carcinoma, non-small-cell lung carcinoma, and adenocarcinoma of the lung. The V600E mutation of the BRAF gene has been associated with hairy cell leukemia in numerous studies and has been suggested for use in screening for Lynch syndrome to reduce the number of patients undergoing unnecessary MLH1 sequencing [106].

BRCA2 (breast cancer type 2 susceptibility protein) is a protein found inside cells. In humans, the instructions to make this protein are carried by a gene, also

called BRCA2 [107]. BRCA2 belongs to the tumor suppressor gene family,[108, 109] and orthologs have been identified in most mammals for which complete genome data are available. The protein encoded by this gene is involved in the repair of chromosomal damage with an important role in the error-free repair of DNA double strand breaks.

3.2. Disease predisposition patterns of healthy humans in the 1000 Genomes Project

The 1000 Genomes Project is an international collaboration to produce an extensive public catalog of human genetic variation, including SNPs and structural variants, and their haplotype contexts. This resource will support genome-wide association studies and other medical research studies. The genomes of about 2500 unidentified people from about 25 populations around the world will be sequenced using next-generation sequencing technologies. The results of the study will be freely and publicly accessible to researchers worldwide.

The project goals are explicit: they want to produce a catalog of human variation down to variants that occur at 1% frequency or less over the genome, and 0.5–0.1% in genes. The intention of the project is to provide a resource that will greatly increase the ability of scientists to do genetic studies on common human disease. The initial goal was to discover most of the genetic variation that occurs at a population frequency greater than 1% by deep sequencing at least 1,000 individuals from different worldwide populations using next-generation platforms and technologies. The genomes of approximately 2,000 individuals, from at least 20 different populations representing Africa, Europe, East Asia, and the American, are being collected and sequenced. The populations included will each have approximately 60 to 100 samples sequenced. For some populations, trios (both

biological parents and an adult child) have been collected. Many of the samples, including some from the children, are going to be densely genotyped using genome-wide arrays. The goal of this study design is to reconstruct the parental chromosomal phase using the information provided by the child.

One will sequence the entire genome of six individuals: two adults and both sets of their parents. DNA in these six genomes will be analyzed repeatedly up to 20 times to ensure almost complete coverage. A second project will sequence 180 individual genomes at light (2×) coverage, leaving gaps. The third project will be to fully sequence (20× coverage) the protein-coding regions of 1000 genes (5% of the total) in about 1000 genomes. The samples, all anonymous and with no clinical information, will mainly be drawn from those collected for the HapMap [5], which includes people of European, Asian, and African descent.

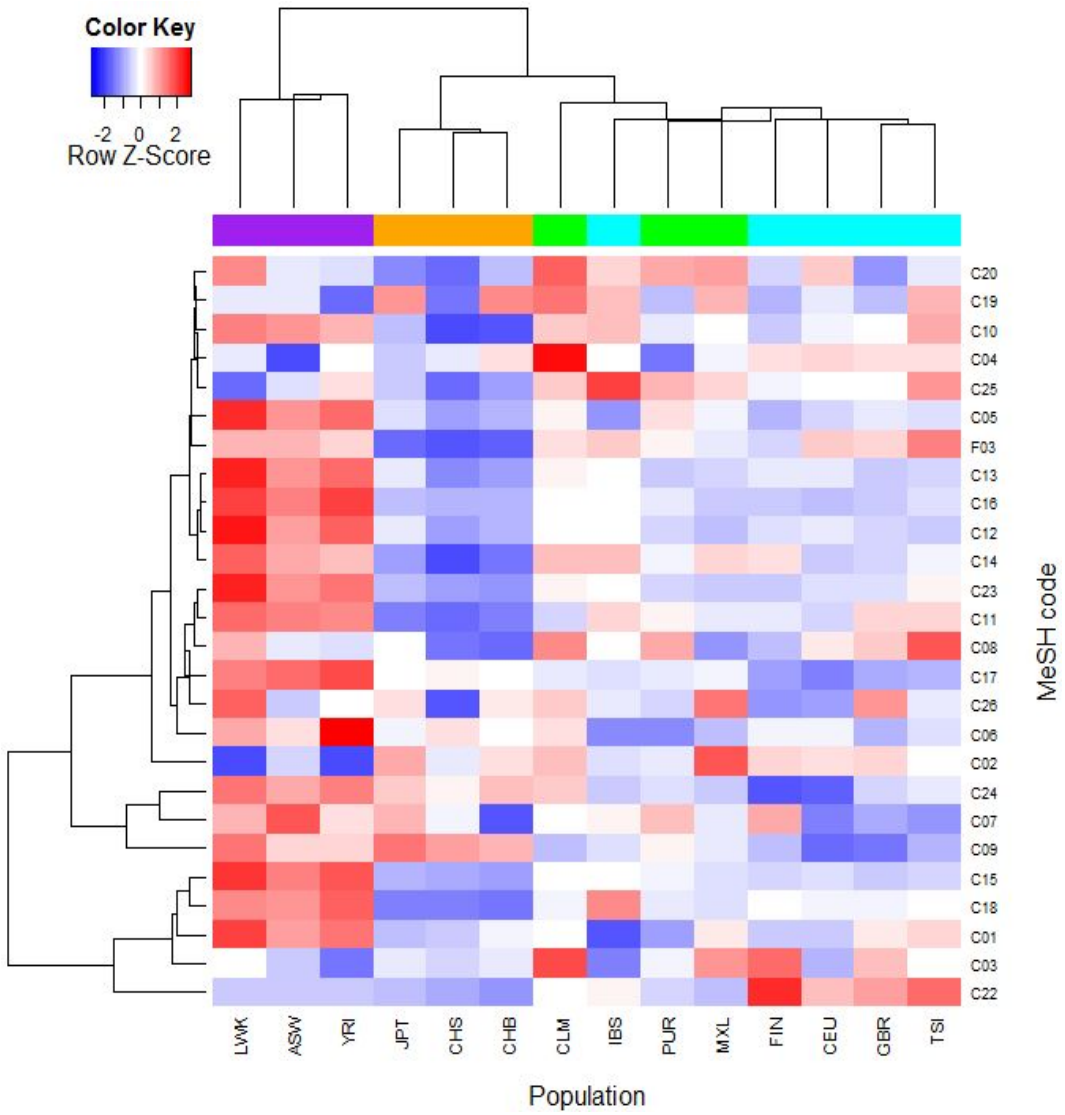

Figure 3.5. Heatmap of mutual information of diseases in the MeSH tree according to the population in the 1000 Genomes project. Distance measure: Manhattan, Linkage method: Median

The 1000 Genomes Project Consortium reported that populations with African ancestry contributed the largest number of variants and contained the highest fraction of novel variants, reflecting the greater diversity in African populations [23]. For example, 63% of novel SNPs in the low-coverage project and 44% in the exon project were discovered in the African populations, compared to 33% and 22% in the European ancestry populations. In the heatmap of mutual information of diseases in the MeSH tree according to the population, Africans have higher mutual information than other populations. It suggests that in order to interpret personal genome properly, I may consider population information together.

In order to find MeSH categories which distinguish populations in terms of mutual information, I applied ANOVA test for healthy people in the 1000 Genomes Project (Table 3.7). MeSH category C15: Hemic and Lymphatic Diseases is statically significant (P value = 2.58e-07). Hematologic diseases and diseases of the lymphatic system collectively. Hemic diseases include disorders involving the formed elements (e.g., ERYTHROCYTE AGGREGATION, INTRAVASCULAR) and chemical components (e.g., BLOOD PROTEIN DISORDERS); lymphatic diseases include disorders relating to lymph, lymph nodes, and lymphocytes.

Sickle-cell disease (SCD), or sickle-cell anaemia (or anemia, SCA) or drepanocytosis, is an autosomal recessive genetic blood disorder with overdominance, characterized by red blood cells that assume an abnormal, rigid,

sickle shape. Sickling decreases the cells' flexibility and results in a risk of various complications. The sickling occurs because of a mutation in the hemoglobin gene. Three quarters of sickle-cell cases occur in Africa. A recent WHO report estimated that around 2% of newborns in Nigeria were affected by sickle cell anaemia, giving a total of 150,000 affected children born every year in Nigeria alone. The carrier frequency ranges between 10% and 40% across equatorial Africa, decreasing to 1–2% on the North African coast and <1% in South Africa.

Table 3.7. Statistically significant difference in MeSH codes among populations

Code	Term	P value
C01	Bacterial Infections and Mycoses	0.01400
C02	Virus Diseases	0.01435
C03	Parasitic Diseases	0.22684
C04	Neoplasms	0.55150
C05	Musculoskeletal Diseases	0.00003
C06	Digestive System Diseases	0.02709
C07	Stomatognathic Diseases	0.20351
C08	Respiratory Tract Diseases	0.23714
C09	Otorhinolaryngologic Diseases	0.00055
C10	Nervous System Diseases	0.00139
C11	Eye Diseases	0.00002
C12	Male Urogenital Diseases	0.00013
C13	Female Urogenital Diseases and Pregnancy Complications	0.00013
C14	Cardiovascular Diseases	0.00058
C15	Hemic and Lymphatic Diseases	2.58E-07
C16	Congenital, Hereditary, and Neonatal Diseases and Abnormalities	1.19E-06
C17	Skin and Connective Tissue Diseases	1.19E-06
C18	Nutritional and Metabolic Diseases	0.00009
C19	Endocrine System Diseases	0.51232
C20	Immune System Diseases	0.00852
C22	Animal Diseases	0.00130
C23	Pathological Conditions, Signs and Symptoms	0.00005
C24	Occupational Diseases	0.00149
C25	Substance-Related Disorders	0.04237
C26	Wounds and Injuries	0.59611
F03	Mental Disorders	0.00019

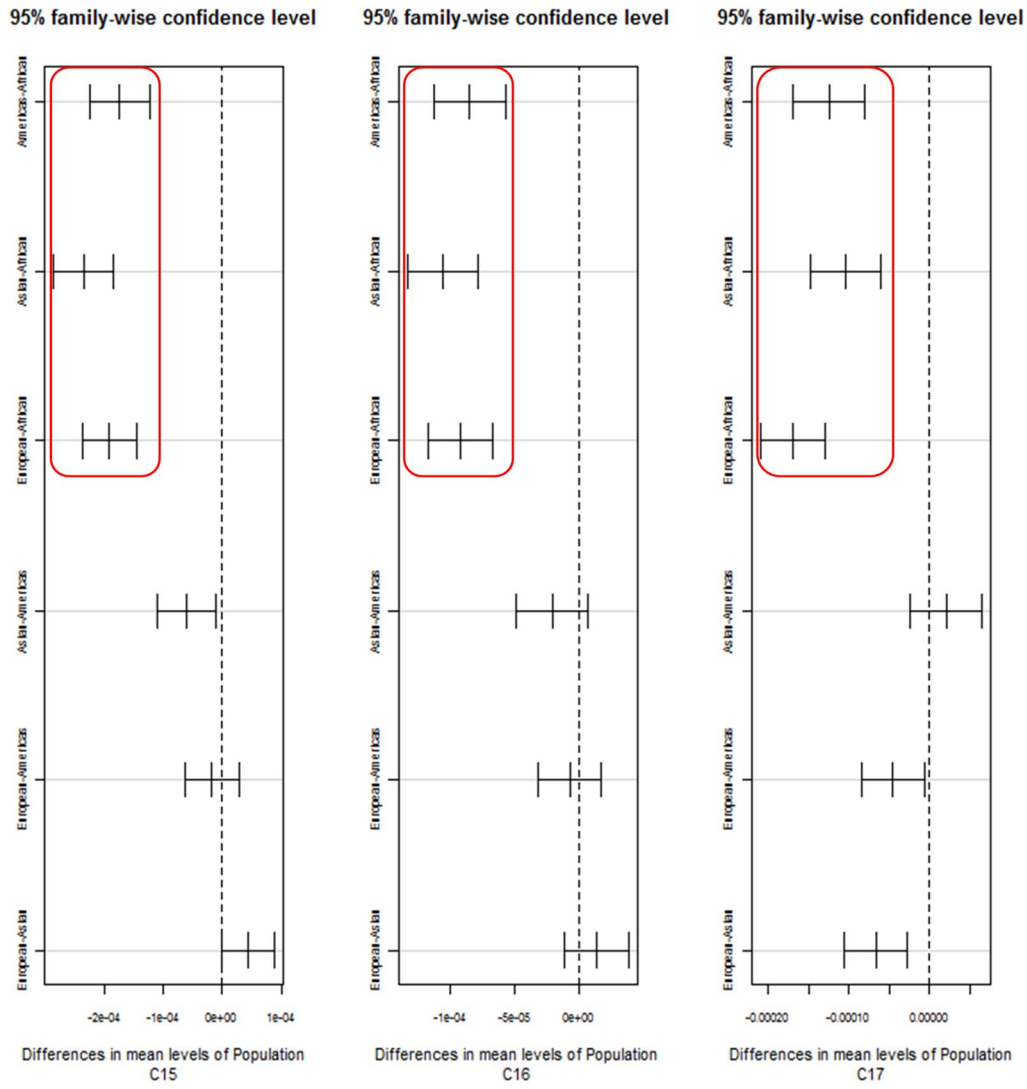

Figure 3.6. Post hoc test in ANOVA in order to identify statistical different diseases in the MeSH category in the sub-populations of the 1000 Genome project

MeSH category C15: Hemic and Lymphatic Diseases, MeSH category C16: Congenital, Hereditary, and Neonatal Diseases and Abnormalities and MeSH category C17: Skin and Connective Tissue Diseases showed statistical difference between Africans and other populations in the 1000 Genomes (Figure 3.6, Table 3.8).

Table 3.8. Statistically different diseases in the MeSH category in the sub-populations

Population	C15	C16	C17
American-African	4.8E-06	1.37E-05	2.96E-05
Asian-African	3E-07	1.8E-06	0.000146
European-African	7E-07	2.3E-06	6E-07
Asian-American	0.018112	0.158669	0.497189
European-American	0.657678	0.811637	0.024881
European-Asian	0.058047	0.37135	0.002112

Differences between African and American are relatively high (Table 3.8). Human evolution refers to the evolutionary process leading up to the appearance of modern humans. While it began with the last common ancestor of all life, the topic usually only covers the evolutionary history of primates, in particular the genus

Homo, and the emergence of Homo sapiens as a distinct species of hominids (or "great apes"). The study of human evolution involves many scientific disciplines, including embryology and genetics. Out of Africa has gained support from research using female mitochondrial DNA (mtDNA) and the male Y chromosome. After analyzing genealogy trees constructed using 133 types of mtDNA, researchers concluded that all were descended from a female African progenitor, dubbed Mitochondrial Eve. Out of Africa is also supported by the fact that mitochondrial genetic diversity is highest among African populations [110].

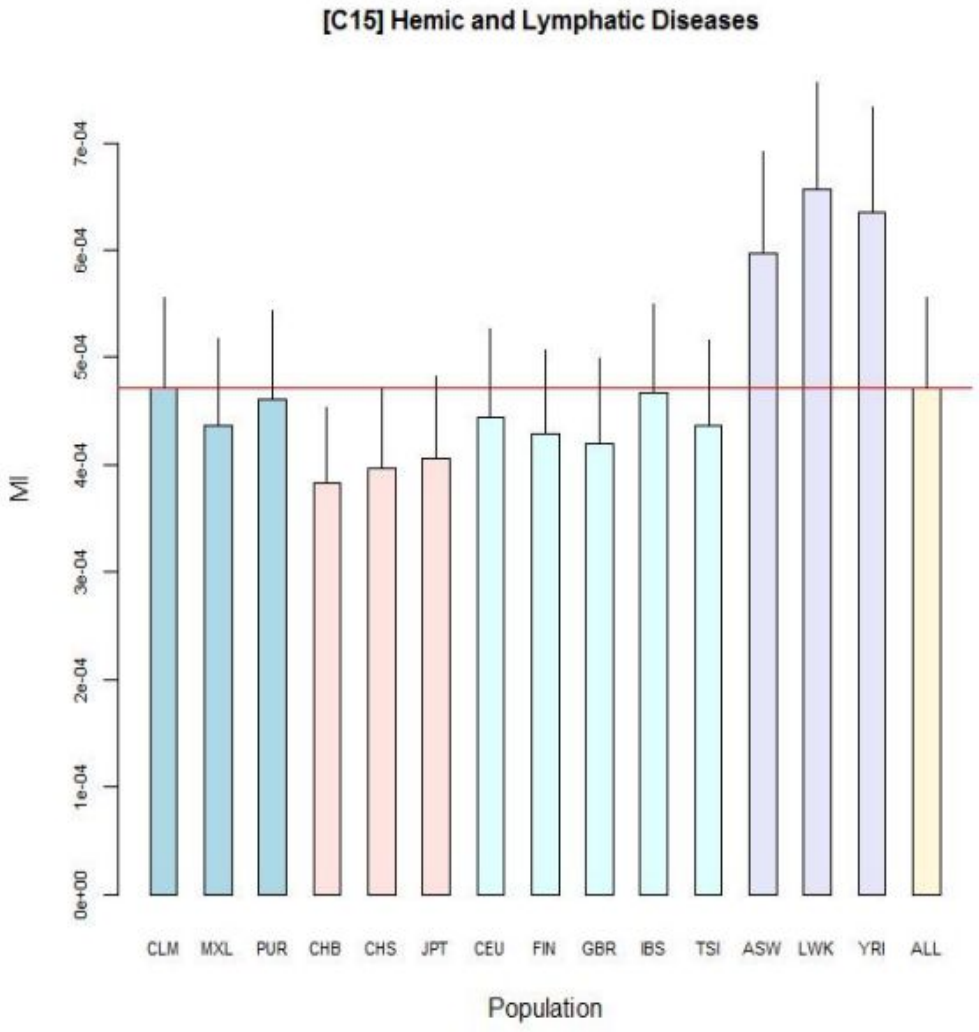

Figure 3.7. Barplot of mutual information of MeSH category 15: Hemic and Lymphatic Diseases according to population in the 1000 Genomes Project

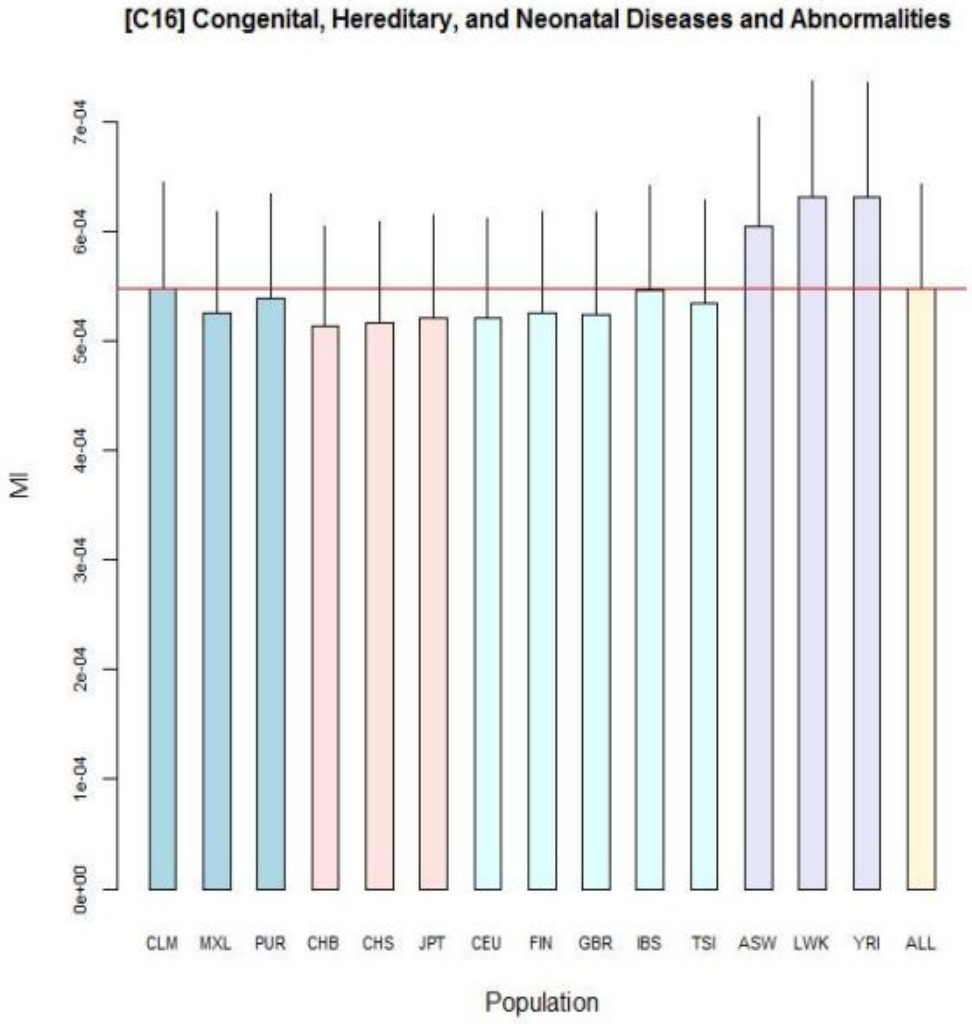

Figure 3.8. Barplot of mutual information of MeSH category C16: Congenital, Hereditary, and Neonatal Diseases and Abnormalities according to population in the 1000 Genomes Project

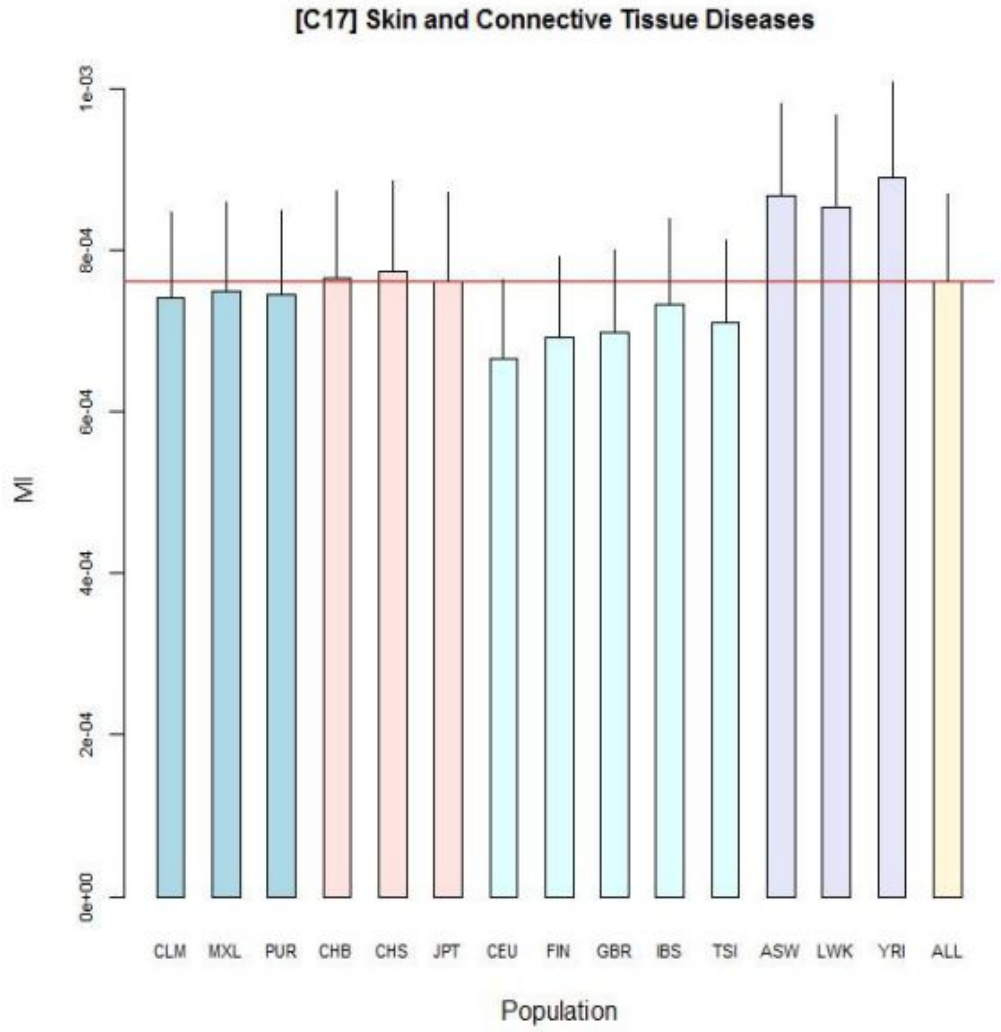

Figure 3.9. Barplot of mutual information of MeSH category C17: Skin and Connective Tissue Diseases according to population in the 1000 Genomes Project

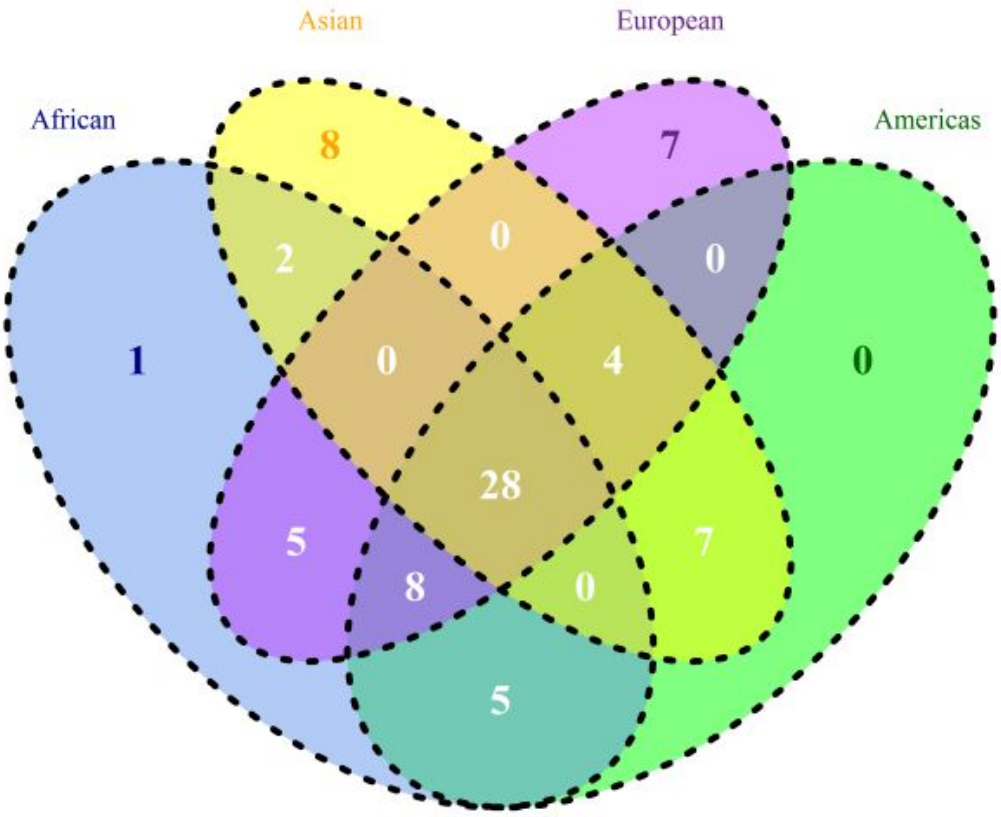

Figure 3.10. Venn diagram of top 50 diseases in the 1000 Genome project

In Venn diagram of top 50 diseases in the 1000 Genome project (Figure 3.10), Macular Degeneration (C11.768.585.439) belongs to only African. In addition, MeSH term Arthritis (C05.550.114), MeSH term Ichthyosis Vulgaris (C16.131.831.512.410, C16.320.850.405, C17.800.428.333.410, C17.800.804.512.410, and C17.800.827.405) and Viremia (C02.937, C23.550.470.790.500.900) belongs to only Asian. AIDS-Related Complex (C02.782.815.616.400.080, C02.800.801.400.080, C02.839.080, and C20.673.480.080) and Hematuria (C12.777.934.442, C13.351.968.934.442, and C23.550.414.849) belong to only European.

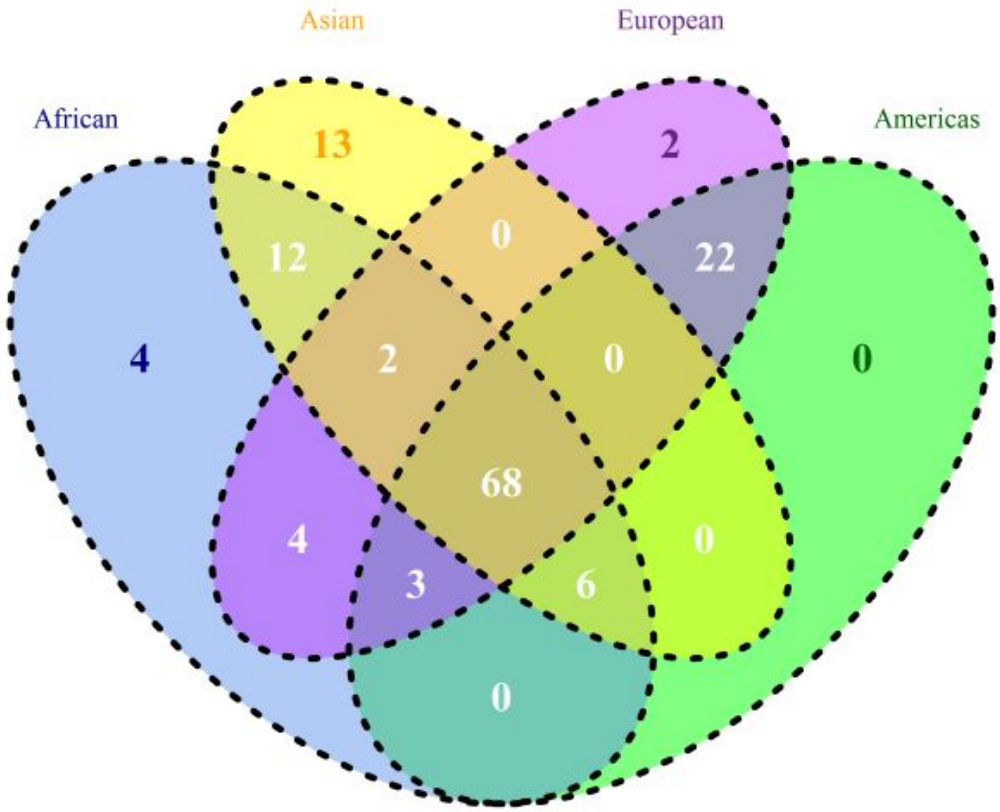

Figure 3.11 Venn diagram of top 100 diseases in the 1000 Genome project

Before comparing ranks of disease terms in the patient sequencing data, I ranked MeSH disease terms in healthy people in the 1000 Genome Projects (Figure 3.12). In healthy people, distribution of disease terms ranked randomly. In other words, if distributions of disease terms are ranked high in the patient groups, this method can show to find disease predisposition patterns.

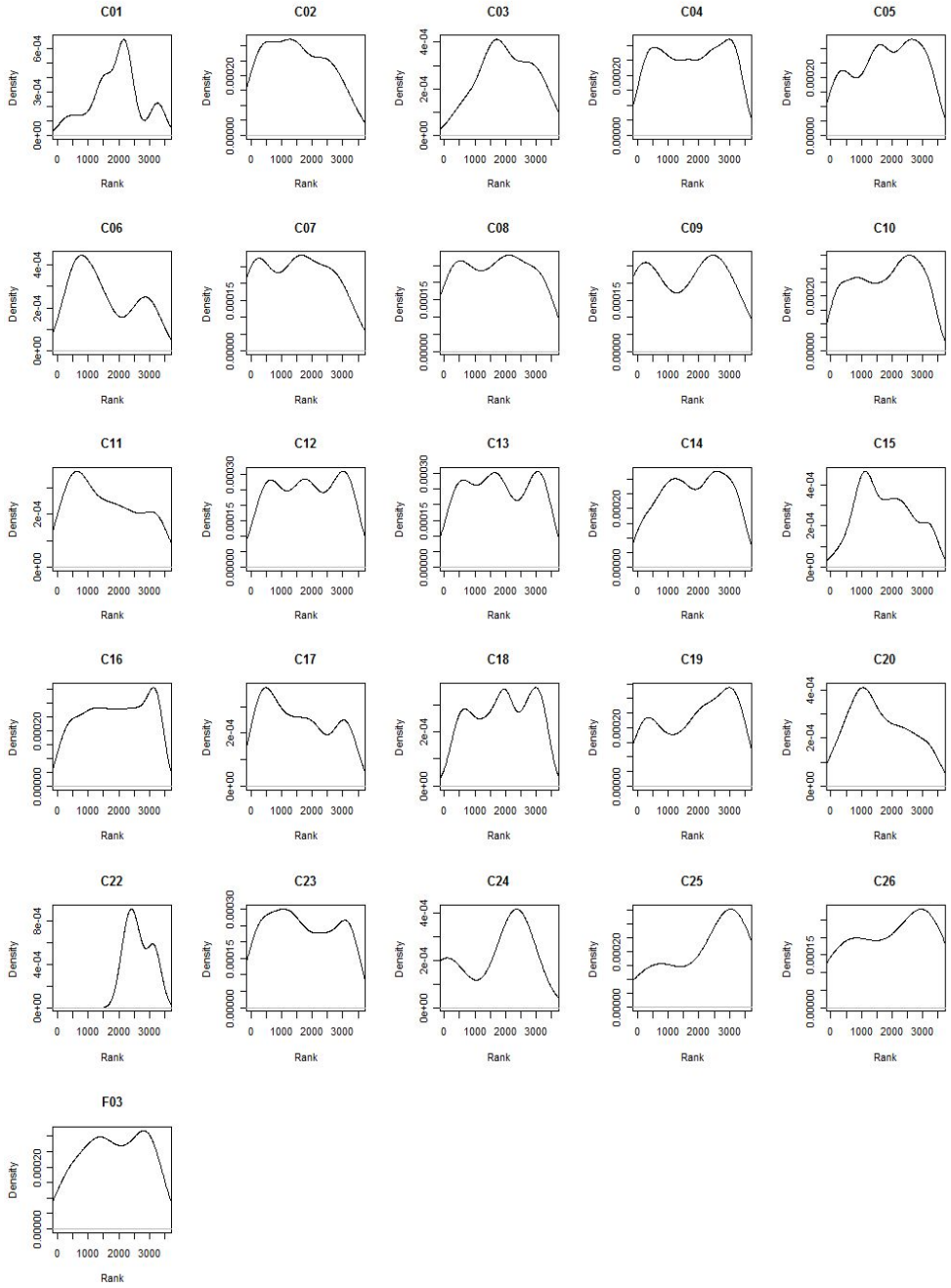

Figure 3.12. Distribution of rank in the 1000 Genomes according to MeSH category

3.3. Disease predisposition patterns in the disease group

Glioblastoma multiforme (GBM) is the most common and aggressive type of brain tumor in humans and the first cancer with comprehensive genomic profiles mapped by TCGA project. A central challenge in large-scale genome projects, such as the TCGA GBM project, is the ability to distinguish cancer-causing ‘‘driver’’ mutations from passively selected ‘‘passenger’’ mutations.

Nearly all GBM tumors contain alterations in the p53 tumor suppressor pathway, but individual tumors exhibit diverse mechanisms for pathway alteration – mutation or homozygous deletion of TP53, mutation or homozygous deletion of CDKN2A/ARF, or amplification of MDM2/MDM4. If tumors frequently target biological modules that execute key biological processes, and network knowledge about such modules is available, I hypothesized that it would be possible to algorithmically identify frequently perturbed modules, and from these modules identify candidate driver mutations.

Acute myeloid leukemia (AML), also known as acute myelogenous leukemia, is a cancer of the myeloid line of blood cells, characterized by the rapid growth of abnormal white blood cells that accumulate in the bone marrow and interfere with the production of normal blood cells. The symptoms of AML are caused by replacement of normal bone marrow with leukemic cells, which causes a drop in red blood cells, platelets, and normal white blood cells. These symptoms

include fatigue, shortness of breath, easy bruising and bleeding, and increased risk of infection. Several risk factors and chromosomal abnormalities have been identified, but the specific cause is not clear.

I provide the visualization using the 336 MeSH categories which are associated variants detected in the AML patient group. The more important an item is, the larger its label and its circle are. “Neuroectodermal Tumors, Primitive” is notable term with respect to the size (i.e., statistical significance). BAALC, the human member of a novel mammalian neuroectoderm gene lineage, is implicated in hematopoiesis and acute leukemia [111]. Next, “Mucinoses” is reported as a presenting sign of acute myeloblastic leukemia [112].

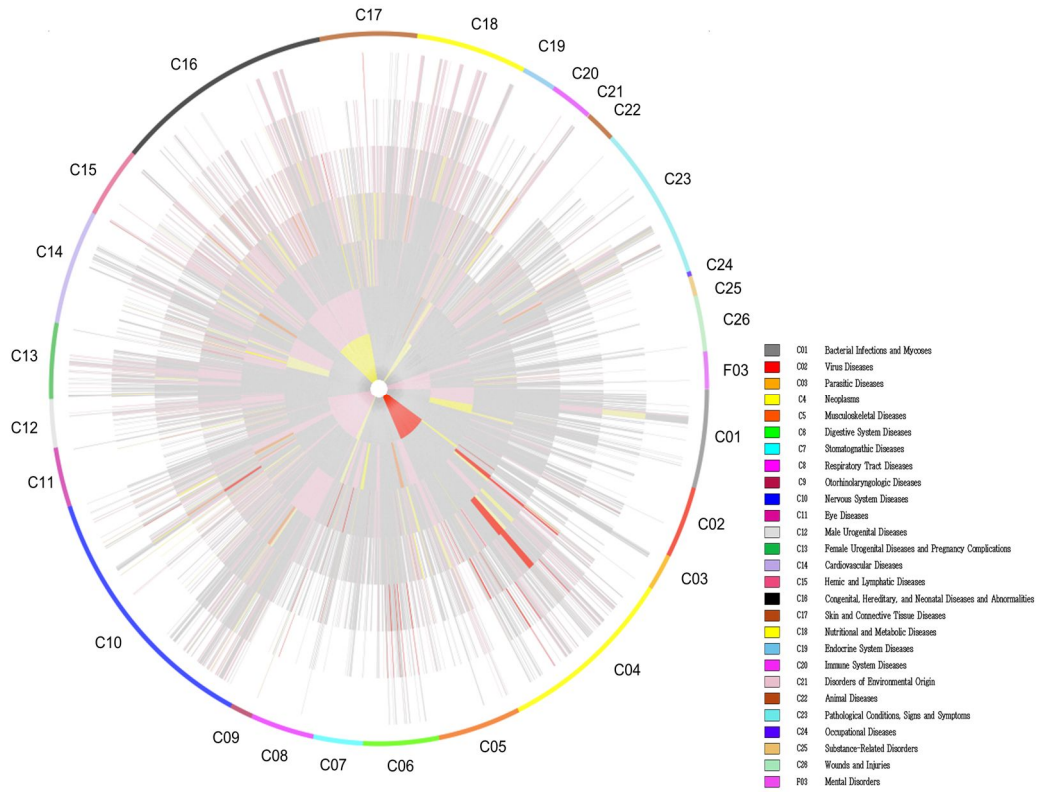

Figure 3.13 Circle plot of mutual information in the Melanoma patient

In order to find disease predisposition patterns from disease sequencing data, I visualized mutual information according to MeSH categories. In a melanoma patient, circle plot showed high mutual information values concentrate on relatively MeSH category C04: Neoplasm compared with other MeSH categories (Figure 3.13). The tree number of Melanoma is C04.557.465.625.650.510.515, C04.557.580.625.650.510.515 and C04.557.665.510.515. The children nodes of Melanoma are Hutchinson's Melanotic Freckle [C04.557.465.625.650.510.385], Melanoma, Amelanotic [C04.557.465.625.650.510.515] and Melanoma, Experimental [C04.557.465.625.650.510.525]

Table 3.9. MeSH codes of used sequencing data

Disease	MeSH Disease terms	No. of MeSH codes	MeSH code
Acute Myeloid Leukemia	-Leukemia, Myeloid, Acute	1	C04.557.337.539.550
Bladder Urothelial Carcinoma	-Urinary Bladder Neoplasms	5	C04.588.945.947.960, C12.758.820.968, C12.777.829.813, C13.351.937.820.945, C13.351.968.829.707
Breast Carcinoma	-Breast Neoplasms	2	C04.588.180, C17.800.090.500
Colon adenocarcinoma	-Colonic Neoplasms	5	C04.588.274.476.411.307.180 , C06.301.371.411.307.180, C06.405.249.411.307.180, C06.405.469.158.356.180,

			C06.405.469.491.307.180
Glioblastoma	-Glioblastoma	3	C04.557.465.625.600.380.080.335, C04.557.470.670.380.080.335, C04.557.580.625.600.380.080.335
Kidney renal clear cell carcinoma	-Carcinoma, Renal Cell	6	C04.557.470.200.025.390, C04.588.945.947.535.160, C12.758.820.750.160, C12.777.419.473.160, C13.351.937.820.535.160, C13.351.968.419.473.160
Small Cell Lung Carcinoma	-Lung Neoplasms -Small Cell Lung Carcinoma	6	C04.588.894.797.520, C08.381.540, C08.785.520, C04.588.894.797.520.109.220.624, C08.381.540.140.750, C08.785.520.100.220.750
Lung adenocarcinoma	-Lung Neoplasms	3	C04.588.894.797.520, C08.381.540, C08.785.520
Lung squamous cell carcinoma	-Lung Neoplasms -Carcinoma, Squamous Cell	5	C04.588.894.797.520, C08.381.540, C08.785.520, C04.557.470.200.400, C04.557.470.700.400
Melanoma	-Melanoma	3	C04.557.465.625.650.510, C04.557.580.625.650.510, C04.557.665.510
Ovarian Neoplasms	-Ovarian Neoplasms	5	C04.588.322.455, C13.351.500.056.630.705, C13.351.937.418.685, C19.344.410, C19.391.630.705
Prostate adenocarcinoma	-Prostatic Neoplasms	4	C04.588.945.440.770, C12.294.260.750, C12.294.565.625, C12.758.409.750
Rectum adenocarcinoma	-Rectal Neoplasms	5	C04.588.274.476.411.307.790, C06.301.371.411.307.790, C06.405.249.411.307.790, C06.405.469.491.307.790,

			C06.405.469.860.180.500
Crohn Disease	-Crohn Disease	2	C06.405.205.731.500, C06.405.469.432.500
Glomerulosclerosis Focal Segmental	-Glomerulosclerosis, Focal Segmental	2	C12.777.419.570.363.660, C13.351.968.419.570.363.640
Retinitis Pigmentosa	-Retinitis Pigmentosa	3	C11.270.684, C11.768.585.658.500, C16.320.290.684

In order to compare rank of diseases, I displayed rank density for each disease groups of patients (Figure 3.14). Most of MeSH disease terms are highly ranked in the corresponding patients sequencing data. Specifically, cancer data are highly ranked compared with non-cancer data. The reason that non-cancer MeSH disease terms are low ranked is the lack of knowledge of association for non-cancer disease. On the other hand, cancer diseases have been studied, so cancer diseases have relatively a lot of knowledge of disease-gene association.

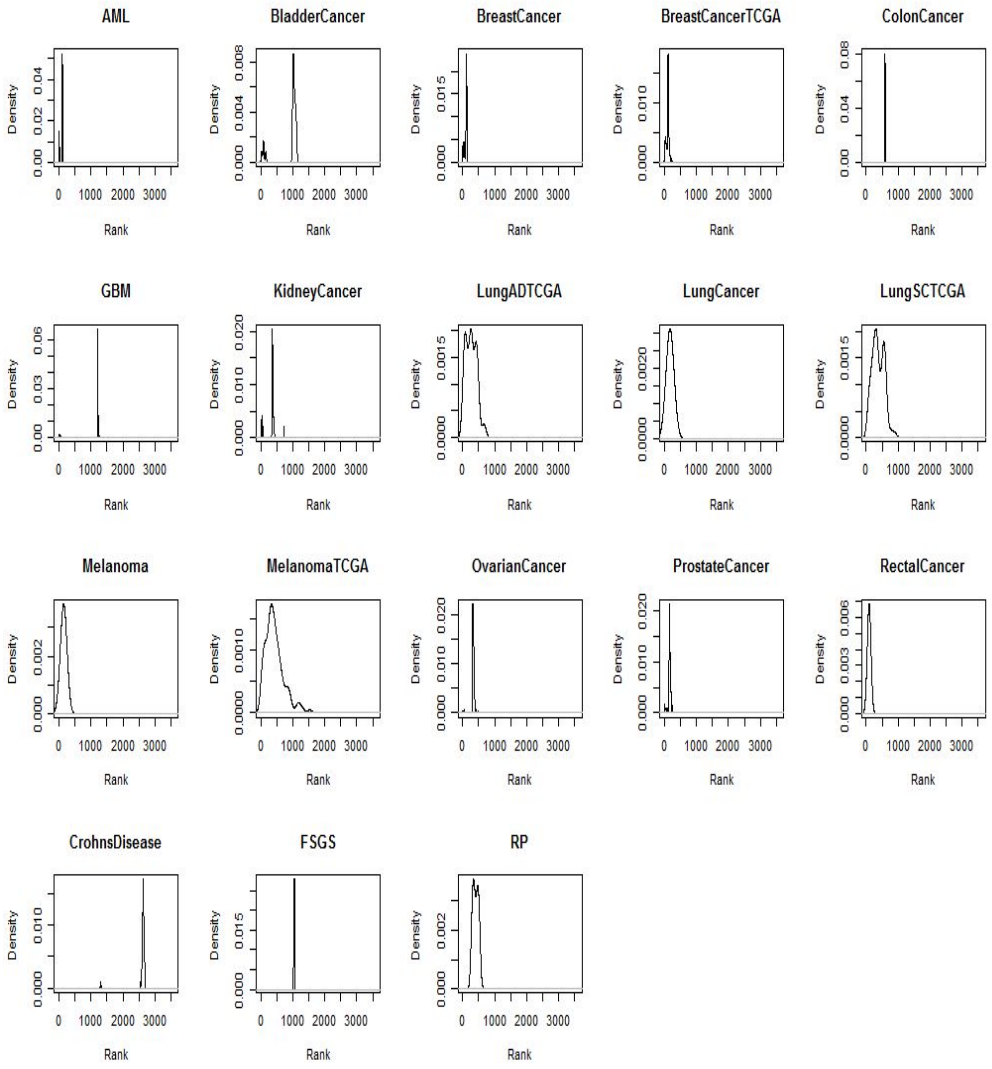

Figure 3.14 Distribution of rank of the corresponding diseases

In order to identify the rank patterns of diseases, I visualized rank plots according to disease sequencing data. MeSH codes of used sequencing data are summarized in Table 3.14. Figure 3.14 shows the rank pattern of the corresponding MeSH disease terms in the disease sequencing data. For example, the plot displays rank density of Leukemia, Myeloid, Acute [C04.557.337.539.550] as the MeSH Disease term in sequencing data of 50 AML patients. In other words, the corresponding MeSH disease terms in the disease sequencing data are ranked high.

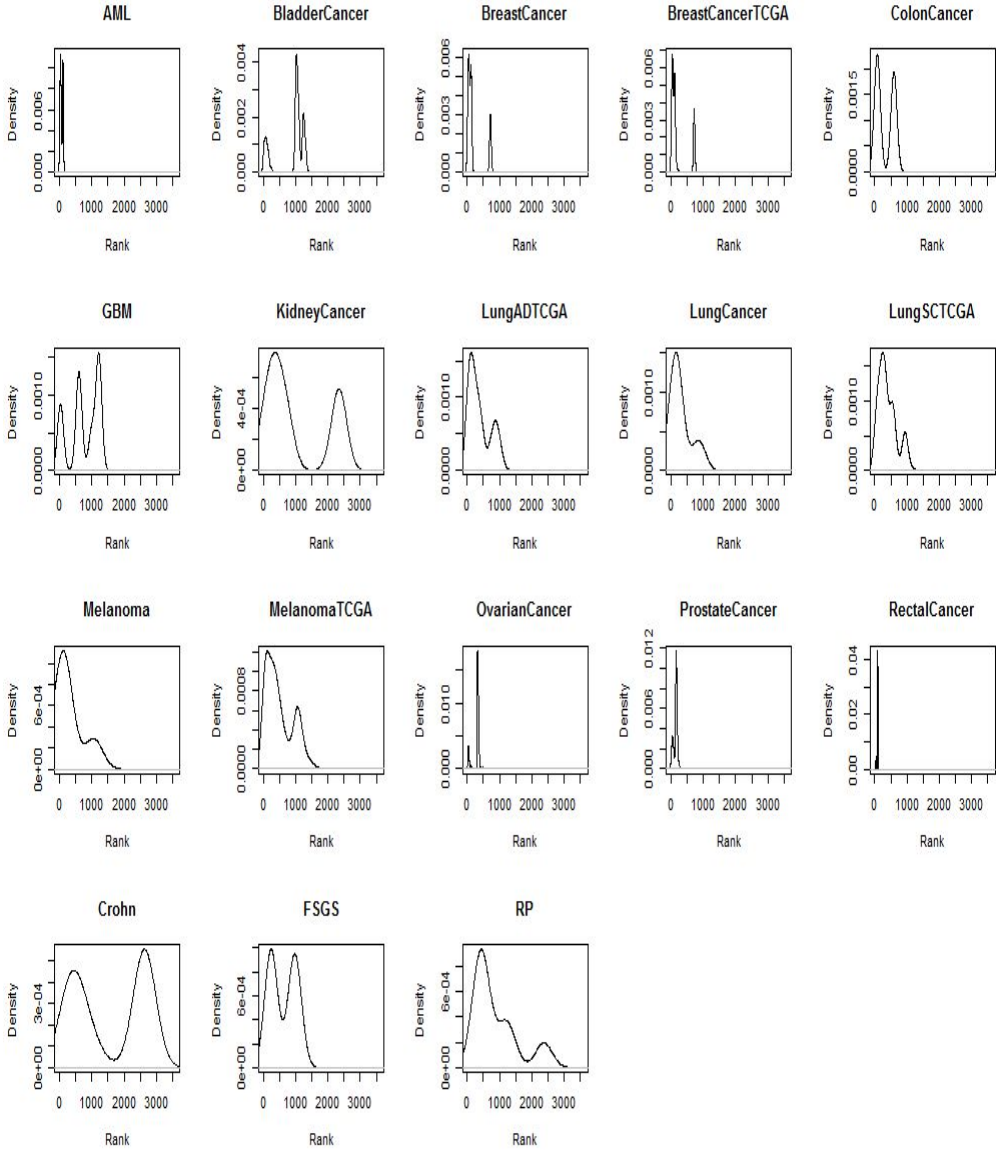

Figure 3.15 Distribution of rank of the corresponding diseases and their parent diseases in MeSH hierarchy

Because MeSH tree is hierarchically organized, I considered the parents MeSH disease terms of patient sequencing data (Figure 3.15). Distribution of rank of the corresponding diseases and their parent diseases showed somewhat blunt patterns compared with direct MeSH disease terms.

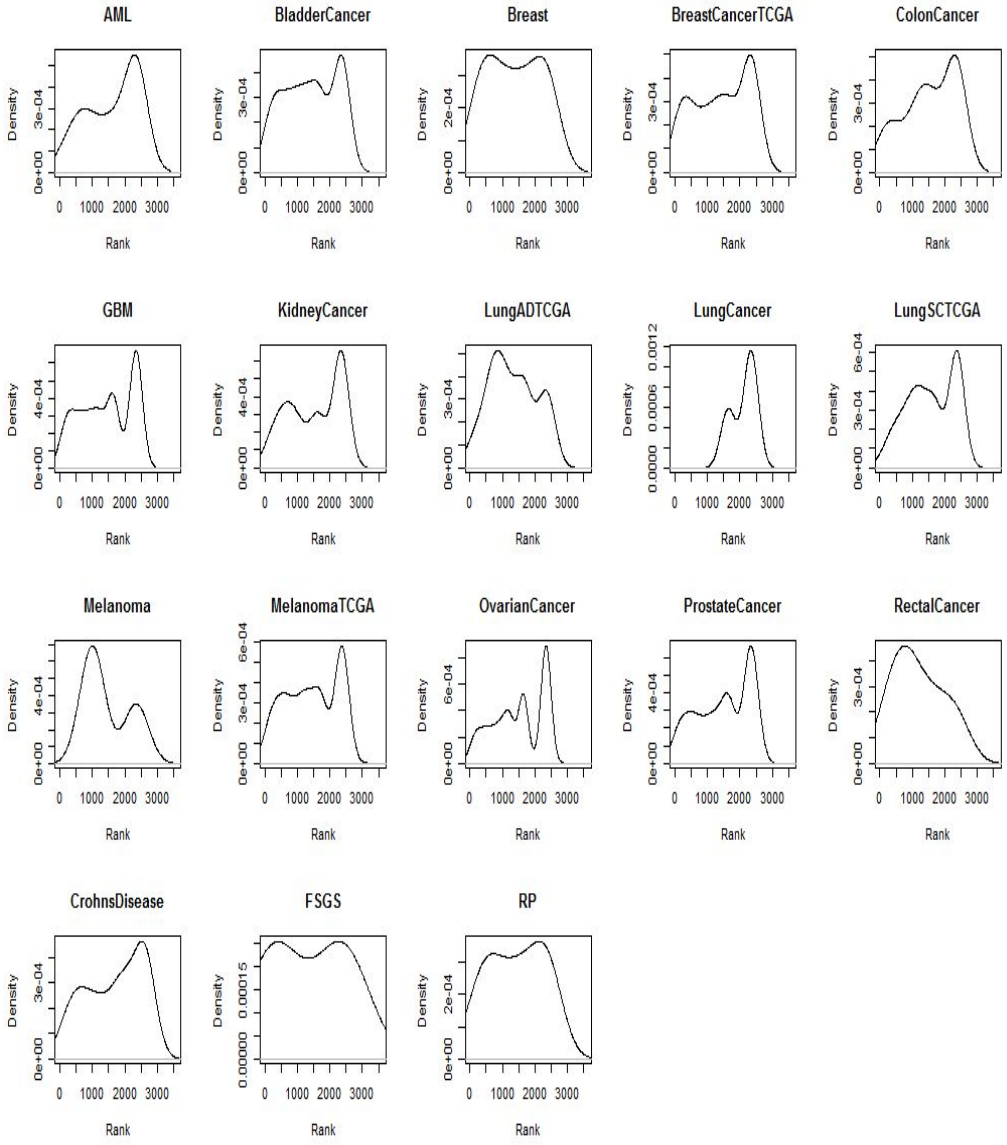

Figure 3.16 Distribution of rank among random diseases

In order to compare random disease terms with corresponding disease terms, I visualized random MeSH disease terms for each disease sequencing data of patients (Figure 3.16). As expected, distribution of rank showed inconsistent patterns. It is a certain extent supportive that this method can identify disease predisposition patterns of sequencing data.

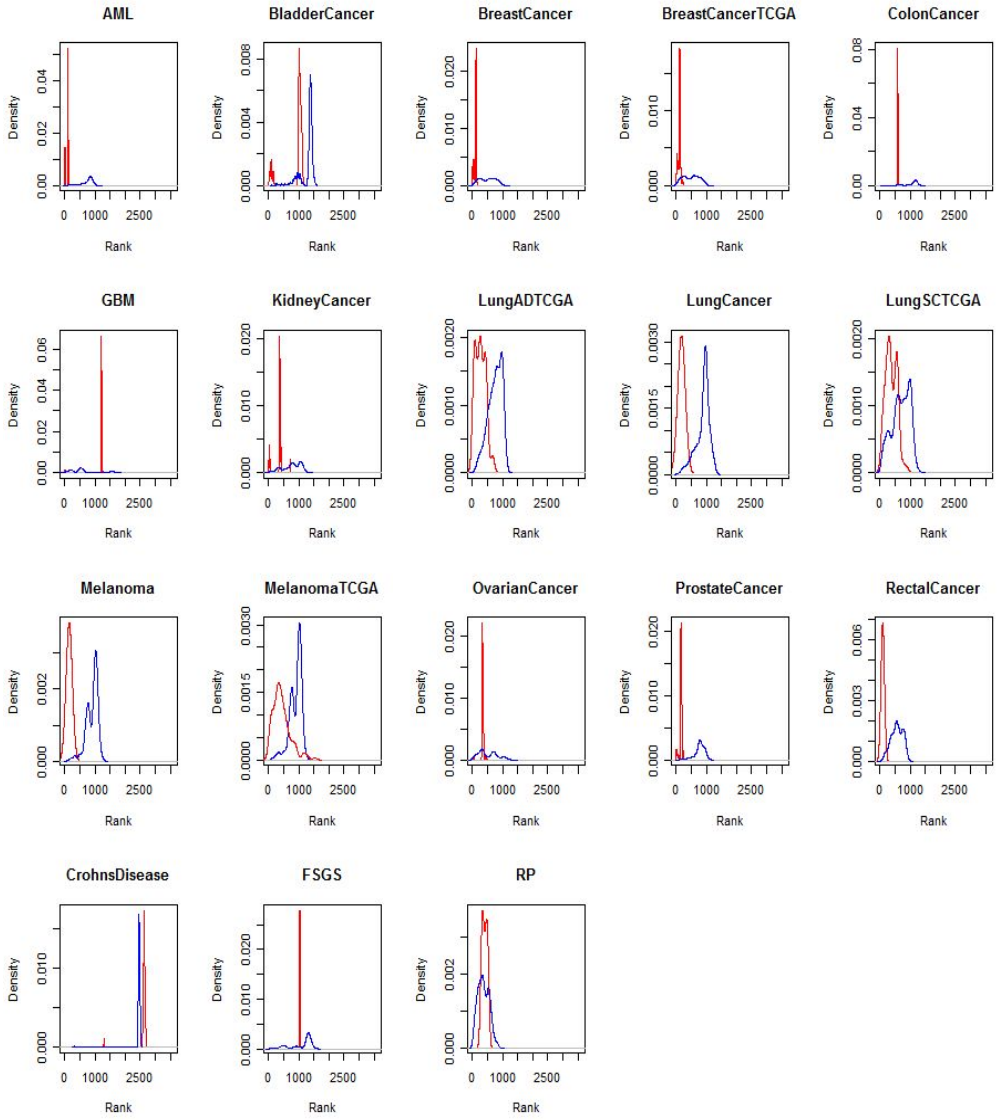

Figure 3.17. Comparison of rank between patients and average total healthy people in 1000 Genomes

To compare disease terms in patients groups with healthy groups, I showed the ranks of the same disease terms between disease group and 1000 Genome data (Figure 3.17). Red line indicates patients group and blue line indicates healthy group. In the most of cases of cancer diseases, the corresponding disease terms are highly ranked. On the other hand, non-cancer disease terms are low ranked or similar to the healthy people.

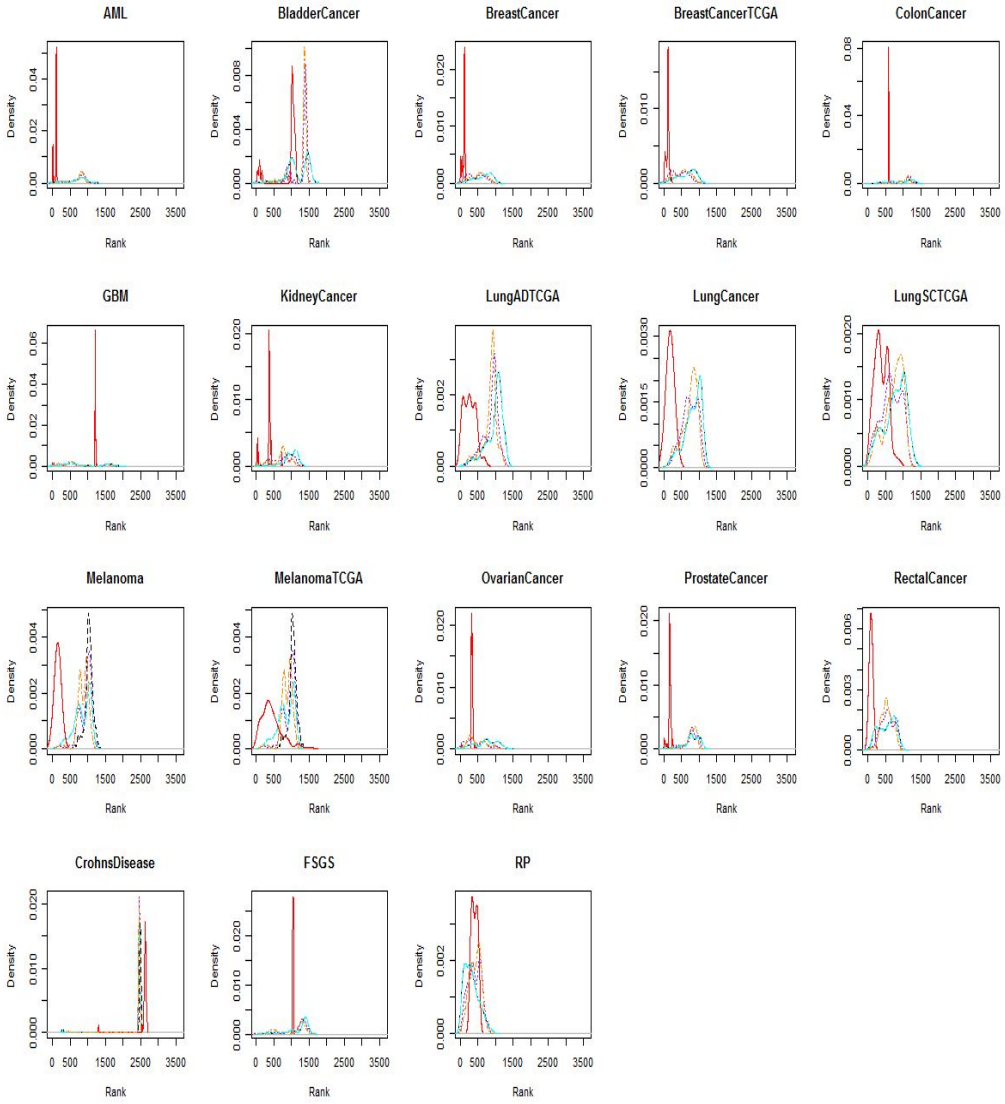

Figure 3.18. Comparison of rank between patients and healthy sub-population in 1000 Genomes

Next, I compared the ranks of the same disease terms between disease group and the sub-population in the 1000 Genome data (Figure 3.18). Consistently, cancer disease MeSH terms are highly ranked in the patient sequencing data. In addition, non-cancer disease terms are relatively somewhat low ranked or similar to healthy people.

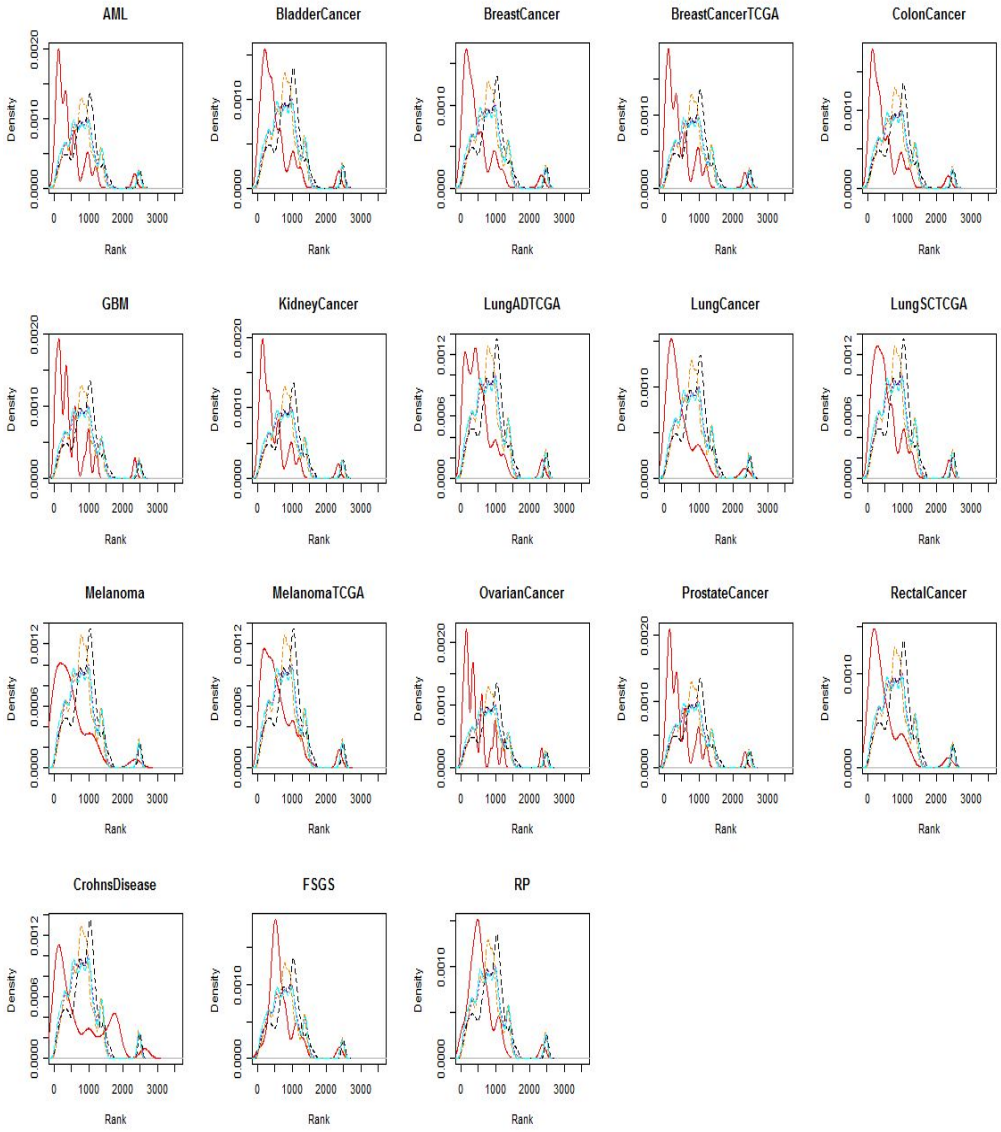

Figure 3.19. Distribution of rank of the total disease terms according disease data

To find whether the distribution of cancer disease terms are biased or not, I compared all the corresponding disease terms (total 60 MeSH disease terms) for each patient sequencing data. Consistently, the corresponding MeSH disease terms are highly ranked in disease groups compared with the healthy people in the 1000 Genome Project.

3.4. Disease rank patterns according to the tree extension

After OMIM-MeSH associations, MeSH tree has sparse mapped values. To overcome this problem, I applied tree extension using triad, in other words, parent node and children nodes of unmapped nodes in MeSH tree. The assumption is that MeSH tree is hierarchically organized. The MeSH disease terms of extended tree are highly ranked compared with direct mapping (Figure 3.20, Figure3.21). In other words, tree extension improves rank patterns of disease quenching data.

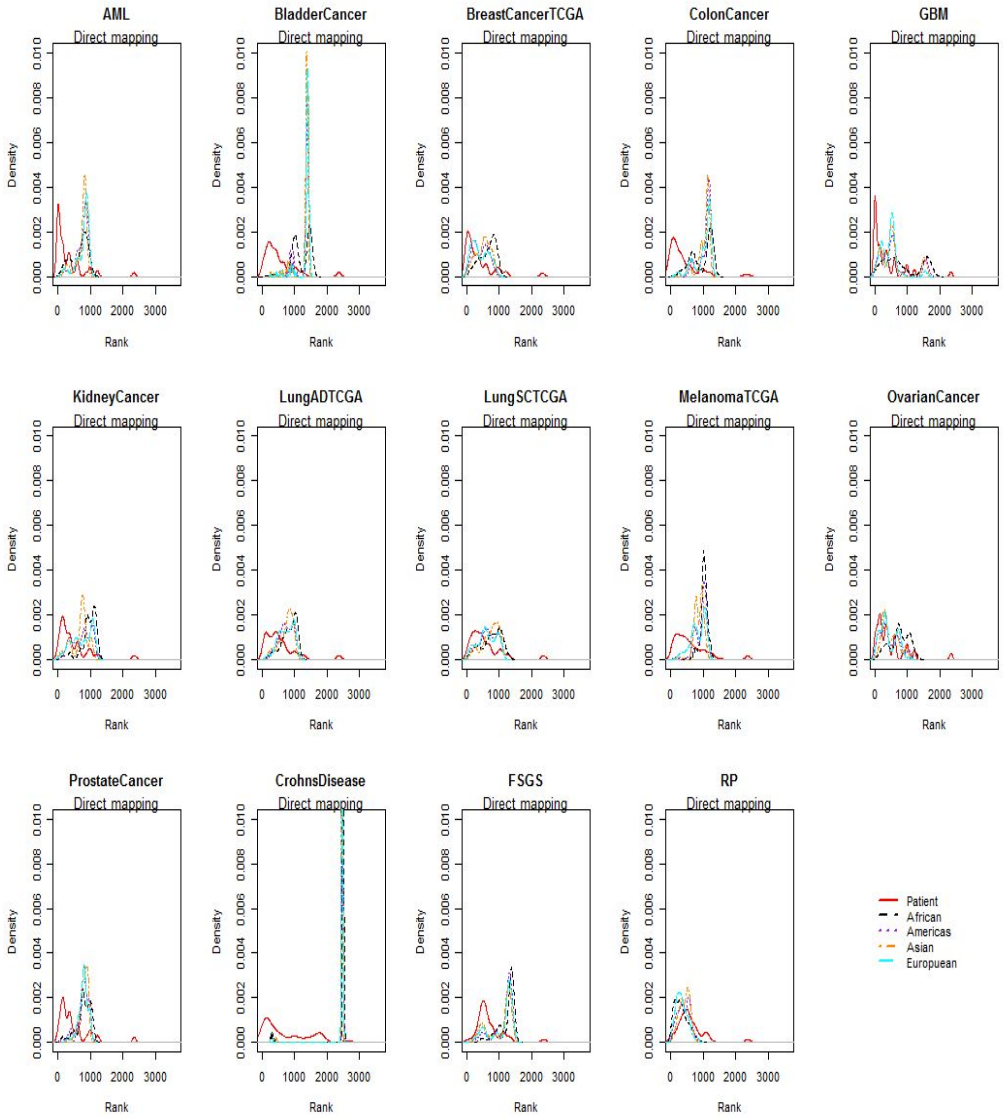

Figure 3.20. At direct mapping, distribution of rank of the corresponding diseases

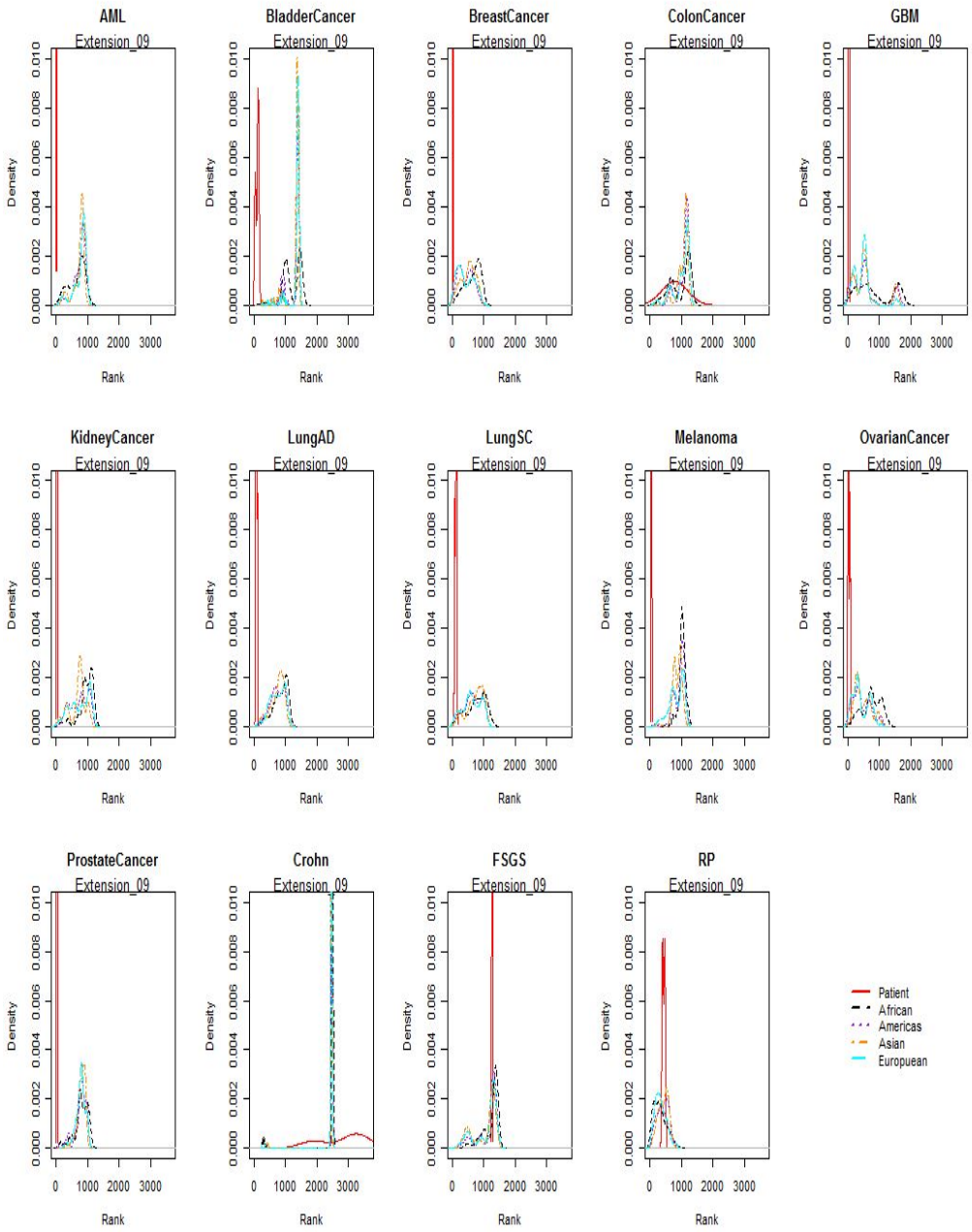

Figure 3.21. At tree extension, distribution of rank of the corresponding diseases

3.5. Disease enrichment analysis using the Diseasome

Although disease rank patterns are an indicator of disease predisposition, it is important to know what variants in the patients sequencing data are enriched. So using different disease knowledge source, “Diseasome”, I applied disease enrichment analysis for patient sequencing data and healthy people in the 1000 Genome Project. Figure 3.22 showed Cancer, Neurological, Endocrine disease categories clustered. The healthy people as well as the patients groups have variants related these categories.

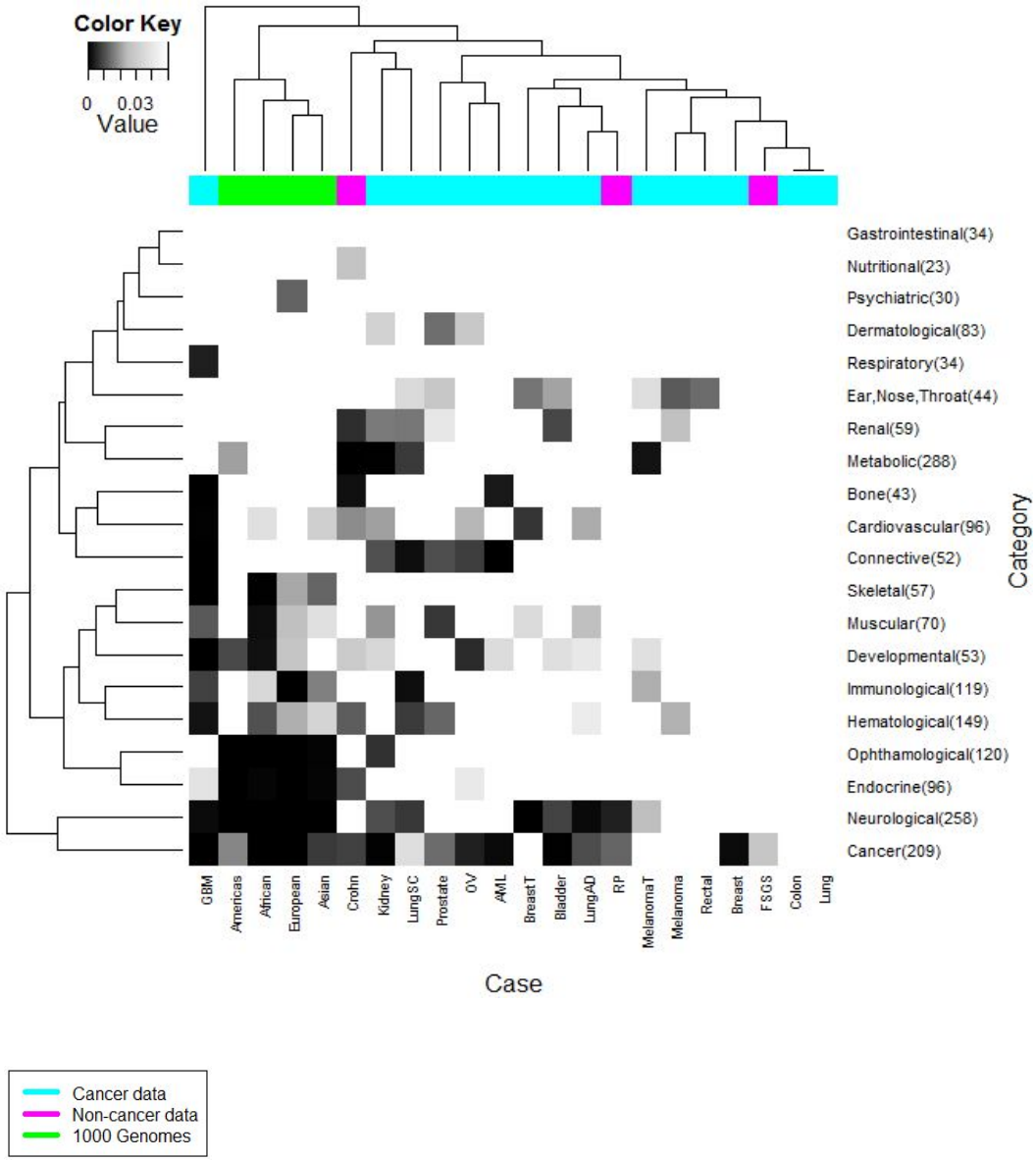

Figure 3.22. Enrichment Heatmap using Diseasome

Next, I used category-specific genes in order to identify category-specific enriched patterns (Figure 3.23). Different from category-overlapped genes (Figure 3.22), category-specific enrichment analysis showed Immunological category clustered into Cancer, Neurological and Endocrine. Specifically, the healthy people and disease groups are distinguished. However, cancer data and non-cancer data aren't distinguished so well.

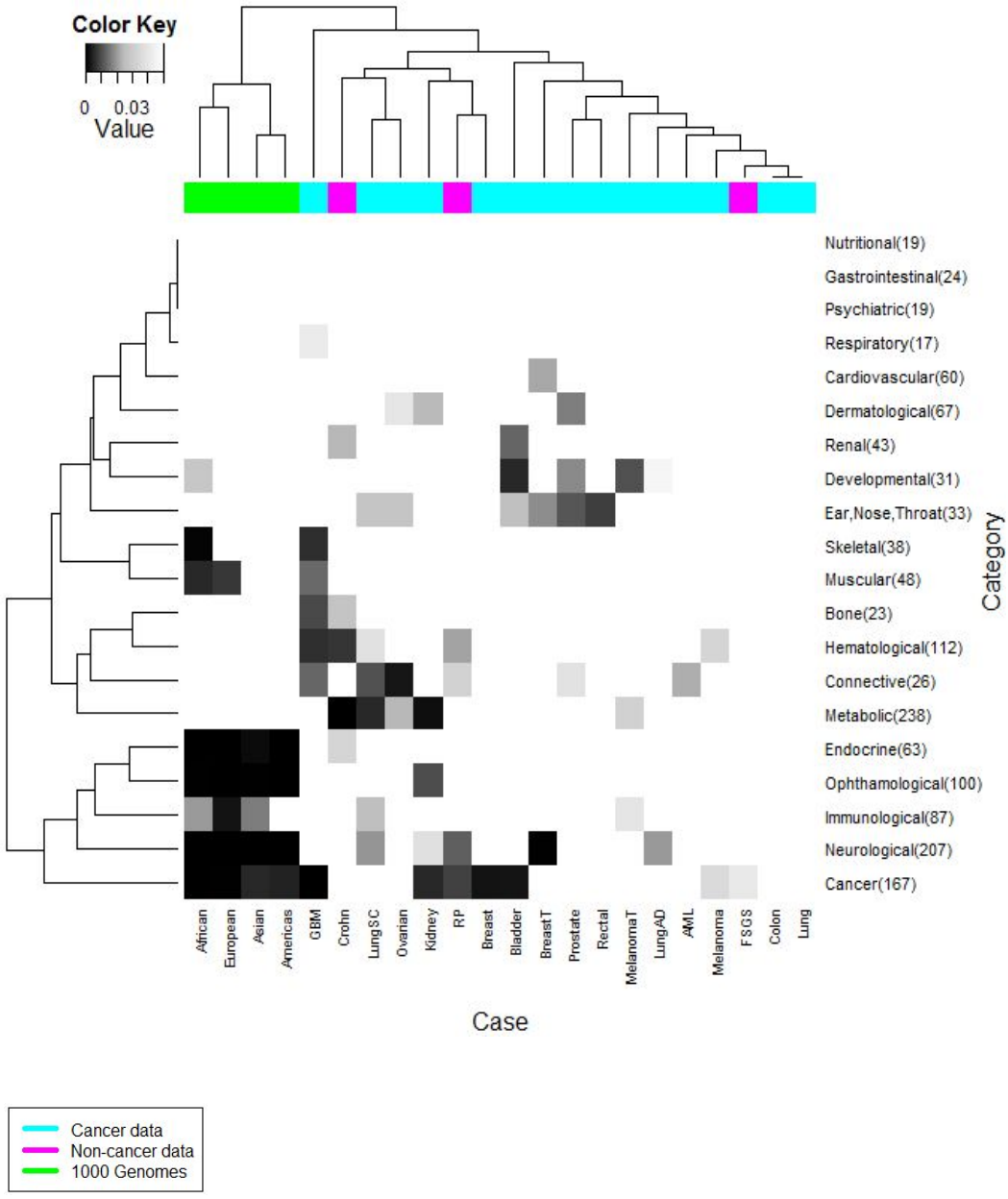

Figure 3.23. Enrichment Heatmap using category-specific genes in Diseasome

4. Conclusion

4.1. Summary

This research provides the first method to associate the personal genome with diseases in terms of individual genome sequencing studies. I present a computational method for associating variants in the personal genome sequencing data with predispositions to disease. The method works by ranking all variants in the personal genome as potential disease risks, and reporting MeSH terms that are significantly associated with highly ranked genes. To identify genetic variants associated diseases, I obtained high-throughput sequencing data in several cancer types (acute myeloid leukemia, bladder cancer, breast cancer, colon cancer, glioblastoma multiforme, kidney cancer, lung adenocarcinoma, lung squamous cell carcinoma, malignant melanoma, ovarian serous cystadenocarcinoma and prostate cancer) and non-cancer types (Crohn's disease, focal segmental glomerulosclerosis, and retinitis pigmentosa). From disease-gene association in the OMIM, I reconstructed relations of diseases and genes in the MeSH tree structures in order to consider the human disease hierarchical structure of human disease ontology.

The results showed the distribution of mutual information in the MeSH disease category differs according to the population in the healthy people. It suggests that in order to interpret personal genome properly, we may consider population

information together. In addition, MeSH disease terms are more highly ranked in the patients than healthy people. Disease-enrichment analysis showed Cancer, Neurological, Endocrine, and Immunological categories were over-represented in the patients as well as healthy people. Namely, it is possible to speculate systemic response patterns to diseases: Neuro-Endocrine-Immune Circuitry. In conclusion, although this study could not answer accurately the disease risk assessment, this study can provide data analysis scheme for the personal genome sequencing data. Rich disease-gene association knowledge can enable to accurately predict disease predisposition patterns. The scheme of this method has extendibility in genomic-based knowledge: drug-gene, environmental factor-gene and so on.

4.2. Future work

In the race for the $1,000 genome, the issue of the $1,000,000 interpretation has not been forgotten. Combing through millions of variants in a personal genome has presented numerous challenges for all parties involved: the physician looking to add genomic measurements to inform their diagnoses, the patients trying to figure out what they should be worried about, and the hobbyists interested in what their DNA means to them. Direct-to-consumer genetic testing companies such as 23andme, Lumigenix, and Navigenics offer a glimpse into the interpretation of a genome. These companies curate literature on gene-trait associations and provide attractive

user interfaces to navigating a personal genotype. However, the interpretations offered by these companies often differ, not because of inherent differences in technologies, but in which variants they choose consider in their calculations. While many have taken this fact to indicate a weakness of the genetic testing industry, it is also a reflection of the dynamic nature of the field.

Appendix

Table S.1. KRAS (No. of MeSH Codes=14)

Term	Code
Virus Diseases	C02
Neoplasms	C04
Musculoskeletal Diseases	C05
Digestive System Diseases	C06
Stomatognathic Diseases	C07
Respiratory Tract Diseases	C08
Nervous System Diseases	C10
Male Urogenital Diseases	C12
Female Urogenital Diseases and Pregnancy Complications	C13
Cardiovascular Diseases	C14
Congenital, Hereditary, and Neonatal Diseases and Abnormalities	C16
Skin and Connective Tissue Diseases	C17
Immune System Diseases	C20
Pathological Conditions, Signs and Symptoms	C23

Table S.2. CDH1 (No. of MeSH Codes=12)

Term	Code
Virus Diseases	C02
Neoplasms	C04
Musculoskeletal Diseases	C05
Digestive System Diseases	C06
Stomatognathic Diseases	C07
Nervous System Diseases	C10
Male Urogenital Diseases	C12
Female Urogenital Diseases and Pregnancy Complications	C13
Congenital, Hereditary, and Neonatal Diseases and Abnormalities	C16
Skin and Connective Tissue Diseases	C17
Immune System Diseases	C20
Pathological Conditions, Signs and Symptoms	C23

Table S.3. NRAS (No. of MeSH Codes=12)

Term	Code
Virus Diseases	C02
Neoplasms	C04
Musculoskeletal Diseases	C05
Digestive System Diseases	C06
Stomatognathic Diseases	C07
Cardiovascular Diseases	C14
Hemic and Lymphatic Diseases	C15
Congenital, Hereditary, and Neonatal Diseases and Abnormalities	C16
Skin and Connective Tissue Diseases	C17
Endocrine System Diseases	C19
Immune System Diseases	C20
Pathological Conditions, Signs and Symptoms	C23

Table S.4. BRAF (No. of MeSH Codes=12)

Term	Code
Virus Diseases	C02
Neoplasms	C04
Musculoskeletal Diseases	C05
Digestive System Diseases	C06
Stomatognathic Diseases	C07
Respiratory Tract Diseases	C08
Nervous System Diseases	C10
Cardiovascular Diseases	C14
Congenital, Hereditary, and Neonatal Diseases and Abnormalities	C16
Skin and Connective Tissue Diseases	C17
Immune System Diseases	C20
Pathological Conditions, Signs and Symptoms	C23

Table S.5. BRCA2 (No. of MeSH Codes=12)

Term	Code
Neoplasms	C04
Digestive System Diseases	C06
Nervous System Diseases	C10
Male Urogenital Diseases	C12
Female Urogenital Diseases and Pregnancy Complications	C13
Hemic and Lymphatic Diseases	C15
Congenital, Hereditary, and Neonatal Diseases and Abnormalities	C16
Skin and Connective Tissue Diseases	C17
Nutritional and Metabolic Diseases	C18
Endocrine System Diseases	C19
Immune System Diseases	C20
Pathological Conditions, Signs and Symptoms	C23

Bibliography

1. Ashley, E.A., et al., *Clinical assessment incorporating a personal genome.* Lancet, 2010. **375**(9725): p. 1525-35.
2. Chen, R., et al., *Personal omics profiling reveals dynamic molecular and medical phenotypes.* Cell, 2012. **148**(6): p. 1293-307.
3. Roberts, N.J., et al., *The predictive capacity of personal genome sequencing.* Sci Transl Med, 2012. **4**(133): p. 133ra58.
4. Sherry, S.T., et al., *dbSNP: the NCBI database of genetic variation.* Nucleic Acids Res, 2001. **29**(1): p. 308-11.
5. *The International HapMap Project.* Nature, 2003. **426**(6968): p. 789-96.
6. Church, D.M., et al., *Public data archives for genomic structural variation.* Nat Genet, 2010. **42**(10): p. 813-4.
7. Zhang, J., et al., *Development of bioinformatics resources for display and analysis of copy number and other structural variants in the human genome.* Cytogenet Genome Res, 2006. **115**(3-4): p. 205-14.
8. Amberger, J., et al., *McKusick's Online Mendelian Inheritance in Man (OMIM).* Nucleic Acids Res, 2009. **37**(Database issue): p. D793-6.
9. Mottaz, A., et al., *Easy retrieval of single amino-acid polymorphisms and phenotype information using SwissVar.* Bioinformatics, 2010. **26**(6): p. 851-2.
10. Stenson, P.D., et al., *Human Gene Mutation Database: towards a comprehensive central mutation database.* J Med Genet, 2008. **45**(2): p. 124-6.
11. Becker, K.G., et al., *The genetic association database.* Nat Genet, 2004. **36**(5): p. 431-2.
12. Mailman, M.D., et al., *The NCBI dbGaP database of genotypes and phenotypes.* Nat Genet, 2007. **39**(10): p. 1181-6.
13. Collins, F.S. and A.D. Barker, *Mapping the cancer genome. Pinpointing the genes involved in cancer will help chart a new course across the complex landscape of human malignancies.* Sci Am, 2007. **296**(3): p. 50-7.
14. Hudson, T.J., et al., *International network of cancer genome projects.* Nature, 2010. **464**(7291): p. 993-8.
15. Futreal, P.A., et al., *A census of human cancer genes.* Nat Rev Cancer, 2004. **4**(3): p. 177-83.
16. Forbes, S.A., et al., *COSMIC (the Catalogue of Somatic Mutations in Cancer): a resource to investigate acquired mutations in human cancer.* Nucleic Acids Res, 2010. **38**(Database issue): p. D652-7.
17. Gong, L., et al., *PharmGKB: an integrated resource of pharmacogenomic*

data and knowledge. Curr Protoc Bioinformatics, 2008. **Chapter 14**: p. Unit14 7.
18. Knox, C., et al., *DrugBank 3.0: a comprehensive resource for 'omics' research on drugs.* Nucleic Acids Res, 2011. **39**(Database issue): p. D1035-41.
19. Huss, J.W., 3rd, et al., *The Gene Wiki: community intelligence applied to human gene annotation.* Nucleic Acids Res, 2010. **38**(Database issue): p. D633-9.
20. Hoffmann, R., *A wiki for the life sciences where authorship matters.* Nat Genet, 2008. **40**(9): p. 1047-51.
21. *The ENCODE (ENCyclopedia Of DNA Elements) Project.* Science, 2004. **306**(5696): p. 636-40.
22. Bernstein, B.E., et al., *The NIH Roadmap Epigenomics Mapping Consortium.* Nat Biotechnol, 2010. **28**(10): p. 1045-8.
23. *A map of human genome variation from population-scale sequencing.* Nature, 2010. **467**(7319): p. 1061-73.
24. Lappalainen, I., et al., *dbVar and DGVa: public archives for genomic structural variation.* Nucleic Acids Res, 2012.
25. Hindorff, L.A., 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): p. 9362-7.
26. International Cancer Genome, C., et al., *International network of cancer genome projects.* Nature, 2010. **464**(7291): p. 993-8.
27. Thorn, C.F., T.E. Klein, and R.B. Altman, *Pharmacogenomics and bioinformatics: PharmGKB.* Pharmacogenomics, 2010. **11**(4): p. 501-5.
28. Ng, P.C. and S. Henikoff, *Predicting the effects of amino acid substitutions on protein function.* Annu Rev Genomics Hum Genet, 2006. **7**: p. 61-80.
29. Ramensky, V., P. Bork, and S. Sunyaev, *Human non-synonymous SNPs: server and survey.* Nucleic Acids Res, 2002. **30**(17): p. 3894-900.
30. Yue, P., E. Melamud, and J. Moult, *SNPs3D: candidate gene and SNP selection for association studies.* BMC Bioinformatics, 2006. **7**: p. 166.
31. Ferrer-Costa, C., et al., *PMUT: a web-based tool for the annotation of pathological mutations on proteins.* Bioinformatics, 2005. **21**(14): p. 3176-8.
32. Kumar, S., et al., *Phylomedicine: an evolutionary telescope to explore and diagnose the universe of disease mutations.* Trends Genet, 2011. **27**(9): p. 377-86.
33. Davydov, E.V., et al., *Identifying a high fraction of the human genome to be under selective constraint using GERP++.* PLoS Comput Biol, 2010. **6**(12): p. e1001025.
34. Dehal, P., et al., *The draft genome of Ciona intestinalis: insights into chordate and vertebrate origins.* Science, 2002. **298**(5601): p. 2157-67.
35. Margulies, E.H., et al., *Identification and characterization of multi-species*

conserved sequences. Genome Res, 2003. **13**(12): p. 2507-18.
36. Siepel, A., et al., *Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes.* Genome Res, 2005. **15**(8): p. 1034-50.
37. Cooper, G.M., et al., *Distribution and intensity of constraint in mammalian genomic sequence.* Genome Res, 2005. **15**(7): p. 901-13.
38. Asthana, S., et al., *Analysis of sequence conservation at nucleotide resolution.* PLoS Comput Biol, 2007. **3**(12): p. e254.
39. Prabhakar, S., et al., *Close sequence comparisons are sufficient to identify human cis-regulatory elements.* Genome Res, 2006. **16**(7): p. 855-63.
40. Cooper, G.M. and J. Shendure, *Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data.* Nat Rev Genet, 2011. **12**(9): p. 628-40.
41. Macintyre, G., et al., *is-rSNP: a novel technique for in silico regulatory SNP detection.* Bioinformatics, 2010. **26**(18): p. i524-30.
42. Andersen, M.C., et al., *In silico detection of sequence variations modifying transcriptional regulation.* PLoS Comput Biol, 2008. **4**(1): p. e5.
43. Zhao, Y., et al., *Prediction of functional regulatory SNPs in monogenic and complex disease.* Hum Mutat, 2011. **32**(10): p. 1183-90.
44. Woolfe, A., J.C. Mullikin, and L. Elnitski, *Genomic features defining exonic variants that modulate splicing.* Genome Biol, 2010. **11**(2): p. R20.
45. Wang, K., M. Li, and H. Hakonarson, *ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data.* Nucleic Acids Res, 2010. **38**(16): p. e164.
46. McLaren, W., et al., *Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor.* Bioinformatics, 2010. **26**(16): p. 2069-70.
47. Sana, M.E., et al., *GAMES identifies and annotates mutations in next-generation sequencing projects.* Bioinformatics, 2011. **27**(1): p. 9-13.
48. Shetty, A.C., et al., *SeqAnt: a web service to rapidly identify and annotate DNA sequence variations.* BMC Bioinformatics, 2010. **11**: p. 471.
49. Ge, D., et al., *SVA: software for annotating and visualizing sequenced human genomes.* Bioinformatics, 2011. **27**(14): p. 1998-2000.
50. Schwarz, J.M., et al., *MutationTaster evaluates disease-causing potential of sequence alterations.* Nat Methods, 2010. **7**(8): p. 575-6.
51. Murphy, S.N., et al., *Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2).* J Am Med Inform Assoc, 2010. **17**(2): p. 124-30.
52. *The Cancer Biomedical Informatics Grid (caBIG): infrastructure and applications for a worldwide research community.* Stud Health Technol Inform, 2007. **129**(Pt 1): p. 330-4.
53. Goecks, J., A. Nekrutenko, and J. Taylor, *Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent*

computational research in the life sciences. Genome Biol, 2010. **11**(8): p. R86.
54. Hull, D., et al., *Taverna: a tool for building and running workflows of services.* Nucleic Acids Res, 2006. **34**(Web Server issue): p. W729-32.
55. Gurwitz, D. and Y. Bregman-Eschet, *Personal genomics services: whose genomes?* Eur J Hum Genet, 2009. **17**(7): p. 883-9.
56. Ng, P.C., et al., *An agenda for personalized medicine.* Nature, 2009. **461**(7265): p. 724-6.
57. Foster, M.W., J.J. Mulvihill, and R.R. Sharp, *Evaluating the utility of personal genomic information.* Genet Med, 2009. **11**(8): p. 570-4.
58. Yang, Q., et al., *Using lifetime risk estimates in personal genomic profiles: estimation of uncertainty.* Am J Hum Genet, 2009. **85**(6): p. 786-800.
59. McGowan, M.L., J.R. Fishman, and M.A. Lambrix, *Personal genomics and individual identities: motivations and moral imperatives of early users.* New Genet Soc, 2010. **29**(3): p. 261-290.
60. Manolio, T.A., et al., *Finding the missing heritability of complex diseases.* Nature, 2009. **461**(7265): p. 747-53.
61. Pritchard, J.K. and N.J. Cox, *The allelic architecture of human disease genes: common disease-common variant...or not?* Hum Mol Genet, 2002. **11**(20): p. 2417-23.
62. Manolio, T.A., *Genomewide association studies and assessment of the risk of disease.* N Engl J Med, 2010. **363**(2): p. 166-76.
63. Hemminki, K., A. Forsti, and J.L. Bermejo, *The 'common disease-common variant' hypothesis and familial risks.* PLoS One, 2008. **3**(6): p. e2504.
64. Marchini, J., et al., *The effects of human population structure on large genetic association studies.* Nat Genet, 2004. **36**(5): p. 512-7.
65. Clayton, D.G., et al., *Population structure, differential bias and genomic control in a large-scale, case-control association study.* Nat Genet, 2005. **37**(11): p. 1243-6.
66. Mackay, T.F., *The genetic architecture of quantitative traits.* Annu Rev Genet, 2001. **35**: p. 303-39.
67. Stram, D.O., *Tag SNP selection for association studies.* Genet Epidemiol, 2004. **27**(4): p. 365-74.
68. de Bakker, P.I., et al., *Efficiency and power in genetic association studies.* Nat Genet, 2005. **37**(11): p. 1217-23.
69. *Comprehensive genomic characterization defines human glioblastoma genes and core pathways.* Nature, 2008. **455**(7216): p. 1061-8.
70. *Comprehensive molecular characterization of human colon and rectal cancer.* Nature, 2012. **487**(7407): p. 330-7.
71. *Comprehensive molecular portraits of human breast tumours.* Nature, 2012. **490**(7418): p. 61-70.
72. *Integrated genomic analyses of ovarian carcinoma.* Nature, 2011.

474(7353): p. 609-15.
73. Nickerson, D.A., V.O. Tobe, and S.L. Taylor, *PolyPhred: automating the detection and genotyping of single nucleotide substitutions using fluorescence-based resequencing.* Nucleic Acids Res, 1997. **25**(14): p. 2745-51.
74. Chen, K., et al., *PolyScan: an automatic indel and SNP detection approach to the analysis of human resequencing data.* Genome Res, 2007. **17**(5): p. 659-66.
75. Zhang, J., et al., *SNPdetector: a software tool for sensitive and accurate SNP detection.* PLoS Comput Biol, 2005. **1**(5): p. e53.
76. Bamford, S., et al., *The COSMIC (Catalogue of Somatic Mutations in Cancer) database and website.* Br J Cancer, 2004. **91**(2): p. 355-8.
77. Forbes, S.A., et al., *The Catalogue of Somatic Mutations in Cancer (COSMIC).* Curr Protoc Hum Genet, 2008. **Chapter 10**: p. Unit 10 11.
78. Forbes, S.A., et al., *COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer.* Nucleic Acids Res, 2011. **39**(Database issue): p. D945-50.
79. Pleasance, E.D., et al., *A comprehensive catalogue of somatic mutations from a human cancer genome.* Nature, 2010. **463**(7278): p. 191-6.
80. Pleasance, E.D., et al., *A small-cell lung cancer genome with complex signatures of tobacco exposure.* Nature, 2010. **463**(7278): p. 184-90.
81. Siva, N., *1000 Genomes project.* Nat Biotechnol, 2008. **26**(3): p. 256.
82. Overbeek, R., et al., *The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes.* Nucleic Acids Res, 2005. **33**(17): p. 5691-702.
83. Kaiser, J., *DNA sequencing. A plan to capture human diversity in 1000 genomes.* Science, 2008. **319**(5862): p. 395.
84. Via, M., C. Gignoux, and E.G. Burchard, *The 1000 Genomes Project: new opportunities for research and social challenges.* Genome Med, 2010. **2**(1): p. 3.
85. Kuehn, B.M., *1000 Genomes Project promises closer look at variation in human genome.* JAMA, 2008. **300**(23): p. 2715.
86. Li, H., J. Ruan, and R. Durbin, *Mapping short DNA sequencing reads and calling variants using mapping quality scores.* Genome Res, 2008. **18**(11): p. 1851-8.
87. Ning, Z., A.J. Cox, and J.C. Mullikin, *SSAHA: a fast search method for large DNA databases.* Genome Res, 2001. **11**(10): p. 1725-9.
88. McKenna, A., et al., *The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data.* Genome Res, 2010. **20**(9): p. 1297-303.
89. Depristo, A., *Science standards: averages deceive.* Science, 2010. **329**(5999): p. 1598; author reply 1599.

90. Le, S.Q. and R. Durbin, *SNP detection and genotyping from low-coverage sequencing data on multiple diploid samples.* Genome Res, 2011. **21**(6): p. 952-60.
91. Li, Y., et al., *MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes.* Genet Epidemiol, 2010. **34**(8): p. 816-34.
92. Goh, K.I., et al., *The human disease network.* Proc Natl Acad Sci U S A, 2007. **104**(21): p. 8685-90.
93. Vogelstein, B., D. Lane, and A.J. Levine, *Surfing the p53 network.* Nature, 2000. **408**(6810): p. 307-10.
94. Hainaut, P. and M. Hollstein, *p53 and human cancer: the first ten thousand mutations.* Adv Cancer Res, 2000. **77**: p. 81-137.
95. Olivier, M., et al., *The IARC TP53 database: new online mutation analysis and recommendations to users.* Hum Mutat, 2002. **19**(6): p. 607-14.
96. Ory, K., et al., *Analysis of the most representative tumour-derived p53 mutants reveals that changes in protein conformation are not correlated with loss of transactivation or inhibition of cell proliferation.* EMBO J, 1994. **13**(15): p. 3496-504.
97. Ogata, H., et al., *KEGG: Kyoto Encyclopedia of Genes and Genomes.* Nucleic Acids Res, 1999. **27**(1): p. 29-34.
98. Schubbert, S., et al., *Germline KRAS mutations cause Noonan syndrome.* Nat Genet, 2006. **38**(3): p. 331-6.
99. Niihori, T., et al., *Germline KRAS and BRAF mutations in cardio-facio-cutaneous syndrome.* Nat Genet, 2006. **38**(3): p. 294-6.
100. Burmer, G.C. and L.A. Loeb, *Mutations in the KRAS2 oncogene during progressive stages of human colon carcinoma.* Proc Natl Acad Sci U S A, 1989. **86**(7): p. 2403-7.
101. Almoguera, C., et al., *Most human carcinomas of the exocrine pancreas contain mutant c-K-ras genes.* Cell, 1988. **53**(4): p. 549-54.
102. Tam, I.Y., et al., *Distinct epidermal growth factor receptor and KRAS mutation patterns in non-small cell lung cancer patients with different tobacco exposure and clinicopathologic features.* Clin Cancer Res, 2006. **12**(5): p. 1647-53.
103. Semb, H. and G. Christofori, *The tumor-suppressor function of E-cadherin.* Am J Hum Genet, 1998. **63**(6): p. 1588-93.
104. Wong, A.S. and B.M. Gumbiner, *Adhesion-independent mechanism for suppression of tumor cell invasion by E-cadherin.* J Cell Biol, 2003. **161**(6): p. 1191-203.
105. Roberts, A., et al., *The cardiofaciocutaneous syndrome.* J Med Genet, 2006. **43**(11): p. 833-42.
106. Palomaki, G.E., et al., *EGAPP supplementary evidence review: DNA testing strategies aimed at reducing morbidity and mortality from Lynch syndrome.*

Genet Med, 2009. **11**(1): p. 42-65.
107. Wooster, R., et al., *Localization of a breast cancer susceptibility gene, BRCA2, to chromosome 13q12-13.* Science, 1994. **265**(5181): p. 2088-90.
108. Duncan, J.A., J.R. Reeves, and T.G. Cooke, *BRCA1 and BRCA2 proteins: roles in health and disease.* Mol Pathol, 1998. **51**(5): p. 237-47.
109. Yoshida, K. and Y. Miki, *Role of BRCA1 and BRCA2 as regulators of DNA repair, transcription, and cell cycle in response to DNA damage.* Cancer Sci, 2004. **95**(11): p. 866-71.
110. Cann, R.L., M. Stoneking, and A.C. Wilson, *Mitochondrial DNA and human evolution.* Nature, 1987. **325**(6099): p. 31-6.
111. Tanner, S.M., et al., *BAALC, the human member of a novel mammalian neuroectoderm gene lineage, is implicated in hematopoiesis and acute leukemia.* Proc Natl Acad Sci U S A, 2001. **98**(24): p. 13901-6.
112. Sumner, W.T., et al., *Follicular mucinosis as a presenting sign of acute myeloblastic leukemia.* J Am Acad Dermatol, 1998. **38**(5 Pt 2): p. 803-5.

초 록

차세대 염기서열 분석 방법의 도입은 기능 유전체학에 큰 반향을 일으켰다. 무제한적인 염기 서열 생산은 단 몇 주 만에 인간의 유전체를 분석 가능하게 했다. 이러한 차세대 염기 서열 분석 방법을 인간 유전체에 적용한 것은 몇 차례 연구된 바 있다. 특히 여러 인종에 걸친 1000 명의 대규모 염기 서열 분석 프로젝트인 1000 유전체 프로젝트와 개인 유전체 프로젝트에서 집중적으로 연구되었다. 암 유전체의 염기 서열 분석으로는 암의 발생과 진행과 관련된 변이들을 찾아내기도 하였다. DNA 염기 서열 정보의 해독의 핵심은 개인차 및 민족적 특성을 파악하거나 유전자 이상과 관련된 질환에서 염색체 이상을 포함한 선천성 원인의 규명과 당뇨병, 고혈압과 같은 복합질병의 유전자 결함을 찾기 위한 것이다. 또한 염기 서열 데이터는 유전자 발현, 유전자 다양성 및 그 상호작용 등의 정보들을 분자진단과 치료영역에서 폭넓게 활용할 수 있어 매우 중요하다. 염기 서열 분석과 함께 질병 및 관련 형질들의 정보가 함께 분석에 사용된다면 인간 유전체의 변이 발굴에 큰 향상을 가져올 것이다.

차세대 염기 서열 분석이 각광을 받고는 있지만, 현재 데이터의 분석 방법에는 어려움을 겪고 있다. 이전과는 달리 방대한 데이터를 얻을 수 있지만, 그 데이터에서 실제 임상적으로나 생물학적으로 의미 있는 결과를 도출하기 위해서는 이전의 데이터를 분석하는 방식으로는 다른 방법으로 접근해야 할 필요가 있는 것이다. 특히 환자로부터 수집된 유전체 정보는 서로 다른 증상이나 질병의 진행과 연관된 유전적인 특징을 규명하는데 중요하고 궁극적으로 치료법과 새로운 약물 개발로 연계될 것이다. 이는 차세대 염기 서열 기술의 일반화와 함께 개인 유전체 정보를 알 수 있다면 맞춤의료에 한발 더 다가설 수 있는 것은 사실이지만, 어디까지 질병의 원인이나 발병 등의 메커니즘을 이해하는 것으로 제한되고 해결방법에 대한 기능적인 정보가 미흡한 상태이다. 즉 각 질병과 연관된 많은 부분의 유전적 기능을 알지 못하면 개인 유전체 정보를 알고 있어도 최상의 개인맞춤 의료가 불가능하다. 개인 맞춤의료 실현을 목표로 체계적인

개인 유전체(Personal Genome) 시대의 준비와 동시에 대용량 바이오 정보들을 총체적으로 다룰 수 있는 정보학 및 유전학적 연구환경 조성도 중요하다.

본 논문에서는 개인 유전체 염기 서열 데이터의 변이 정보로부터 질병 위험도를 예측하는 방법론을 제시한다. 유전체가 갖고 있는 변이 정보에서 해당 변이가 단백질 생산에 얼마나 손상을 주는 지에 따른 효과를 이용하는 것이다. 이러한 유전체가 갖고 있는 변이 정보를 이용하여 질병 위험도를 계산하는 것은 이전에는 사용되지 않은 방법으로, 유전체의 직접적인 변이 정보를 활용한다는 점이 강점이다. 또한 기존의 질병과 유전자와의 지식을 이용하여 유전자를 매개로 하여 인간유전체와 질병과의 연관성을 찾고자 한다. 특히, 질병이 단순 구조가 아니라 질병 용어의 계층 구조(MeSH tree structure)를 이용하여 유전체와 질병 계층 구조 간의 패턴을 찾는다. 사용하는 질병 염기 서열 데이터로는 여러 암(급성 골수성 백혈별, 방광암, 유방암, 대장암, 다형신경교아종, 신장암, 폐 선암종, 편평상피성 폐암, 흑색종, 난소암, 전립선암)에 대해서 염기 서열 분석을 한 공개 데이터인 The Cancer Genome Atlas (TCGA) 및 비암 데이터(크론병, 국소분절사구체경화증, 색소성 망막염)를 획득하여 해당 질병 염기 서열 데이터에서의 질병 위험도의 패턴을 찾도록 한다. 질병 계층 구조를 이용한 질병 패턴을 찾기 위해 OMIM 지식 기반의 질병과 유전자의 연관 관계를 질병 용어의 계층 구조로 재구성하여 사용한다.

정상인 데이터에서 본 방법을 적용한 결과, 질병 카테고리 분포를 통해 인종별 질병 특성이 확연히 구분됨을 확인하였다. 즉, 이 방법론으로 개인 유전체 서열 데이터를 해석할 때 인종 별 차이를 고려하여 계산해야 보다 정확한 질병 분포를 계산할 수 있다. 그리고 특정 질병 서열 데이터에서 상위를 차지하는 질병이 해당 질병인데 반해, 동일한 질병이 정상인 서열 데이터에서는 그렇지 않은 패턴을 확인하였다. 또한 정상인 및 질병군의 서열 데이터 내의 질병 분포를 테스트한 결과, 암, 신경계, 내분비계, 면역계 질환들의 enrichment pattern이 유사함을 확인하였다. 이러한 결과들은 간접적이나마 본 방법론이 질병 위험도를 예측할 수 있는 도구로 사용될 수 있음을 시사하는 것이다. 한 명의 서열 데이터에서 관련 질환을 추론해내고자 하는 문제 제기는 서열 데이터 연구의 현 시점

에서 적절하다고 본다. 앞으로 데이터 구조 내의 해결해야 할 과제 및 기존의 질병 유전자 간 지식 정보가 향상된다면 개인 유전체와 질병 간의 관계를 측정하는데 있어 보다 정확한 결과를 제시할 수 있을 것이다. 본 연구의 질병-유전자 간의 분석 구조는 약물-유전자 및 환경인자-유전자 관계로써 확장 가능하므로, 개인 유전체 해석에 있어 그 범위를 확장하는 데에 기여할 수 있을 것이다.

주요어: 차세대 염기서열분석, MeSH 트리 구조, 질병위험도, 개인유전체

학번: 2004-23352

감사의 글

이 순간, 여기까지 인도해주신 하나님께 감사 드립니다. 내 능력이 아니라 하나님께서 주신 은혜로 여기까지 오게 되었습니다.

부족한 저에게 학문의 기회를 주시고 끊임없이 진심 어린 가르침을 주신 김주한 교수님께 감사와 존경을 표하고 싶습니다. 교수님께서 가르쳐주신 연구자로서의 마음가짐을 항상 마음에 품고 학자로서의 걸음을 한걸음씩 내딛도록 하겠습니다.

학위과정 가운데 함께 했던 많은 분들에게도 감사의 인사를 전합니다. 힘들고 지칠 때마다 힘주던 한 사람, 한 사람을 잊지 못할 것입니다. 동력자가 있었기에 다시 힘내어 연구할 수 있게 된 것들을 소중한 추억으로 삼겠습니다.

사랑하는 가족들에게 마음 깊이 감사를 드립니다. 항상 응원하고 지지해주는 가족들이 있었기에 여기까지 올 수 있게 되었습니다.

‘시작’이란 두려움과 기대를 동시에 갖고 있는 단어인 것 같습니다. 이제부터 학자로서 펼쳐질 제 삶의 ‘시작’에서, 가르침을 주신 분들의 귀한 가르침을 따르며 제대로 된 연구자로서의 삶을 살아내도록 하겠습니다.