D-GRIP: DNA Genetic Risk Information Profile

A genotype analysis system to predict a genetic risk profile for an individual

by

Siddhartha Srivastava

B.Sc., Biological Science, The University of Calgary, 2005
B.Sc., Computer Science, The University of Calgary, 2005

A THESIS SUBMITTED IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF

Master of Science

in

The Faculty of Graduate Studies

(Bioinformatics)

The University Of British Columbia

October, 2007

Abstract

New genotyping technologies are producing reliable results with far greater coverage and at dramatically lower cost than previously possible. Given the rapid new discovery of disease associated markers and the new technology for determining the nucleotide sequences of key positions in the DNA of an individual, it is now feasible to apply existing knowledge to generate personalized analyses of genetic risk for diverse diseases. DNA Genetic Risk Information Profile (D-GRIP) is a genotype analysis software system that determines an individual's genetic risk profile given a genotype. The prototype web tool can take, as input, up to a million observed genotypes from single nucleotide positions known to be polymorphic in a human population. The submitted genotype data are compared to a database of disease associated single nucleotide polymorphisms (SNPs) and an output is generated, reporting disease-associated variants for which the individual has a predicted modified risk.

An evaluation of D-GRIP was performed through the direct surveying of potential users of such a system - users such as clinicians, genetic counselors and genetics researchers. Due to ethical issues related to providing a genetic risk profile, the prototype system is kept closed to the general public and reserved for research into the utility and requirements of such software.

The major conclusions drawn direct attention towards the key limitations presently precluding the creation of personalized genetic risk assessment. The lack of computationally exploitable resource for disease associated genetic variants, the inherent statistical complexities involved with risk calculation for large-scale genotyping data and the limited understanding of interactions between genes, environment and complex diseases, are all key factors that need to be overcome in order to create a practical genetic risk assessment tool.

Table of Contents

Appendices

List of Tables

List of Figures

Acknowledgments

I would like to acknowledge my academic supervisor, Dr. Wyeth W. Wasserman. Through his continuous support and guidance I have gained valuable insights in how to conduct and present research. I would also like to acknowledge my thesis committee: Francis Ouellette and Dr. Jan M. Friedman for their advice and support.

In addition, I thank Drs. Cornelius Boerkoel, Angela Brooks-Wilson, Lorne Clarke, Denise Daley, Anita Dircks, Bill Gibson, Jinko Graham, Millan Patel, Colin Ross for providing valuable feedback regarding utility and shortcomings of D-GRIP. I would like to extend further acknowledgments to Drs. William Dana Flanders, Muin Khoury and Quanhe Yang for their feedback on the statistical model. I want to also thank Francis Ouellette and Dr. Artem Cherkasov for giving their advice and encouragement during my training in the CIHR/MSFHR Strategic Training Program in Bioinformatics.

I would like to thank the members of the Wasserman Laboratory: David Arenillas, Jochen Brumm, Warren Cheung, Alice Chou, Debra Fulton, Shannan J. Ho Sui, Andrew Kwon, Jonathan Lim, Stuart Lithwick, Dora Pak, Elodie Portales-Casamar, Magdalena Swanson, Amy Ticoll, Tony Wong and Dimas Yusuf, for creating a friendly and enjoyable research environment.

I greatly appreciate the financial support that was provided by the Canadian Institutes of Health Research (CIHR) and the Michael Smith Foundation for Health Research (MSFHR) Strategic Training Program in Bioinformatics.

Chapter 1

Introduction

This thesis describes the exploration of how bioinformatics can be applied in the field of genetics, specifically to the prediction of disease risk. The causes of human diseases range from simple Mendelian inheritance patterns to complex combination of genetic and non-genetic (environmental) factors. With the availability of the entire human genome sequence and the common variation map (HapMap project), the understanding of genetic contributions to diseases is increasing rapidly. We are approaching a time where prediction of disease risk on a personalized level will become a reality.

1.1 Variations and Diseases

Variations in DNA sequences occur throughout the genome at a frequency of approximately 4-5 in 1000 bases ($0.4 - 0.5\%$) on average between two unrelated individuals [3]. These differences or variations in sequences include both mutations and polymorphisms, which are distinguished by their frequency within a population. Mutations are by definition rarely observed in a population and while they can cause disease, are not generally relevant to the prediction of disease risk in the general population. The simplest and most common form of polymorphism is called a Single Nucleotide Polymorphism (SNP). At a particular site on the human genomic sequence, a

SNP is defined by the existence of a certain percentage of individuals with a nucleotide differing from the norm. For instance, in two copies of a chromosome at one site, one chromosome might have an A at that position (the 'A' allele) and the other might have a C (a 'C' allele). The minimum threshold percentage for classifying a position as being a SNP rather than a mutation is generally defined as 1%of tested choromosomes, although some reports use other values. In the human populations, there are approximately 10 million SNPs that occur with greater than 1% frequency and these 10 million sites constitute 90% of the variation in the population [3, 21]. In short, SNPs constitute a dramatic portion of the genetic variation between two individuals. A genotype is then defined as the combination of the two alleles at a particular locus for a given SNP. For instance, at a known polymorphic position with A and C forms, genotypes would be AA, AC or CC. SNPs occur throughout the genome (promoter region, coding and intronic regions) where those variations situated in proten coding regions are of two types, synonymous (not altering the encoded amino acid sequence) and non-synonymous (causing a change to the encoded amino acid sequence).

In the study of human genetics there have been a litany of examples of links between sequence variations (also referred to as markers) and specific traits or diseases [27]. Disorders where genetics plays an important role, the so called genetic diseases, can be classified into single gene defects, chromosomal disorders or multifactorial. Single gene disorders (or Mendelian disorders) such as Cystic Fibrosis, are usually rare and identifying the causal genetic variant has helped understand the disease. Chromosomal disorders are caused by excess or deficiency of genes [8]. Most common diseases are multifactorial such as diabetes or heart disease and it is generally accepted that these phenotypic effects are based on direct genetic effects, multiple gene-gene interactions and gene-environment interactions [27, 30]. Recently, through new technologies and genome-wide association surveys, there has been a strong effort towards finding disease susceptibility variations (especially SNPs) for complex disorders [13].

1.2 Discovery of new markers

Recently, there has been a surge in new discovery of disease susceptibility genes and variations. Traditionally, in human genetics, a discovery involved identifying a gene for susceptibility of disease. That notion, however, comes from working on rare diseases in which single studies have reported strong statistical associations between a mutation in a gene and a disease [13]. In contrast, for common diseases, the oligogenic model is usually accepted. The model states that the genetic component of complex diseases are more likely to be a result of a few genes with moderate effect or a large number of genes with smaller effect [11]. With the development of large-scale genotyping technologies, it has now become feasible to perform genome-wide association studies [11, 13] to identify contributing loci by surveying a large set of known variable sites.

Several large-scale genome-wide association studies have been recently published, including studies of diabetes Mellitus type II [26, 28, 31, 33], bipolar disorder [1], Alzheimer's disease [4], Crohn's (inflammatory bowel) disease [6, 22] and coronary artery disease [24]. Given the small sample of diseases listed here and the short timeframe in which they were published, a large number of markers are being discovered at a very rapid rate. A more detailed analysis on the recent advances of genome wide association studies and a count of newly discovered markers for several common diseases can be found in [5].

1.3 Genotyping technologies

New genotyping technologies are driving the burst of genetic studies. For studies where a small number of SNPs are analyzed, Sequenom's MassARRAY® system, TaqMan® and Pyrosequencing™ have been widely used. These methods provide flexibility in study design for investigators prepared to work on a small set of candidate genes. For studies where thousands of SNPs need to be analyzed simultaneously (i.e., multiplexed) for each sample, platforms such as the Illumina BeadArray™ and the Affymetrix GeneChip® can be used. These systems have dramatically increased the throughput of genotyping and substantially reduced genotyping costs [23].

To illustrate the underlying technology, a brief description of the original Illumina BeadArray™ platform and the GoldenGate™ assay follows. The array-based technology comes in a 96 well plate format. Each well contains an optical fiber bundle where an array of 50,000 randomly placed beads, each $\sim$3 microns in diameter, exist. There are 1520 bead types, each representing a different oligonucleotide sequence. This gives $\sim$30 copies of each bead type providing (on average) 30 replicate genotyping experiments for each SNP and can screen up to 100,000 genotypes in one sample [10].

The GoldenGate® Assay is used with the BeadArray platform and has the advantage of allowing high multiplexing during amplification steps while minimizing reagent volumes and time. Genomic DNA is normalized and then chemically reacted to incorporate biotin to make activated DNA. Three oligonucleotides are designed for each SNP. Two are allele-specific oligonucleotides (ASO) and one is locus-specific oligonucleotide (LSO). Each ASO has a 3' base complementary to one of the two SNP alleles. The LSO hybridizes downstream of the ASOs. Each of the three oligonucleotide sequences contain regions of genomic complementary for polymerase chain reaction (PCR): P1 and P2 on the ASOs and P3 on the LSO. The LSO also contains a unique address sequence that targets a particular bead type on the well plate. After extension and ligation, activated genomic DNA is amplified using PCR and labeled P1 and P2. The primers P1 and P2 are labeled with Cy3 and Cy5 respectively. The PCR products are then hybridized to array matrix plate where the Cy5 and Cy3 labeled materials bind in proportion to the relative abundance of the two alleles in the sample such that a homozygote for the allele has only one color and a heterozygote has two. The labels are detected and analyzed using the fluorescence signal and using software for genotype clustering and calling. Based on the color distribution of each allele, the genotype of the samples for the designated SNPs can be determined. For a more thorough and detailed description of the assay, refer to [19] and [32].

Both Illumina and Affymetrix systems have challenged the technological limit of genotyping analysis. For instance, Illumina's Sentrix® HumanHap650Y BeadChip and whole-genome Human1M BeadChip can respectively genotype over 650,000 tag SNPs and over one million genetic variations on a single array, whereas the Affymetrix's GeneChip® Genome-wide human SNP array 5.0 can genotype approximately 500,000 SNPs in one sample. Both platforms can genotype fixed set of SNPs as well as customized panels of SNPs. Illumina's SNP selection is based on the HapMap project while Affymetrix's SNPs selection is based on feasibility of SNPs to be genotyped. For both systems, the cost of genotyping is less than $0.01 per SNP. A general recent summary of the various methods is shown in Table 1.1. A more detailed review of various genotyping technologies is available in [32] and [23]. Given the new technologies and the high throughput of genotypes at substantially low costs, genotyping an individual has become increasingly feasible and led to a shift from investigation of a few candidate polymorphisms at a time to comprehensive whole-genome studies [23].

1.4 Bioinformatic Tools

There are many different open source and commercial systems available that manage, organize and analyze large-scale genotype data and/or provide risk assessments for disease. In order to determine whether any currently available systems integrate the analysis of many genotypes to provide personalized risk assessments for diseases, a survey of the risk prediction systems follows.

	Assay design	Multiplexing capability	Throughput (no. of samples per assay)	Cost per genotype
TaqMan®	By manufacturer or investigator	No	Up to 10,000+	>US$0.30
Pyrosequencing™	By investigator	1 to 3	Up to 4,000+	>US$0.30
Sequenom's MassARRAY®	By investigator	1 to 29	Up to 3,000+	US$0.05-0.10
Illumina's Sentrix®	By manufacturer	1,536 to 1,000,000	Up to 96	<US$0.01
Affymetrix's GeneChip®	By manufacturer	10,000 to 500,000	Up to 96	<US$0.01

Table 1.1: A summary of genotyping technologies currently available: The cost per genotype is an estimate of maximal multiplexing capability. A note, Illumina's Sentrix® numbers in the table are based on the Human1M BeadChip which will be released in the second quarter of 2007.

1.4.1 Commercial Systems

Genetics and genetic testing companies such as GeneSage [16], GeneTracks [17] and Genelex [14], provide or attempted to provide a variety of products and services. For instance, GeneSage, which now appears defunct, offered secure storage of genetic information for its users as well as access to genetic information and clinical information on genetic medicine for health professionals such as physicians and nurses. Also, risk assessments for specific diseases were provided through a team of in-house genetic counselors. An advantage of GeneSage was that risk assessments were provided by qualified genetic counselors, but the assessments were not based on genotype information.

GeneTracks, on the other hand, provides various forms of DNA testing such as Paternity, Twin or Sibship and Maternity. The strength of Gene-

Tracks lies in its DNA testing capability while the disadvantage is the lack of genetic assessment. In addition, two facets of GeneTracks are the DNA Bank and DNA Ancestry project. The DNA Bank acts as a storage facility for the customer's genetic data while the DNA Ancestry project provides a way to trace an individual's ancestry based on 20-40 Y-chromosome DNA markers. One advantage of such a service is the incorporation of genotype data in tracing ancestry but the disadvantages are the lack of genetic risk assessment and the lack of flexibility because only males can be tested since the test uses markers from Y-chromosome.

Lastly, Genelex provides a diverse range of services. For health professionals, genetic information, drug information, pharmacogenetic testing for specific drugs and nutrigenetic tests (dietary consultation) are provided. Also for clinicians, a software called GeneMedRx, which provides drug-drug and drug-gene interaction risk prediction for cytochrome P450 metabolism and genetic testing [15]. For the general public, adverse drug reaction testing, nutritional testing (dietary consultation), ancestry DNA testing and predictive testing for four diseases are provided. All the testing services utilize genetic information from the customer and test a set of known genotypes, genes or set of phenotypes. One advantage of Genelex is the GeneMedRx software. It incorporates genetic testing with risk prediction to ensure drug efficacy and prevent adverse drug reaction. One disadvantage is that GeneMedRx only incorporates one genetic test with risk prediction (the cytochrome 450P metabolism).

1.4.2 Open Source Systems

There are many open source systems that provide management and analysis of genotype data and a disease risk assessment. For brevity, only recently published tools will be discussed. The open source systems can be broken down into three categories. There are data management tools, visualization tools and risk assessment tools.

In the realm of data management, IGS, Integrated Genotype Analysis [12], stores, edits and analyzes genotype and phenotype data. IGS can handle large-scale genotype data, stores the data and meta data in various formats and can be used for genetic analysis (e.g. pedigree checks, Hardy-Weinberg tests, allele frequency tests, etc). The system is freely available on-line and the underlying database structure can be easily re-created. IGS is useful for storing raw genotype as well as processed genotype data (simply the genotype and the sample). Another tool is called SNPP, Single Nucleotide Polymorphism Processor [36]. SNPP's strength lies in handling massive amounts of raw SNP genotyping data, using a backend database framework for storage and it can also be used as a tool for data format conversion. The disadvantage lies in the minimal analysis of the genotype data since it only provides Mendelian inheritance checks for SNP data obtained from families.

For visualization tools, there are several programs which provide an integrated environment for visualization and analysis of genotype data. SNP-VISTA, an interactive SNP visualization tool [29] allows visualization of large-scale genotype data for disease related genes. The software maps SNPs to gene structure, classifies SNPs based on location, frequency and allele composition, clusters SNPs according to user criteria and includes protein evolutionary conservation visualization. The strength of SNP-VISTA is the graphical interface and visual representation of large scale data. SNPAnalyzer, a workbench for SNP analysis [35], performs data manipulation, statistical analysis on genotype data and visualization. Another recent tool is GEVALT, GEnotype Visualization and ALgorithmic Tool [7], which provides phasing and tag SNP selection algorithms, along with visualization of LD plots and haplotype data. All of the functionality is available in one integrated viewer. The advantage of GEVALT is in the integration of analysis tools and the visualization in one environment. There are other visualization tools that provide various features but are not mentioned here. All the visualization software provides analysis of genotype data but does not provide any disease risk assessments.

Risk assessment tools can be broken down into two categories, non-family-based and family-based. For non-family-based risk assessments, the tools are classified as expert systems or knowledge-based systems. An expert system is a computer system, based on artificial intelligence(AI) principles, which uses an organized body of knowledge, heuristics and inference to suggest solutions in a particular domain of expertise, for instance in medicine. A review of various expert systems and currently used systems is done in [25] and [18]. Therefore, for brevity, only one of the originally developed systems will be mentioned here. MYCIN [2] was developed to provide assistance to physicians in the diagnosis and treatment of meningitis and bacterial infections. MYCIN conducts a question and answer dialog where it ask questions such as suspected sites of infection, symptoms and results of other laboratory tests. Then, MYCIN recommends a course of antibiotics and can also provide its reasoning behind its answers. The advantage of an expert system is its diagnostic support capability to a physician. A potential disadvantage is the purely computational basis of prediction and no incorporation of genetic history in diagnosis of diseases.

For family-based risk assessments, there are many tools available, of which the majority target cancers. A tool for prediction of breast cancer risk is BRCAPRO [9]. The BRCAPRO model incorporates information on all family members (affected and unaffected) for breast and ovarian cancer and then calculates the probability of carrying the BRCA gene mutation using Bayes theorem. BRCAPRO's strength is its accuracy to predict BRCA gene mutation. BRCAPRO was validated by comparing to genetic counselors and it was found that BRCAPRO had similar sensitivity and higher specificity to experienced genetic counselors in identifying BRCA mutation carriers. A similar system has been created for identifying high risk individuals of familial pancreatic cancer called PancPRO [34]. The underlying framework of PancPRO is similar to BRCAPRO. Again, a validation of PancPRO indicated its accuracy in risk assessment. In a recent review [20], a set of cancer risk assessment tools (CRATs) which were available on the Internet. The five tools discussed in the paper determined the risk of various types of cancers based on family history. One of the disadvantages of these tools is the focus on purely familial-based Mendelian model diseases and not on other more complex diseases such as Diabetes Mellitus or Alzheimer's.

1.5 Overview of project

Given the rapid new discoveries of disease-associated markers and the advent of new genotyping technology, a question arises: is it now possible to apply existing knowledge of genetic diseases to create disease risk profiles for individuals? This thesis project was motivated by such a question and was designed to ascertain the bioinformatic limitations that must be overcome to facilitate a genotyping-based analysis of disease risk. We created a web tool called D-GRIP, DNA Genetic Risk Information Profile, which is a genotype analysis system that determines an individual's genetic risk profile given a genotype as input. The on-line tool can take, as input, up to one million observed genotypes from known SNPs in human populations. The submitted genotype data are then compared to validated disease-associated SNPs (a DNA-Disease database) and then outputs a list of diseases for which the individual has modified (up or down) risk. D-GRIP is intended to serve as an early prototype of a prognostic tool for use by genetic counselors. D-GRIP went through a testing phase where clinical geneticists, genetic counselors, genetic researchers and biostatisticians were consulted on the utility of D-GRIP and their feedback was recorded. One major conclusion drawn from the project is that the level of current knowledge for disease-causing SNPs is limited. There are only a few diseases that had strong supporting evidence causally linking SNPs to the disease. Given this scarcity of data, substantial studies on disease-causing variations are needed, especially for complex diseases.

Bibliography

[1] A E Baum, N Akula, M Cabanero, I Cardona, W Corona, B Klemens, T G Schulze, S Cichon, M Rietschel, M M Nöthen, A Georgi, J Schumacher, M Schwarz, R Abou Jamra, S Höfels, P Propping, J Satagopan, S D Detera-Wadleigh, J Hardy, and F J McMahon. A genome-wide association study implicates diacylglycerol kinase eta (dgkh) and several other genes in the etiology of bipolar disorder. *Mol Psychiatry*, May 2007.

[2] Bruce G. Buchanan and Edward H. Shortliffe. *Rule-Based Expert Systems: The MYCIN Experiments of the Stanford Heuristic Programming Project.* AAAI Press, Available at: http://www.aaaipress.org/Classic/Buchanan/buchanan.html, ebook edition, 1984.

[3] International Hapmap Consortium. The international hapmap project. *Nature*, 426(6968):789–96, December 2003.

[4] Keith D Coon, Amanda J Myers, David W Craig, Jennifer A Webster, John V Pearson, Diane Hu Lince, Victoria L Zismann, Thomas G Beach, Doris Leung, Leslie Bryden, Rebecca F Halperin, Lauren Marlowe, Mona Kaleem, Douglas G Walker, Rivka Ravid, Christopher B Heward, Joseph Rogers, Andreas Papassotiropoulos, Eric M Reiman, John Hardy, and Dietrich A Stephan. A high-density whole-genome association study reveals that apoe is the major susceptibility gene for sporadic late-onset alzheimer's disease. *J Clin Psychiatry*, 68(4):613–8, April 2007.

[5] Jennifer Couzin and Jocelyn Kaiser. Genome-wide association. closing the net on common disease genes. *Science*, 316(5826):820–2, May 2007.

[6] J R Fraser Cummings, Rachel Cooney, Saad Pathan, Carl A Anderson, Jeffrey C Barrett, John Beckly, Alessandra Geremia, Laura Hancock, Changcun Guo, Tariq Ahmad, Lon R Cardon, and Derek P Jewell.

Confirmation of the role of atg16l1 as a crohn's disease susceptibility gene. *Inflamm Bowel Dis*, April 2007.

[7] Ofir Davidovich, Gad Kimmel, and Ron Shamir. Gevalt: an integrated software tool for genotype analysis. *BMC Bioinformatics*, 8:36, 2007.

[8] A.D.A.M. Medical Encyclopedia. Genetics. [updated 2005 apr 20; cited 2007 may 2007] available at: http://www.nlm.nih.gov/medlineplus/ency/article/002048.htm, May 2007.

[9] David M Euhus, Kristin C Smith, Linda Robinson, Amy Stucky, Olufunmilayo I Olopade, Shelly Cummings, Judy E Garber, Anu Chittenden, Gordon B Mills, Paula Rieger, Laura Esserman, Beth Crawford, Kevin S Hughes, Connie A Roche, Patricia A Ganz, Joyce Seldon, Carol J Fabian, Jennifer Klemp, and Gail Tomlinson. Pretest prediction of brca1 or brca2 mutation by risk counselors and the computer model brcapro. *J Natl Cancer Inst*, 94(11):844–51, June 2002.

[10] J.B. Fan, A. Qliphant, R. Shen, B.G. Kermani, F. Garcia, K.L. Gunderson, M. Hansen, F. Steemers, S.L. Butler, P. Deloukas, L. Galver, S. Hunt, C. McBride, M. Bibikova, T. Rubano, J. Chen, E. Wickham, D. Doucet, W. Chang, D. Campbell, B. Zhang, S. Kruglyak, D. Bently, J. Haas, P. Rigault, L. Zhou, J. Stuelpnagel, and M.S. Chee. Highly parallel snp genotyping. *Cold Springs Harbor Symposia on Quantitative Biology*, 68:69–78, 2003.

[11] Martin Farrall and Andrew P Morris. Gearing up for genome-wide gene-association studies. *Hum Mol Genet*, 14 Spec No. 2:R157–62, October 2005.

[12] Simon Fiddy, David Cattermole, Dong Xie, Xiao Yuan Duan, and Richard Mott. Igs: An integrated system for genetic analysis. *BMC Bioinformatics*, 7:210, 2006.

[13] Nelson B Freimer and Chiara Sabatti. Human genetics: variants in common diseases. *Nature*, 445(7130):828–30, February 2007.

[14] Genelex. Genelex website. available at `http://www.genelex.com/`, May 2007.

[15] Genelex. Genemedrx: Drug-drug and drug-gene interaction software. available at: http://genemedrx.com/, May 2007.

[16] GeneSage. Genesage website. avalable at `http://www.genesage.com`, July 2006.

[17] GeneTrack. Genetrack website. available at `http://www.genetrack.bc.ca`, July 2006.

[18] Leigh S Goggin, Robert H Eikelboom, and Marcus D Atlas. Clinical decision support systems and computer-aided diagnosis in otology. *Otolaryngol Head Neck Surg*, 136(4 Suppl):S21–6, April 2007.

[19] Illumina. Illumina godengate assay workflow. available at: http://www.illumina.com/downloads/goldengateassay.pdf, May 2007.

[20] K M Kelly and K Sweet. In search of a familial cancer risk assessment tool. *Clin Genet*, 71(1):76–83, January 2007.

[21] L Kruglyak and D A Nickerson. Variation is the spice of life. *Nat Genet*, 27(3):234–6, March 2001.

[22] Cécile Libioulle, Edouard Louis, Sarah Hansoul, Cynthia Sandor, Frédéric Farnir, Denis Franchimont, Séverine Vermeire, Olivier Dewit, Martine de Vos, Anna Dixon, Bruno Demarche, Ivo Gut, Simon Heath, Mario Foglio, Liming Liang, Debby Laukens, Myriam Mni, Diana Zelenika, André Van Gossum, Paul Rutgeerts, Jacques Belaiche, Mark Lathrop, and Michel Georges. Novel crohn disease locus identified by genome-wide association maps to a gene desert on 5p13.1 and modulates expression of ptger4. *PLoS Genet*, 3(4):e58, April 2007.

[23] Yen-Ling Low, Sara Wedrén, and Jianjun Liu. High-throughput genomic technology in research and clinical management of breast cancer. evolving landscape of genetic epidemiological studies. *Breast Cancer Res*, 8(3):209, 2006.

[24] Ruth McPherson, Alexander Pertsemlidis, Nihan Kavaslar, Alexandre Stewart, Robert Roberts, David R Cox, David A Hinds, Len A Pennacchio, Anne Tybjaerg-Hansen, Aaron R Folsom, Eric Boerwinkle, Helen H Hobbs, and Jonathan C Cohen. A common allele on chromosome 9 associated with coronary heart disease. *Science*, May 2007.

[25] K S Metaxiotis and J E Samouilidis. Expert systems in medicine: academic exercise or practical tool? *J Med Eng Technol*, 24(2):68–72, 2000.

[26] Richa Saxena, Benjamin F Voight, Valeriya Lyssenko, Noel P Burtt, Paul I W de Bakker, Hong Chen, Jeffrey J Roix, Sekar Kathiresan, Joel N Hirschhorn, Mark J Daly, Thomas E Hughes, Leif Groop, David Altshuler, Peter Almgren, Jose C Florez, Joanne Meyer, Kristin Ardlie, Kristina Bengtsson, Bo Isomaa, Guillaume Lettre, Ulf Lindblad, Helen N Lyon, Olle Melander, Christopher Newton-Cheh, Peter Nilsson, Marju Orho-Melander, Lennart Råstam, Elizabeth K Speliotes, Marja-Riitta Taskinen, Tiinamaija Tuomi, Candace Guiducci, Anna Berglund, Joyce Carlson, Lauren Gianniny, Rachel Hackett, Liselott Hall, Johan Holmkvist, Esa Laurila, Marketa Sjögren, Maria Sterner, Aarti Surti, Margareta Svensson, Malin Svensson, Ryan Tewhey, Brendan Blumenstiel, Melissa Parkin, Matthew Defelice, Rachel Barry, Wendy Brodeur, Jody Camarata, Nancy Chia, Mary Fava, John Gibbons, Bob Handsaker, Claire Healy, Kieu Nguyen, Casey Gates, Carrie Sougnez, Diane Gage, Marcia Nizzari, Stacey B Gabriel, Gung-Wei Chirn, Qicheng Ma, Hemang Parikh, Delwood Richardson, Darrell Ricke, and Shaun Purcell. Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. *Science*, April 2007.

[27] NJ Schork, D Fallin, and D. Lanchbury. Single nucleotide polymorphism and the future of genetic epidemiology. *Clinical Genetics*, 58:250–264, 2000.

[28] Laura J Scott, Karen L Mohlke, Lori L Bonnycastle, Cristen J Willer, Yun Li, William L Duren, Michael R Erdos, Heather M Stringham, Peter S Chines, Anne U Jackson, Ludmila Prokunina-Olsson, Chia-Jen Ding, Amy J Swift, Narisu Narisu, Tianle Hu, Randall Pruim, Rui Xiao, Xiao-Yi Li, Karen N Conneely, Nancy L Riebow, Andrew G Sprau, Maurine Tong, Peggy P White, Kurt N Hetrick, Michael W Barnhart, Craig W Bark, Janet L Goldstein, Lee Watkins, Fang Xiang, Jouko Saramies, Thomas A Buchanan, Richard M Watanabe, Timo T Valle, Leena Kinnunen, Goncalo R Abecasis, Elizabeth W Pugh, Kimberly F Doheny, Richard N Bergman, Jaakko Tuomilehto, Francis S Collins, and Michael Boehnke. A genome-wide association study of type 2 diabetes in finns detects multiple susceptibility variants. *Science*, April 2007.

[29] Nameeta Shah, Michael V Teplitsky, Simon Minovitsky, Len A Pennacchio, Philip Hugenholtz, Bernd Hamann, and Inna L Dubchak. Snp-vista: an interactive snp visualization tool. *BMC Bioinformatics*, 6:292, 2005.

[30] Barkur S. Shastry. Snp allels in human disease and evolution. *American Journal of Human Genetics*, 47:561–566, 2002.

[31] Robert Sladek, Ghislain Rocheleau, Johan Rung, Christian Dina, Lishuang Shen, David Serre, Philippe Boutin, Daniel Vincent, Alexandre Belisle, Samy Hadjadj, Beverley Balkau, Barbara Heude, Guillanume Charpentier, Thomas J. Hudson, Alexandre Montpetit, Alexey V. Pshezhetsky, Marc Prentki, Barry I. Posner, David J. Balding, David Meyre, Constantin Polychronakos, and Philippe Froguel. A genome-wide association study identifies novel risk loci for type 2 diabetes. *Nature*, 445:881–885, February 2007.

[32] Beatriz Sobrino, María Brión, and Angel Carracedo. Snps in forensic genetics: a review on snp typing methodologies. *Forensic Sci Int*, 154(2-3):181–94, November 2005.

[33] Valgerdur Steinthorsdottir, Gudmar Thorleifsson, Inga Reynisdottir, Rafn Benediktsson, Thorbjorg Jonsdottir, G Bragi Walters, Unnur Styrkarsdottir, Solveig Gretarsdottir, Valur Emilsson, Shyamali Ghosh, Adam Baker, Steinunn Snorradottir, Hjordis Bjarnason, Maggie C Y Ng, Torben Hansen, Yu Bagger, Robert L Wilensky, Muredach P Reilly, Adebowale Adeyemo, Yuanxiu Chen, Jie Zhou, Vilmundur Gudnason, Guanjie Chen, Hanxia Huang, Kerrie Lashley, Ayo Doumatey, Wing-Yee So, Ronald C Y Ma, Gitte Andersen, Knut Borch-Johnsen, Torben Jorgensen, Jana V van Vliet-Ostaptchouk, Marten H Hofker, Cisca Wijmenga, Claus Christiansen, Daniel J Rader, Charles Rotimi, Mark Gurney, Juliana C N Chan, Oluf Pedersen, Gunnar Sigurdsson, Jeffrey R Gulcher, Unnur Thorsteinsdottir, Augustine Kong, and Kari Stefansson. A variant in cdkal1 influences insulin response and risk of type 2 diabetes. *Nat Genet*, April 2007.

[34] Wenyi Wang, Sining Chen, Kieran A Brune, Ralph H Hruban, Giovanni Parmigiani, and Alison P Klein. Pancpro: risk assessment for individuals with a family history of pancreatic cancer. *J Clin Oncol*, 25(11):1417–22, April 2007.

[35] Jinho Yoo, Bonghee Seo, and Yangseok Kim. Snpanalyzer: a web-based integrated workbench for single-nucleotide polymorphism analysis. *Nucleic Acids Res*, 33(Web Server issue):W483–8, July 2005.

[36] Lan-Juan Zhao, Miao-Xin Li, Yan-Fang Guo, Fu-Hua Xu, Jin-Long Li, and Hong-Wen Deng. Snpp: automating large-scale snp genotype data management. *Bioinformatics*, 21(2):266–8, January 2005.

Chapter 2

D-GRIP: DNA Genetic Risk Information Profile[1]

2.1 Introduction

Genetics knowledge is being transformed through whole-genome association studies enabled by new high-throughput genotyping and re-sequencing technologies. In the past, genetics research focused on the identification of individual genes directly responsible for a disease or phenotype, based on Mendellian genetics. Common diseases such as diabetes, heart disease, asthma and cancer are caused by a combination of genetic and environmental factors [17, 25]. For complex diseases, the genetic component may be provided by a few genes with moderate effects or a large number of genes with smaller effects [22].

To identify genes that contribute to susceptibility but are not definitively causal has emerged as the focus of many large genetics studies. With the completion of the Human Genome project [21], the uncovering of common genetic variants through the International Haplotype Map (HapMap) [6] has enabled the susceptibility studies for common diseases [32]. The analysis of

[1] A version of this chapter will be submitted for publication: Srivastava S and Wasserman W. 2007

large sets of common genetic variants in association to specific diseases are called genome-wide association (GWA) studies. The GWA studies typically utilize a set of single nucleotide polymorphisms (SNPs) from the HapMap project known to represent blocks of linked variations (so called 'tag' SNPs) along with nonsynonymous SNPs and SNPs situated within evolutionarily conserved regions of the genome. A large number of GWAs were recently published for diseases such as Diabetes type 2 [34, 35, 37, 39], bipolar disorder [4], Alzheimer's disease [7], Crohn's (inflammatory bowel) disease [8, 26] and coronary artery disease [29]. Given the rate of these new discoveries, there is much excitement in the scientific community for the potential to discover new links between genes and diseases - links which could pave the road for predictive genetic screening [13].

Facilitating the GWA studies are several new high-throughput genotyping platform technologies such as the Affymetrix GeneChip® and Illumina BeadArray™ which can simultaneously analyze thousands of variable positions (i.e. SNPs). The advantages of such platforms lie in their high multiplexing capability, increased reliability and the low genotyping cost per SNP. Both platforms allow genotyping of 500,000 SNPs per sample at a cost of less than $0.01 per SNP with greater than 95% accuracy [12, 27]. Due to such advancements, it is now economically feasible to perform large-scale whole-genome studies. It is even possible for an individual to obtain one's own genotype information covering many known common sequence variants for an affordable price.

Suppose a geneticist was provided with results of a large-scale genotyping experiment. It would be natural for that scientist to seek insights into the data from computational services. There are several open source systems and commercial systems currently available for application to genotype data. In the realm of open source systems, three categories exist: data management tools, visualization tools and risk assessment tools. Data management and visualization tools such as Integrated Genotype Analysis [14], SNPP [45], SNP-VISTA [36] and GEVALT [9] can process and store large amounts of raw genotype data and provide intuitive visualization of results from many samples. The majority of the risk prediction tools, such as BRCAPRO [11], PancPRO [41] and other Cancer Risk Assessment Tools (CRATs) [24], utilize family history to predict risk of disease with relatively high accuracy. However, no system allows an individual to explore genetic risk for many diseases given a single individual's genotype.

A few commercial systems handle genetic data and/or perform risk prediction. In addition to performing DNA tests, GeneTracks [19] provides software to trace family ancestry based on markers and offers secure storage of personal genetic information. Genelex [18] uses a small set of markers for predicting adverse drug reactions and Mendelian model diseases. For the commercial risk prediction services, genome-scale genotype data are not utilized and risk predictions are, in general, very specific to a small set of Mendelian diseases.

Given the advent of new genotyping technologies and the flow of new discoveries of disease-associated variants, is it now possible to use existing knowledge of diseases to create disease risk profiles for individuals? This paper is concerned with exploring such a concept in order to identify key limitations which must be addressed in genetics, bioinformatics and statistics. In addition, the research raises ethical and societal implications. We created a prototype web tool called D-GRIP, DNA Genetic Risk Information Profile, which attempts to generate an individual's genetic risk profile given a genotype as input. The prototype system accepts up to one million genotypes, compares the submitted data to a DNA-Disease database and then outputs a report for those diseases for which the individual has a predicted modified risk. In order to test the utility and ascertain the limitations of D-GRIP, a survey of potential users, such as genetic counselors, was performed and their feedback was recorded. The development and subsequent assessment of D-GRIP revealed several key weaknesses which must be addressed before wide use of a predictive system should be attempted.

2.2 Methods

In order to create a practical prototype risk assessment tool, several components are required. An intuitive and easy to use interface is essential. At the core of the software, two aspects are needed: a DNA-Disease database and a statistical model for risk assessment. Lastly, the software needs to be passed through a testing phase to assess both usability and predictive performance. For D-GRIP, a schematic overview is shown in Figure 2.1. In subsequent sections, the various components of the software are detailed and in the results, a walkthrough is performed to illustrate the features of D-GRIP.

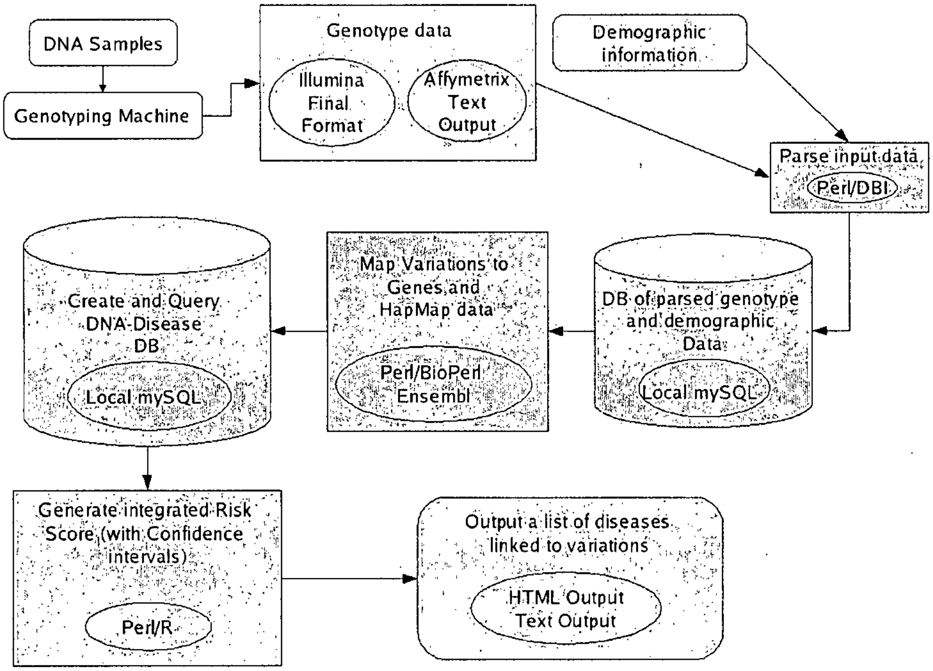

DNA Samples
Genotyping Machine
Genotype data
Illumina Final Format
Affymetrix Text Output
Demographic information
Parse input data
Perl/DBI
Create and Query DNA-Disease DB
Local mySQL
Map Variations to Genes and HapMap data
Perl/BioPerl Ensembl
DB of parsed genotype and demographic Data
Local mySQL
Generate integrated Risk Score (with Confidence intervals)
Perl/R
Output a list of diseases linked to variations
HTML Output Text Output

Figure 2.1: A schematic overview detailing the flow of information across the various components of D-GRIP is illustrated.

2.2.1 D-GRIP Overview

The overall flow of information occurs in three steps. The first step involves entering demographic information and the user's genotype data. In the second step the genotype data are compared to a genotype-phenotype database and a risk is calculated for the individual to develop each disease represented in the database. The last step is the reporting of any disease associated variations found in the user's genotype and the relevant statistical measures.

In the first step, the user enters demographic information such as age, gender and ethnic background. Due to the complexities involved in classifying ethnicity [23], a geographical generic grouping was used as follows: European, Asian, African, Pacific, Mixed and First Nations/Aboriginals. It is also possible to infer ethnicity based on ancestry informative markers (AIMs) [43], especially for admixed individuals but for simplicity, D-GRIP uses user-identified ethnicity. The user is also required to input one's genotype data, either by uploading the processed genotype file or copying and pasting the file. D-GRIP accepts two types of genotype file formats from widely used instruments (Illumina Final format and Affymetrix text output). Each row of the genotype file contains a SNP identifier (the 'rs' number provided by dbSNP [38]) and the two alleles that make up the observed genotype. The software is capable of handling up to one million genotypes at a time.

The second step processes the genotype data, based on the defined ethnic background and compares each of the user's genotypes to the entries in a genotype-phenotype database. For corresponding matches of SNP id and genotype, D-GRIP uses the specific SNP from the genotype-phenotype database in the statistical model to calculate probability of developing a specific disease. The details of the database and statistical model are explained in subsequent sections.

The final step involves reporting all matching SNPs between the user's genotype data and the disease-associated SNPs. The analysis results are reported in a tabular format which includes for each disease, the particular gene, the particular SNP (and genotype) associated with the disease, the population in which the association was observed and links to relevant studies supporting the association between the disease and genotype. In addition, for each SNP, odds ratio and confidence intervals, risk and major allele's homozygous genotypes, the case and control genotype frequencies and set of SNPs found to be in high linkage disequilibrium based on the HapMap data are reported. Finally, an overall probability of developing a disease is shown, based on the statistical model used. As per the model, the overall probability is calculated over the whole set of observed disease-associated genotypes.

D-GRIP was implemented for browser-based access over a network. Since there are many social, ethical and legal implications associated with the use of such a risk assessment tool, access to D-GRIP is restricted. D-GRIP is envisioned to be used in a guided setting, for example, in the presence of a genetic counselor. In addition, to respect privacy and confidentiality, user submitted information (e.g. demographic and genotype data) is not stored in the system. Once a report is generated, all user data are removed from memory.

2.2.2 Genotype-Phenotype Database

Existing genotype-phenotype databases are not sufficient for large-scale disease risk prediction due to deficiencies in the organization and/or extent of genetic risk knowledge [33]. Currently, the majority of genetic disease databases use free text for disease information (rather than a more structured format) and thus are not suited for large scale computational analyses. Due to this deficiency, we created a D-GRIP DNA-Disease database for the testing of the system. The DNA-Disease database contains information pertaining to a limited set of complex diseases. The information represented primarily includes validated markers (SNPs) either confirmed in multiple studies or emerging from studies performed with samples from large numbers of participants.

For each available disease, the DNA-Disease database contains associated and validated SNPs. For each SNP, the case and control allele and genotypic frequencies from different populations is recorded. We decided to model the information in the database on an existing system, AlzGene, which was developed for genetic markers predictive for Alzheimer's disease risk [5]. AlzGene was created to house the results of a meta-analysis for each polymorphism with known genotype data in at least three case-control studies. For each polymorphism, allele and genotypic frequencies on a per population basis are provided in a well organized structure. In addition to Alzheimer's data, the D-GRIP DNA-Disease database contains information from a Parkinson's disease database (PDGene), created by the developers of AlzGene [1]. The data for Diabetes Type II was manually extracted from a recent large scale genome wide association study [37]. A summary of the contents of the database is represented in Table 2.1.

Diseases	Number of Genes	Number of Polymorphisms
Alzheimer's Disease	38	76
Parkinson Disease	8	17
Diabetes Type II	5	8

Table 2.1: A summary of the number of genes and number of polymorphisms for each of the diseases in the DNA-Disease database.

2.2.3 Disease Risk Model

The implemented statistical model in D-GRIP was defined by Yang et al. [44]. The original model includes two steps. First, a likelihood ratio was calculated using logistic regression and then a posterior probability of disease was estimated using the likelihood ratio. The likelihood ratio is defined as the ratio of the probability for an individual with a disease to have an observed genotype to the probability for an individual without the disease to have the genotype [44]. While full details can be obtained in the cited paper, a brief summary follows.

Likelihood Ratio

For an individual with a set of genetic tests, G, where G is a vector of n disease susceptibility genes or alleles $(g_1,g_2,\ldots,g_n)$. Let $g_i = 1$ for positive genetic test result and $g_i = 0$ for negative test result, then, let the individual who is tested for one allele be represented as a combination of 0s and 1s.

Also, let D represent the diseased (case) population and let $\bar{D}$ represent the non-diseased (control) population. Then, the likelihood ratio for any observed value of G can be defined as:

$$LR(G) = \frac{P(G|D)}{P(G|\bar{D})}. \tag{2.1}$$

As stated, G is a set of genetic tests $G = (g_1, g_2, \ldots, g_n)$. The implemented model assumes that each genetic test is acting independently, thus the joint probability of a given result is the product of the individual probabilities, $P(G|D) = P(g_1|D)P(g_2|D)\ldots P(g_n|D)$. This is also true for $P(G|\bar{D})$ and thus it follows that $LR(G) = LR(g_1)LR(g_2)\ldots LR(g_n)$. Thus, the likelihood ratio for a panel of independent tests is simply the product of the likelihood ratios of the individual test results. The assumption of independence will be discussed in a later section.

Since the DNA-Disease database contains case-control studies from various populations for each disease, a logistic model can be used to estimate the likelihood ratio. For a binary disease outcome ($D = 0, 1$), for a logistic model in the population, logistic regression can be used to calculate the likelihood ratio from case-control studies in a population, as follows:

$$\ln LR(G) = \ln\left(\frac{N_{CO}}{N_{CA}}\right) + \alpha_{CC} + \beta G^T, \tag{2.2}$$

where α_{CC} and β are the intercept term and the logistic regression coefficient of the odds of developing the disease respectively. N_{CA} is the number of case subjects in the study sample and N_{CO} is the number of control subjects in the study sample. It is worth nothing that the likelihood-ratio calculation assumes each gene is acting independently. However, realistically, gene-gene interactions and gene-environment interactions should be included in the model. The likelihood-ratio equation (Equation 2.2) can be modified by including a vector of covariates as well as interaction effects of multiple binary tests (gene-gene or gene-environment interactions). However, for brevity, the equation is not shown here and for prototype development, is not used in D-GRIP.

Posterior Probability

The statistical model uses a set of genetic tests to predict the probability that the multifactorial disease will develop in people with allele-positive result, or $P(D|G)$. By using the a pretest risk of disease, $P(D)$, or the average risk of disease in the population, the posterior probability can be defined as:

$$P(D|G) = \frac{LR(G)P(D)}{[1 - P(D)] + LR(G)P(D)}. \tag{2.3}$$

2.2.4 Haplotype Data

In addition to utilizing validated disease associated variations, we incorporated the use of haplotype blocks in the statistical model. For each of the SNPs that are associated with disease in our DNA-Disease database, we extracted SNPs (1Mb on either side), in corresponding HapMap populations that were in high linkage disequilibrium (threshold of $r^2 > 0.8$ [2]). To extract the HapMap SNPs and linkage disequilibrium values, Ensembl (build 45) was used. Due to the complexity involved in defining and classifying populations, a simplification was made when incorporating the hapmap data: the populations from the HapMap project were generalized to match the populations found in the DNA-Disease database. The population categories from the DNA-Disease database were Caucasians, Asians, African and other/Mixed. The corresponding matches from the HapMap project were European ancestry (CEPH) grouped as Caucasians, the Tokyo (JPT) and Han Chinese (CHB) ethnic groups represented as Asians and the Nigeria (YRI) ethnic group matched to Africans.

D-GRIP uses the HapMap data in two different ways during the generation of a disease risk profile. First, for the reported disease-associated SNP, an integrated analysis is performed in which multiple disease associated SNPs in high linkage disequilibrium (LD) are clustered together during the probability calculation. Rather than treating these high LD SNPs independently in the calculated overall disease probability, a simplification is made. The SNP with the highest effect (highest odds ratio) is used to represent the other SNPs in high LD and thus only one SNP (with strongest effect) is used in the posterior probability calculation.

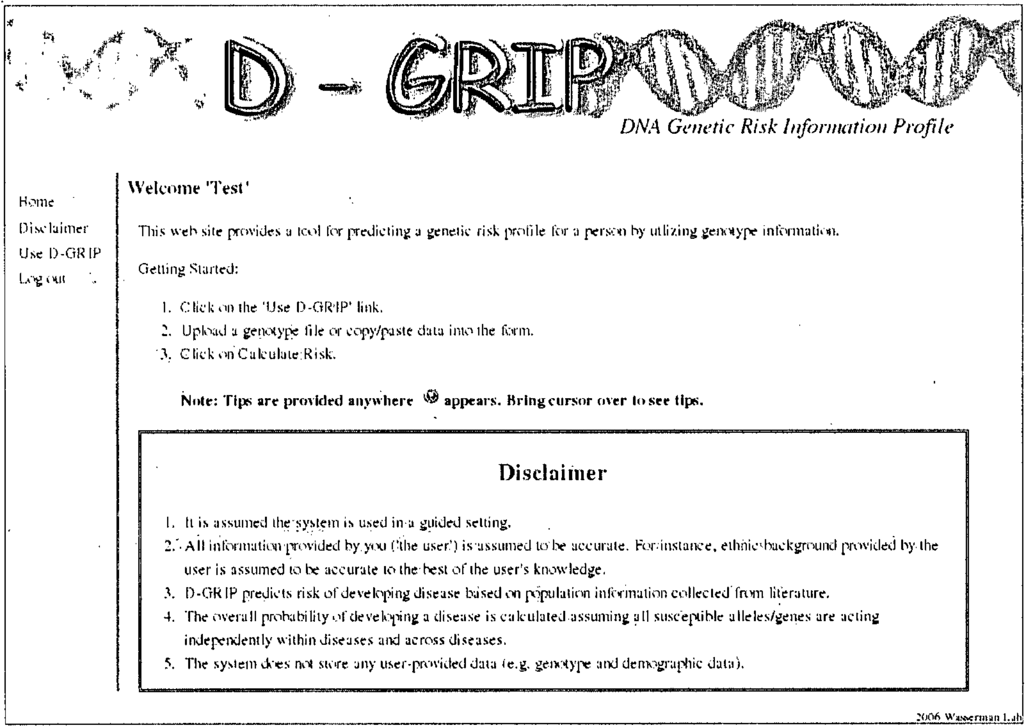
Second, an inferred analysis is reported with the observed genotypes in the final risk profile output. The inferred analysis reports SNPs that were present in the user's genotype but did not have a direct association to a disease. These inferred SNPs are in high LD with known disease associated SNPs which are present in the DNA-Disease database. The Hapmap Genome Browser (Release 21) [40] was used to extract the phased genotype data. Subsequently, Haploview (version 3.32) [3] software was used to calculate the haplotype blocks, using the default method on Haploview software for haplotype block calculation, in order to infer phase information. Since the inferred analysis is highly predictive in nature and untested, it is provided as an option for the user, which by default is turned off during analysis. Also, the inferred SNPs are not used in overall posterior probability calculation.

2.2.5 Software Evaluation

After a working prototype was created, D-GRIP underwent a series of critical evaluations. The evaluation was structured as a survey where D-GRIP was demonstrated to experts and their feedback was recorded. A total of 21 scientists, clinicians or counselors were surveyed including clinical geneticists, molecular geneticists, biostatisticians and genetic counseling students.

2.3 Results

A walkthrough of D-GRIP illustrates the user interface features as well as the underlying DNA-Disease database. Figure 2.2 shows the first page of D-GRIP after the user logs in. The opening page explains how to use D-GRIP and warns the user via a disclaimer which outlines the assumptions and limitations of D-GRIP. Upon clicking the 'Use D-GRIP' link, the user is presented with a form to solicit demographic information and options regarding the analysis. In this example, suppose the user is a male, 47 years old from European ancestry and inference analysis turned on (Figure 2.3).



Welcome 'Test'

This web site provides a tool for predicting a genetic risk profile for a person by utlizing genotype information.

Getting Started:

1. Click on the 'Use D-GRIP' link.
2. Upload a genotype file or copy/paste data into the form.
3. Click on Calculate Risk.

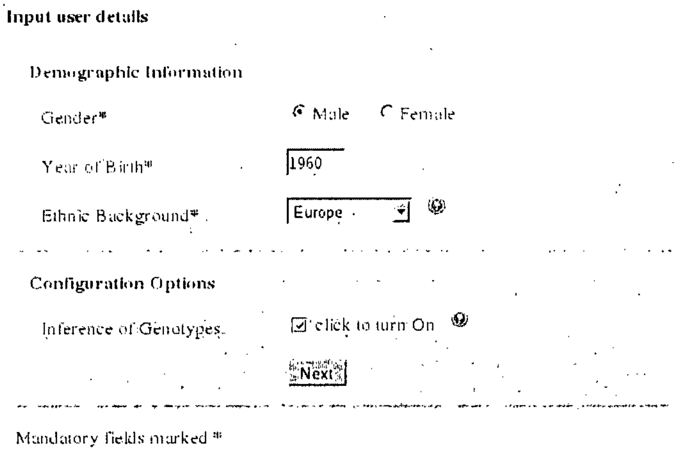
Note: Tips are provided anywhere ☺ appears. Bring cursor over to see tips.

Disclaimer

1. It is assumed the system is used in a guided setting.
2. All information provided by you ('the user') is assumed to be accurate. For instance, ethnic background provided by the user is assumed to be accurate to the best of the user's knowledge.
3. D-GRIP predicts risk of developing disease based on population information collected from literature.
4. The overall probability of developing a disease is calculated assuming all susceptible alleles/genes are acting independently within diseases and across diseases.
5. The system does not store any user-provided data (e.g. genotype and demographic data).

2006 Wasserman Lab

Figure 2.2: The opening page of D-GRIP is shown. The instructions on how to use D-GRIP and a disclaimer explicitly stating the assumptions and limitations inherent in D-GRIP are shown.



Input user details

Demographic Information

Gender*	● Male ○ Female
Year of Birth*	1960
Ethnic Background*	Europe

Configuration Options

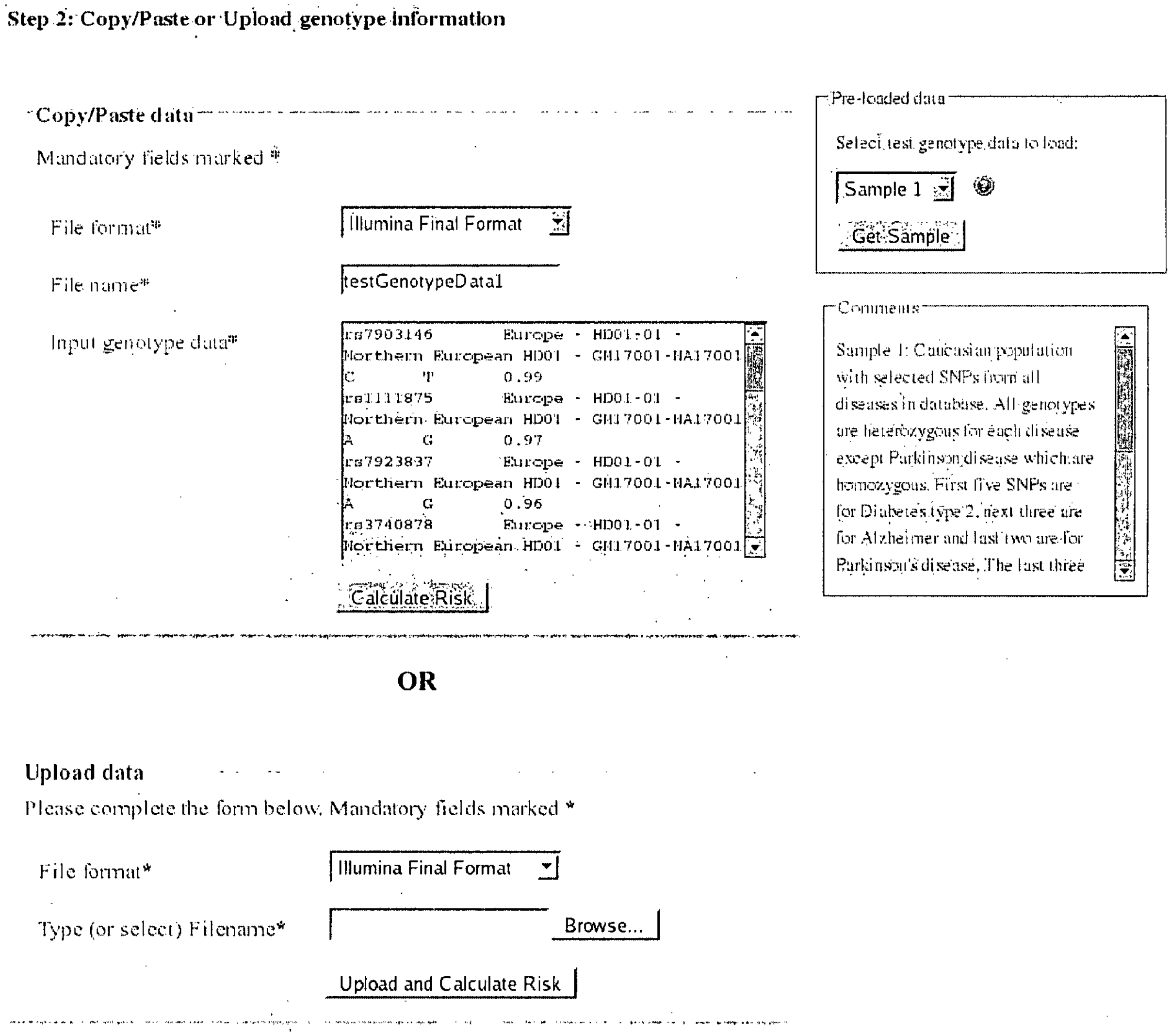
Inference of Genotypes	☑ click to turn On

Mandatory fields marked *

Figure 2.3: The first step in using D-GRIP is illustrated, where the user's demographic information such as gender, age and ethnicity is collected. The hypothetical example above shows a male, 47 years old from European ancestry. The inference option is turned on (checked).

After clicking 'next', the user is asked to submit genotype data. Two options are available, copying and pasting the data into the submission box or uploading a genotype data file. The file formats supported are Illumina's Final format or Affymetrix's Text Output format. In order to illustrate the underlying DNA-Disease database, sample genotype files were created. One such sample genotype file is shown in Figure 2.4 which shows 13 genotypes, all of which are heterozygous for particular SNPs from each of the three diseases. After loading the genotype data, the user clicks on 'Calculate Risk'.

The last step is the output page which displays a disease risk profile report. As seen in Figure 2.5, for each disease, the associated gene, SNP, genotype and population is reported along with a list of scientific articles supporting the association between disease and genotype. In addition, for



Step 2: Copy/Paste or Upload genotype information

Copy/Paste data

Mandatory fields marked *

File format*: Illumina Final Format

File name*: testGenotypeData1

Input genotype data*:

```
rs7903146        Europe - HD01-01 -
Northern European HD01 - GM17001-HA17001
C       T        0.99
rs1111875        Europe - HD01-01 -
Northern European HD01 - GM17001-HA17001
A       G        0.97
rs7923837        Europe - HD01-01 -
Northern European HD01 - GM17001-HA17001
A       G        0.96
rs3740878        Europe - HD01-01 -
Northern European HD01 - GM17001-HA17001
```

Pre-loaded data

Select test genotype data to load: Sample 1

Comments

Sample 1: Caucasian population with selected SNPs from all diseases in database. All genotypes are heterozygous for each disease except Parkinson disease which are homozygous. First five SNPs are for Diabetes type 2, next three are for Alzheimer and last two are for Parkinson's disease. The last three

OR

Upload data

Please complete the form below. Mandatory fields marked *

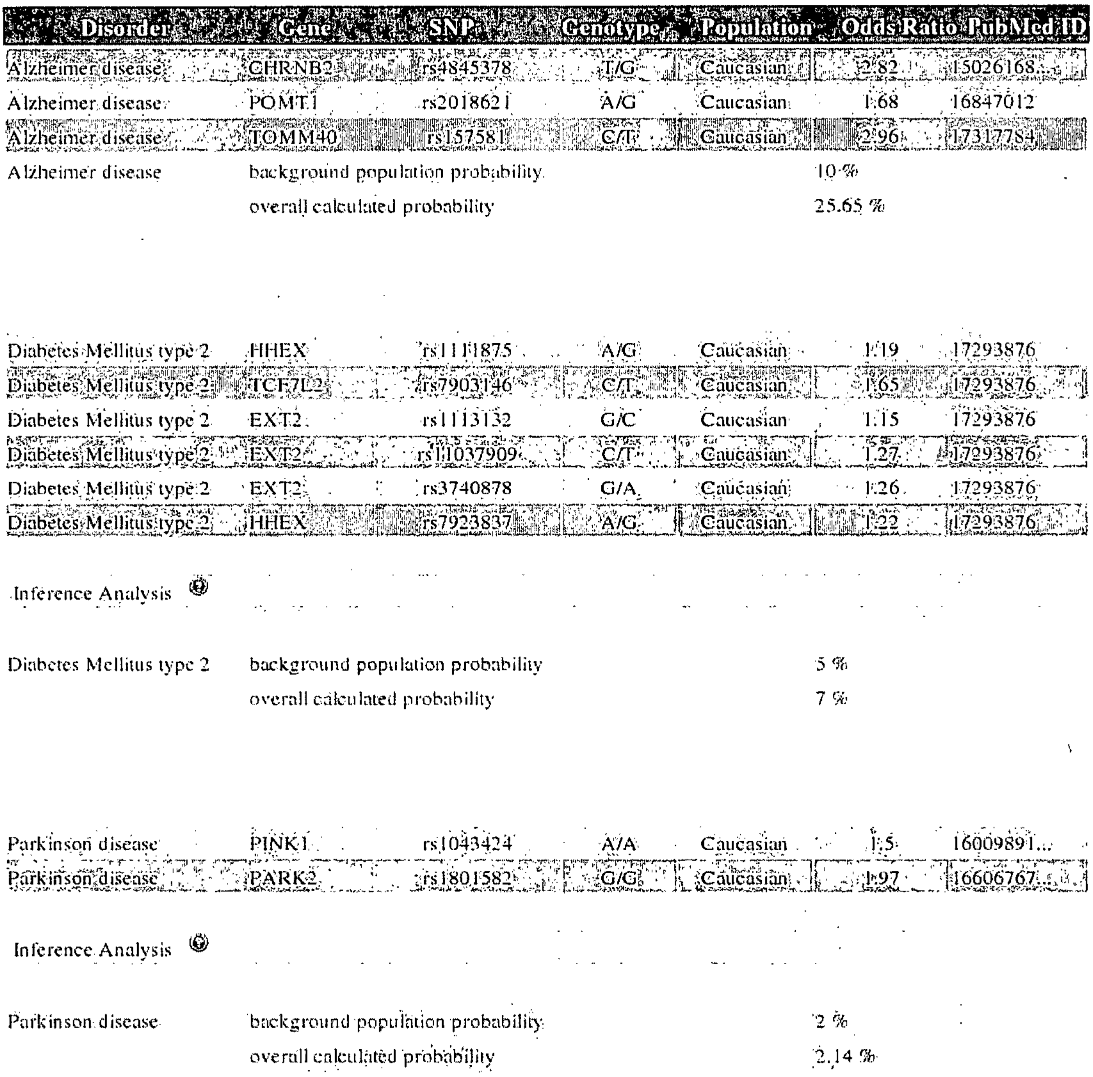
File format*: Illumina Final Format

Type (or select) Filename*:

Figure 2.4: The second step in using D-GRIP is illustrated here. The user has a choice of copying and pasting the genotype data or uploading it. For ease of use, various hypothetical sample genotype files were created to illustrate D-GRIP. The above example contains the 13 highly significant genotypes which are heterozygous for each disease in the DNA-Disease database. A description of the pre-loaded data is shown in the 'Comments' box.

each disease, the background probability and the calculated probability are indicated. For example, based solely on genotype, a 47 year old male from European ancestry is reported to have a slightly elevated risk (7.0%) of developing Diabetes type II given that the background probability in the Caucasian population is 5%. In figure 2.6, further details for each SNP are shown such as risk and major genotype, genotype frequencies for case and control populations, odds and likelihood ratios and confidence intervals. In this example, SNP rs7903146 from gene TCF7L2 for Diabetes Mellitus type 2 is shown. Links to relevant resources such as GenBank, OMIM and db-SNP are available when clicking the gene, disease name and SNP identifier respectively. Also, after clicking on the overall probability row, an integrated analysis is shown which combines disease associated SNPs that are in linkage disequillibrium according to HapMap data (figure 2.7). Lastly, inferred SNPs are shown separately under each disease. The overall probability calculation is done only once, using observed SNPs and the inferred SNPs are not included in the calculation due to their speculative nature.

The evaluation of D-GRIP was performed by surveying experts in the field. The feedback was recorded and is presented in Appendix A.



Disorder	Gene	SNP	Genotype	Population	Odds Ratio	PubMed ID
Alzheimer disease	CHRNB2	rs4845378	T/G	Caucasian	2.82	15026168...
Alzheimer disease	POMT1	rs2018621	A/G	Caucasian	1.68	16847012
Alzheimer disease	TOMM40	rs157581	C/T	Caucasian	2.96	17317784

Alzheimer disease	background population probability	10 %
	overall calculated probability	25.65 %

Diabetes Mellitus type 2	HHEX	rs1111875	A/G	Caucasian	1.19	17293876
Diabetes Mellitus type 2	TCF7L2	rs7903146	C/T	Caucasian	1.65	17293876
Diabetes Mellitus type 2	EXT2	rs1113132	G/C	Caucasian	1.15	17293876
Diabetes Mellitus type 2	EXT2	rs11037909	C/T	Caucasian	1.27	17293876
Diabetes Mellitus type 2	EXT2	rs3740878	G/A	Caucasian	1.26	17293876
Diabetes Mellitus type 2	HHEX	rs7923837	A/G	Caucasian	1.22	17293876

Inference Analysis

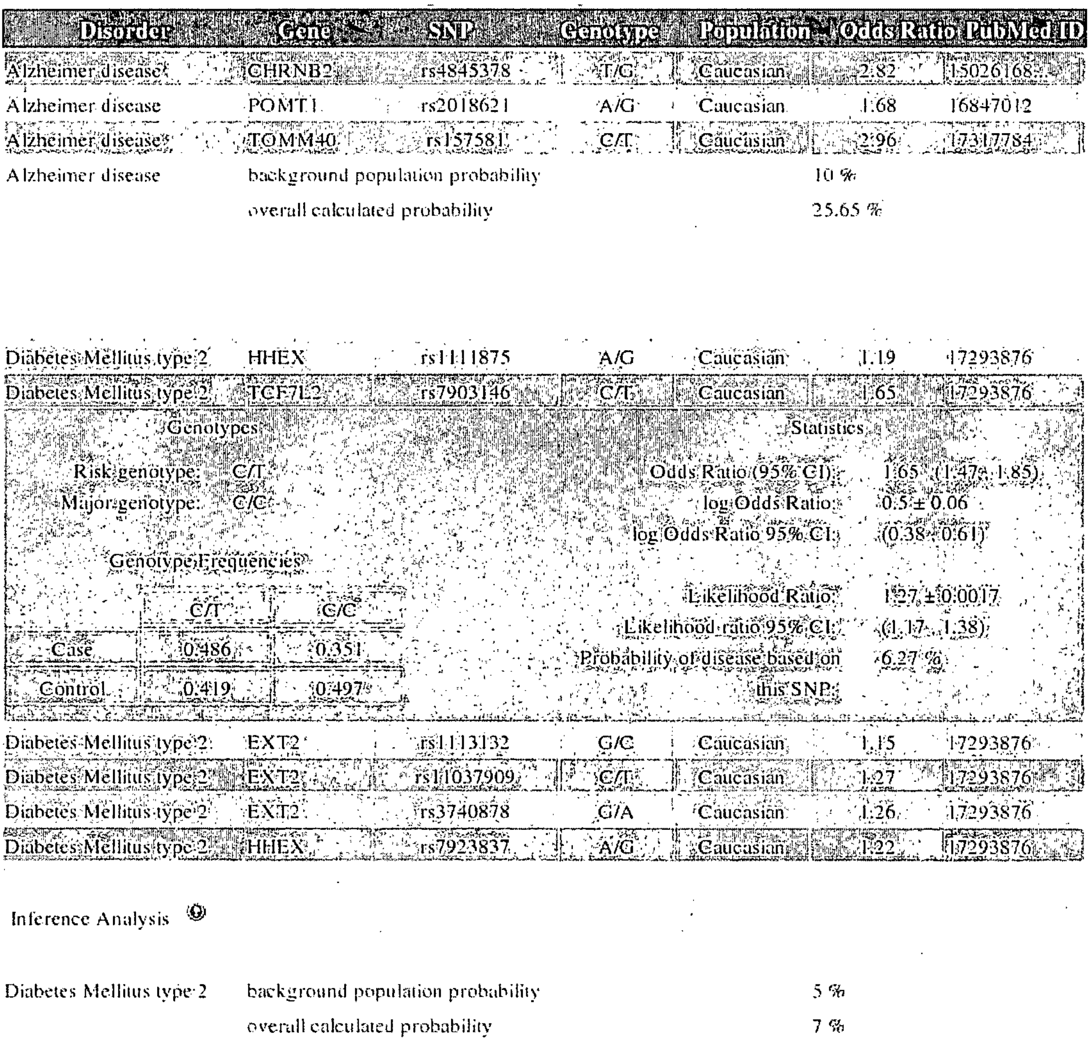
Diabetes Mellitus type 2	background population probability	5 %
	overall calculated probability	7 %

Parkinson disease	PINK1	rs1043424	A/A	Caucasian	1.5	16009891...
Parkinson disease	PARK2	rs1801582	G/G	Caucasian	1.97	16606767...

Inference Analysis

Parkinson disease	background population probability	2 %
	overall calculated probability	2.14 %

Figure 2.5: The last step shows a tabular result for any single nucleotide polymorphisms (SNPs) found to be associated with a disease in the user's genotype data.



Disorder	Gene	SNP	Genotype	Population	Odds Ratio	PubMed ID
Alzheimer disease	CHRNB2	rs4845378	T/G	Caucasian	2.82	15026168
Alzheimer disease	POMT1	rs2018621	A/G	Caucasian	1.68	16847012
Alzheimer disease	TOMM40	rs157581	C/T	Caucasian	2.96	17317784

Alzheimer disease	background population probability	10 %
	overall calculated probability	25.65 %

Disorder	Gene	SNP	Genotype	Population	Odds Ratio	PubMed ID
Diabetes Mellitus type 2	HHEX	rs1111875	A/G	Caucasian	1.19	17293876
Diabetes Mellitus type 2	TCF7L2	rs7903146	C/T	Caucasian	1.65	17293876

Genotypes

Risk genotype: C/T

Major genotype: C/C

Genotype Frequencies

	C/T	C/C
Case	0.486	0.351
Control	0.419	0.497

Statistics

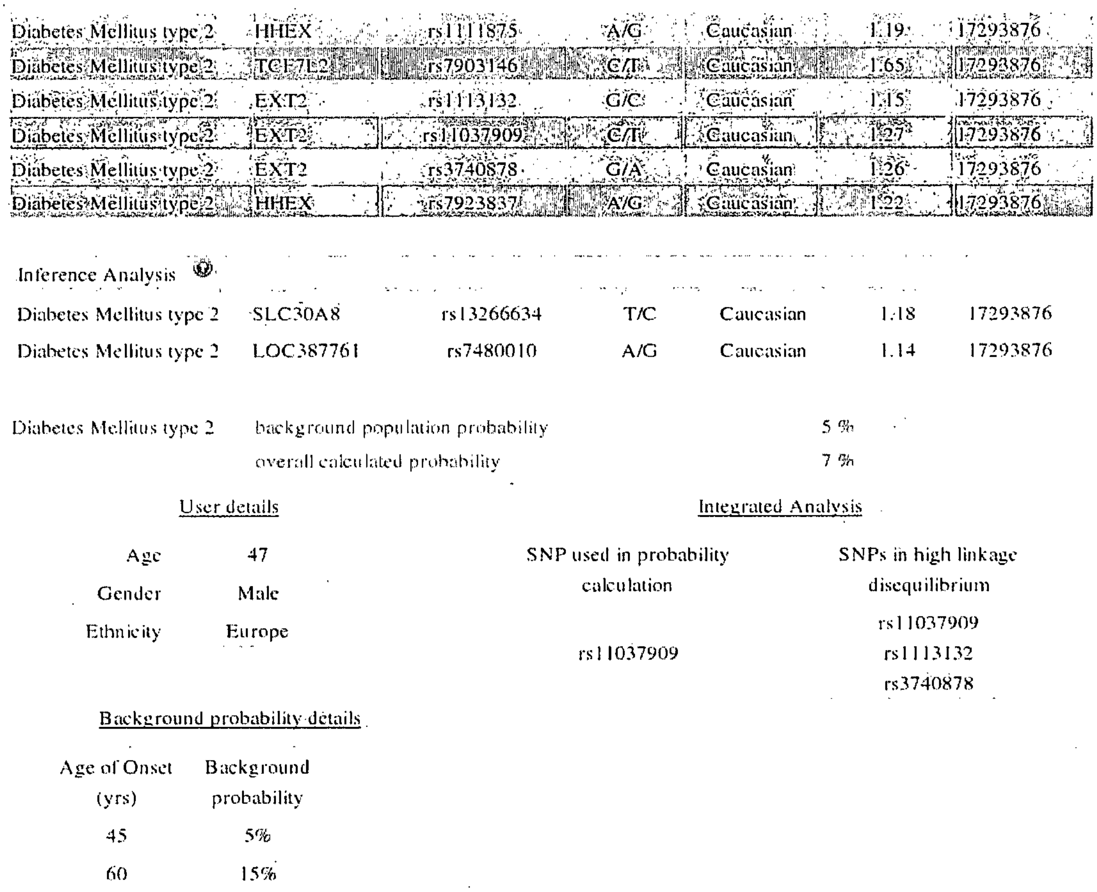
Odds Ratio (95% CI):	1.65 (1.47, 1.85)
log Odds Ratio:	0.5 ± 0.06
log Odds Ratio 95% CI:	(0.38, 0.61)
Likelihood Ratio:	1.27 ± 0.0017
Likelihood ratio 95% CI:	(1.17, 1.38)
Probability of disease based on this SNP:	6.27 %

Disorder	Gene	SNP	Genotype	Population	Odds Ratio	PubMed ID
Diabetes Mellitus type 2	EXT2	rs1113132	G/C	Caucasian	1.15	17293876
Diabetes Mellitus type 2	EXT2	rs11037909	C/T	Caucasian	1.27	17293876
Diabetes Mellitus type 2	EXT2	rs3740878	G/A	Caucasian	1.26	17293876
Diabetes Mellitus type 2	HHEX	rs7923837	A/G	Caucasian	1.22	17293876

Inference Analysis

Diabetes Mellitus type 2	background population probability	5 %
	overall calculated probability	7 %

Figure 2.6: More details are shown for each SNP. As an example, details for SNP rs7903146 is shown from gene TCF7L2 from Diabetes Mellitus type II.



Diabetes Mellitus type 2	HHEX	rs1111875	A/G	Caucasian	1.19	17293876
Diabetes Mellitus type 2	TCF7L2	rs7903146	C/T	Caucasian	1.65	17293876
Diabetes Mellitus type 2	EXT2	rs1113132	G/C	Caucasian	1.15	17293876
Diabetes Mellitus type 2	EXT2	rs11037909	C/T	Caucasian	1.27	17293876
Diabetes Mellitus type 2	EXT2	rs3740878	G/A	Caucasian	1.26	17293876
Diabetes Mellitus type 2	HHEX	rs7923837	A/G	Caucasian	1.22	17293876

Inference Analysis

Diabetes Mellitus type 2	SLC30A8	rs13266634	T/C	Caucasian	1.18	17293876
Diabetes Mellitus type 2	LOC387761	rs7480010	A/G	Caucasian	1.14	17293876

Diabetes Mellitus type 2	background population probability	5 %
	overall calculated probability	7 %

User details

Age	47
Gender	Male
Ethnicity	Europe

Integrated Analysis

SNP used in probability calculation	SNPs in high linkage disequilibrium
rs11037909	rs11037909 rs1113132 rs3740878

Background probability details

Age of Onset (yrs)	Background probability
45	5%
60	15%

Figure 2.7: Details of overall probability calculation, integrated analysis and inferred SNPs are shown for Diabetes Mellitus type 2 disease. The integrated analysis indicates which disease-associated SNPs are in high linkage disequilibrium ($r^2 > 0.8$). For SNPs in high LD, only the SNP with strongest effect (highest odds ratio) is used in the overall calculated probability.

2.4 Discussion

2.4.1 Limitations

In the current state of information and implementation, D-GRIP has several limitations. A key limitation is the narrow scope of the DNA-Disease database. The scarcity reflects two key causes: lack of organization of genotype-phenotype data and the small number of confirmed markers for risk. Even though numerous studies report new DNA marker-disease associations, there is a shortage of databases that organize such information in a comprehensive and computationally accessible manner. Databases such as AlzGene [5] and PDGene [1] are rare examples of organized genotype-phenotype data which are continuously updated when new studies are published and are easy to use computationally. More such genetics databases are required for other common diseases [33]. It should be noted that numerous databases provide information about genetics and disease, such as OMIM [15] and HGVbase [16], but the information is not sufficiently granular and/or formatted to incorporate into the risk calculation procedure of D-GRIP. The second problem, the scarcity of confirmed predictive markers will soon be ameliorated as the rate of publication of such studies is accelerating.

Another limitation of D-GRIP resides in the statistical model. There are several issues regarding the statistical model. When a posterior probability is calculated using the observed SNPs which are associated to a disease, each genetic test (SNP) is assumed to be acting independently. This is a very simplistic view and does not realistically capture the underlying disease process. In order to partially circumvent this limitation, we included haplotype information in the analysis. By including an integrated analysis where if observed SNPs in the output were found in linkage disequilibrium, only the SNP with strongest effect is included in the posterior probability calculation. Again, this is a simplification which is warranted since we could not find other existant suitable statistical models that incorporate haplotype data for risk prediction of disease with SNPs. Furthermore, the lack of consideration for gene-gene interactions and gene-environment interactions is another limitation. Even though the model allows for incorporation of interaction effects, for simplicity, D-GRIP does not utilize that feature.

A second issue with the statistical model is lack of incorporation of age and gender during risk calculation. Even though we require the user to input such demographic information when calculating risk for a particular disease, this information is not utilized. In order to use demographic information appropriately, we require the age and gender distribution for each of the individuals in the case-control studies stored in the DNA-Disease database. Since such raw data are unavailable, a simplification was used. D-GRIP uses a different prior probability (background probability) for specific diseases (e.g. Alzheimer's disease) based on the age of the person. In order to alleviate this scarcity of raw data, currently efforts are under way at the NIH to archive and distribute more detailed information on upcoming genetic association studies. The database, dbGaP is designed to house genetics studies dealing with genotype-phenotype interactions and provide all study documentation as well as pre-computed analysis [30].

Currently, no family history or medical history is used for predicting risk of disease. Incorporation of family history has been shown to improve the predictive accuracy of risk models [11]. Thus future versions of D-GRIP should incorporate family history in the risk model. For any prediction based software, rigorous validation regarding specificity, sensitivity and accuracy is required. Currently, no such validation is performed due to insufficient number of diseases in DNA-Disease database as well as the unavailability of raw genotype data from individuals for testing. D-GRIP was evaluated through a survey in which D-GRIP was demonstrated to various experts in genetics-related field and feedback was recorded. The conclusions from this form of evaluation are discussed in the next section.

2.4.2 Ideal Software

Based on the conclusions drawn from the prototype system and feedback from experts, the features and functionality of an idealized software system can be outlined. The input features of a system should include, as in the prototype, demographic information collected from the user and in addition, an option for collecting family history of any diseases and relevant environmental exposures (e.g. cigarette smoking). Also, the genotype parser should be flexible and accommodate various file formats. Preferably, an widely accepted file format standard should be established for genotyping data which are released from platforms such as Illumina and Affymetrix. By having a standard file format, exchange of genotyping data across studies will be more efficient. Lastly, user information on non-SNP variants, such as insertions/deletions, copy number variations and large-scale structural variants should also be accepted.

At the core of the software, the ideal DNA-Disease database will contain information for as many common diseases as possible. There are two ways to populate such a database. One, create a meta-analysis engine for each disease. When new studies are published for a disease, they can be added to the database and then meta-analysis re-performed over all the studies for a specific disease. This would require continuous updating of the database each time new disease associated markers are found. In the second approach, genotype-phenotype data would be extracted from disease-specific databases such as AlzGene and PD Gene, but currently, such disease specific genetics databases, of suitable format are rare.

Based on recommendations from biostatisticians, an idealized software's statistical approach would include a unique model for each disease (or a range of optional models). Since common diseases are varied and complex, it is crucial to have rigorously tested and validated statistical models. In addition, the statistical models will need to incorporate gene-gene interactions as well as co-variates such as exposure to environmental or behavioral factors.

The user interface, both the input and output of an ideal system will have to be tailored towards the audience. For example, the current disease risk profile report generated from D-GRIP is intended to be read by a trained user such as a genetic counselor. If one were to target use to family physicians, as suggested by one survey participant, it might be more suitable for the output to highlight links to information about prevention. Appropriate training will be required for any user of such a system, be it genetic counselors, family physicians or individual subjects. Lastly, it was highly recommended by the respondants that access to D-GRIP-like tools be restricted - the mixture of complicated interpretation of risk and opportunity for the generation of undue stress on the recipient of information combine to warrant limited user access for the near-term. As a last comment, the average consensus from the feedback for when such an ideal system could be accepted and used clinically was between 5 and 10 years.

2.4.3 Implications

There are many societal, ethical and legal implications involved with using D-GRIP. The immediate issues are discussed here and potential directions are presented. One of the pressing questions deals with data protection. The same level of protection should be provided for genetic data as for sensitive medical data, that is, confidentiality and privacy. In addition, the individual's rights should be respected everytime such a tool is used in professional setting. Currently, D-GRIP ensures protection of the user's rights by not storing any user specified information (demographic and genotype) and ensures confidentiality via anonymous submission of genetic data. However, in the long-term it would be more appropriate for a continuous analysis engine to reassess the DNA each time a new genetic risk marker was deposited into the database. Therefore, encryption and privacy features are required in such a tool.

There is much research needed in how to present and explain genetic risk information to individuals [10]. The effect of inappropriately explaining risks can lead to demoralization and unnecessarily increased anxiety, both of which can decrease an individual's ability to change risk-related behavior [28, 42]. Also, most people find probabilities and relative risk information difficult to comprehend, in part due to poor presentation of statistics [20]. Thus, it is recommended to use standard vocabulary, use a common denominator when explaining odds, provide both positive and negative perspectives and use visual aids for probabilities [31].

Genetic testing for affected or at risk individuals creates serious ethical dilemmas. Concerns such as discrimination from employers and insurers and fear of discrimination can deter individuals who could benefit from genetic testing. It also remains to be seen how third-party use of genetic information and potential will impact the use of predictive tools such as D-GRIP. These issues will have to be discussed and addressed by governments, industries and the public in a transparent manner [22].

2.4.4 Conclusions

The creation of the D-GRIP system for genetic risk prediction was intended to identify bioinformatics, statistical and scientific challenges that must be addressed to create predictive systems of clinical utility. The major bioinformatic limitation is the lack of available data in terms of strongly predictive susceptibility alleles for complex diseases. This is in part due to the lack of organized and computationally exploitable disease databases for complex disorders. The major statistical limitation is the calculation of risk given large-scale genotype data (e.g. incorporating haplotype information into the analysis). The major scientific limitation, despite the flurry of association studies, is our limited understanding of complex diseases and how various genes interact with each other and the environment. Any proposed predictive model (be it for a single disease or a general model) will have to undergo rigorous testing and evaluations in order to ensure clinical utility.

When the proposed limitations are overcome, useful and beneficial predictive software can be created and implemented. The key features include: incorporation of genotype data along with family history of disease, a continuously updated DNA-Disease database with a meta-analysis engine, disease-specific risk models which have been validated and user-oriented risk profile reporting. The use of the software will be under a guided setting, with potential users being genetic counselors and family physicians. Regardless of the user, appropriate training in using the software and interpreting the output will be a necessity. Lastly, implications such as privacy and confidentiality of genetic data, appropriate explanations of risk, discrimination towards individuals via third parties, effect on public health policies and public education are all important challenges to be addressed before implementation of such a predictive tool becomes a reality.

Bibliography

[1] S Bagade, NC Allen, R Tanzi, and L Bertram. The pdgene database. alzheimer research forum. available at: `http://www.pdgene.org/`, Accessed May 2007.

[2] Michael R Barnes. Navigating the hapmap. *Brief Bioinform*, 7(3):211–24, September 2006.

[3] J C Barrett, B Fry, J Maller, and M J Daly. Haploview: analysis and visualization of ld and haplotype maps. *Bioinformatics*, 21(2):263–5, January 2005.

[4] A E Baum, N Akula, M Cabanero, I Cardona, W Corona, B Klemens, T G Schulze, S Cichon, M Rietschel, M M Nöthen, A Georgi, J Schumacher, M Schwarz, R Abou Jamra, S Höfels, P Propping, J Satagopan, S D Detera-Wadleigh, J Hardy, and F J McMahon. A genome-wide association study implicates diacylglycerol kinase eta (dgkh) and several other genes in the etiology of bipolar disorder. *Mol Psychiatry*, May 2007.

[5] Lars Bertram, Matthew B McQueen, Kristina Mullin, Deborah Blacker, and Rudolph E Tanzi. Systematic meta-analysis of alzheimer disease genetic association studies: The alzgene database. *Nature Genetics*, 39:17–23, January 2007.

[6] International Hapmap Consortium. The international hapmap project. *Nature*, 426(6968):789–96, December 2003.

[7] Keith D Coon, Amanda J Myers, David W Craig, Jennifer A Webster, John V Pearson, Diane Hu Lince, Victoria L Zismann, Thomas G Beach, Doris Leung, Leslie Bryden, Rebecca F Halperin, Lauren Marlowe, Mona Kaleem, Douglas G Walker, Rivka Ravid, Christopher B Heward, Joseph Rogers, Andreas Papassotiropoulos, Eric M Reiman, John Hardy, and Dietrich A Stephan. A high-density whole-genome association study reveals that apoe is the major susceptibility gene for

sporadic late-onset alzheimer's disease. *J Clin Psychiatry*, 68(4):613–8, April 2007.

[8] J R Fraser Cummings, Rachel Cooney, Saad Pathan, Carl A Anderson, Jeffrey C Barrett, John Beckly, Alessandra Geremia, Laura Hancock, Changcun Guo, Tariq Ahmad, Lon R Cardon, and Derek P Jewell. Confirmation of the role of atg16l1 as a crohn's disease susceptibility gene. *Inflamm Bowel Dis*, April 2007.

[9] Ofir Davidovich, Gad Kimmel, and Ron Shamir. Gevalt: an integrated software tool for genotype analysis. *BMC Bioinformatics*, 8:36, 2007.

[10] Adrian Edwards, Silvana Unigwe, Glyn Elwyn, and Kerenza Hood. Effects of communicating individual risks in screening programmes: Cochrane systematic review. *BMJ*, 327(7417):703–9, September 2003.

[11] David M Euhus, Kristin C Smith, Linda Robinson, Amy Stucky, Olufunmilayo I Olopade, Shelly Cummings, Judy E Garber, Anu Chittenden, Gordon B Mills, Paula Rieger, Laura Esserman, Beth Crawford, Kevin S Hughes, Connie A Roche, Patricia A Ganz, Joyce Seldon, Carol J Fabian, Jennifer Klemp, and Gail Tomlinson. Pretest prediction of brca1 or brca2 mutation by risk counselors and the computer model brcapro. *J Natl Cancer Inst*, 94(11):844–51, June 2002.

[12] J.B. Fan, A. Qliphant, R. Shen, B.G. Kermani, F. Garcia, K.L. Gunderson, M. Hansen, F. Steemers, S.L. Butler, P. Deloukas, L. Galver, S. Hunt, C. McBride, M. Bibikova, T. Rubano, J. Chen, E. Wickham, D. Doucet, W. Chang, D. Campbell, B. Zhang, S. Kruglyak, D. Bently, J. Haas, P. Rigault, L. Zhou, J. Stuelpnagel, and M.S. Chee. Highly parallel snp genotyping. *Cold Springs Harbor Symposia on Quantitative Biology*, 68:69–78, 2003.

[13] Martin Farrall and Andrew P Morris. Gearing up for genome-wide gene-association studies. *Hum Mol Genet*, 14 Spec No. 2:R157–62, October 2005.

[14] Simon Fiddy, David Cattermole, Dong Xie, Xiao Yuan Duan, and Richard Mott. Igs: An integrated system for genetic analysis. *BMC Bioinformatics*, 7:210, 2006.

[15] McKusick-Nathans Institute for Genetic Medicine and National Center for Biotechnology Information. Online mendelian inheritance in man omim (tm), `http://www.ncbi.nlm.nih.gov/omim/`, July 2006.

[16] D Fredman, M Siegfried, Y P Yuan, P Bork, H Lehväslaiho, and A J Brookes. Hgvbase: a human sequence variation database emphasizing data quality and a broad spectrum of data sources. *Nucleic Acids Res*, 30(1):387–91, January 2002.

[17] Nelson B Freimer and Chiara Sabatti. Human genetics: variants in common diseases. *Nature*, 445(7130):828–30, February 2007.

[18] Genelex. Genelex website. available at `http://www.genelex.com/`, May 2007.

[19] GeneTrack. Genetrack website. available at `http://www.genetrack.bc.ca`, July 2006.

[20] Gerd Gigerenzer and Adrian Edwards. Simple tools for understanding risks: from innumeracy to insight. *BMJ*, 327(7417):741–4, September 2003.

[21] Alan E Guttmacher and Francis S Collins. Welcome to the genomic era. *N Engl J Med*, 349(10):996–8, September 2003.

[22] Wayne D Hall, Katherine I Morley, and Jayne C Lucke. The prediction of disease risk in genomic medicine. *EMBO Rep*, 5 Spec No:S22–6, October 2004.

[23] Lynn B Jorde and Stephen P Wooding. Genetic variation, classification and 'race'. *Nat Genet*, 36(11 Suppl):S28–33, November 2004.

[24] K M Kelly and K Sweet. In search of a familial cancer risk assessment tool. *Clin Genet*, 71(1):76–83, January 2007.

[25] Muin J Khoury, Julian Little, Marta Gwinn, and John Pa Ioannidis. On the synthesis and interpretation of consistent but weak gene-disease associations in the era of genome-wide association studies. *Int J Epidemiol*, 36(2):439–45, April 2007.

[26] Cécile Libioulle, Edouard Louis, Sarah Hansoul, Cynthia Sandor, Frédéric Farnir, Denis Franchimont, Séverine Vermeire, Olivier Dewit, Martine de Vos, Anna Dixon, Bruno Demarche, Ivo Gut, Simon Heath, Mario Foglio, Liming Liang, Debby Laukens, Myriam Mni, Diana Zelenika, André Van Gossum, Paul Rutgeerts, Jacques Belaiche, Mark Lathrop, and Michel Georges. Novel crohn disease locus identified by genome-wide association maps to a gene desert on 5p13.1 and modulates expression of ptger4. *PLoS Genet*, 3(4):e58, April 2007.

[27] Yen-Ling Low, Sara Wedrén, and Jianjun Liu. High-throughput genomic technology in research and clinical management of breast cancer. evolving landscape of genetic epidemiological studies. *Breast Cancer Res*, 8(3):209, 2006.

[28] T M Marteau and R T Croyle. The new genetics. psychological responses to genetic testing. *BMJ*, 316(7132):693–6, February 1998.

[29] Ruth McPherson, Alexander Pertsemlidis, Nihan Kavaslar, Alexandre Stewart, Robert Roberts, David R Cox, David A Hinds, Len A Pennacchio, Anne Tybjaerg-Hansen, Aaron R Folsom, Eric Boerwinkle, Helen H Hobbs, and Jonathan C Cohen. A common allele on chromosome 9 associated with coronary heart disease. *Science*, May 2007.

[30] NCBI. dbgap: Database of genome wide association studies. url: http://www.ncbi.nlm.nih.gov/sites/entrez?db=gap, 2007.

[31] John Paling. Strategies to help patients understand risks. *BMJ*, 327(7417):745–8, September 2003.

[32] Lyle J Palmer and Lon R Cardon. Shaking the tree: mapping complex disease genes with linkage disequilibrium. *Lancet*, 366(9492):1223–34, October 2005.

[33] George P Patrinos and Anthony J Brookes. Dna, diseases and databases: disastrously deficient. *Trends Genet*, 21(6):333–8, June 2005.

[34] Richa Saxena, Benjamin F Voight, Valeriya Lyssenko, Noel P Burtt, Paul I W de Bakker, Hong Chen, Jeffrey J Roix, Sekar Kathiresan, Joel N Hirschhorn, Mark J Daly, Thomas E Hughes, Leif Groop, David Altshuler, Peter Almgren, Jose C Florez, Joanne Meyer, Kristin Ardlie, Kristina Bengtsson, Bo Isomaa, Guillaume Lettre, Ulf Lindblad, Helen N Lyon, Olle Melander, Christopher Newton-Cheh, Peter Nilsson, Marju Orho-Melander, Lennart Råstam, Elizabeth K Speliotes, Marja-Riitta Taskinen, Tiinamaija Tuomi, Candace Guiducci, Anna Berglund, Joyce Carlson, Lauren Gianniny, Rachel Hackett, Liselott Hall, Johan Holmkvist, Esa Laurila, Marketa Sjögren, Maria Sterner, Aarti Surti, Margareta Svensson, Malin Svensson, Ryan Tewhey, Brendan Blumenstiel, Melissa Parkin, Matthew Defelice, Rachel Barry, Wendy Brodeur, Jody Camarata, Nancy Chia, Mary Fava, John Gibbons, Bob Handsaker, Claire Healy, Kieu Nguyen, Casey Gates, Carrie Sougnez, Diane

Gage, Marcia Nizzari, Stacey B Gabriel, Gung-Wei Chirn, Qicheng Ma, Hemang Parikh, Delwood Richardson, Darrell Ricke, and Shaun Purcell. Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. *Science*, April 2007.

[35] Laura J Scott, Karen L Mohlke, Lori L Bonnycastle, Cristen J Willer, Yun Li, William L Duren, Michael R Erdos, Heather M Stringham, Peter S Chines, Anne U Jackson, Ludmila Prokunina-Olsson, Chia-Jen Ding, Amy J Swift, Narisu Narisu, Tianle Hu, Randall Pruim, Rui Xiao, Xiao-Yi Li, Karen N Conneely, Nancy L Riebow, Andrew G Sprau, Maurine Tong, Peggy P White, Kurt N Hetrick, Michael W Barnhart, Craig W Bark, Janet L Goldstein, Lee Watkins, Fang Xiang, Jouko Saramies, Thomas A Buchanan, Richard M Watanabe, Timo T Valle, Leena Kinnunen, Goncalo R Abecasis, Elizabeth W Pugh, Kimberly F Doheny, Richard N Bergman, Jaakko Tuomilehto, Francis S Collins, and Michael Boehnke. A genome-wide association study of type 2 diabetes in finns detects multiple susceptibility variants. *Science*, April 2007.

[36] Nameeta Shah, Michael V Teplitsky, Simon Minovitsky, Len A Pennacchio, Philip Hugenholtz, Bernd Hamann, and Inna L Dubchak. Snpvista: an interactive snp visualization tool. *BMC Bioinformatics*, 6:292, 2005.

[37] Robert Sladek, Ghislain Rocheleau, Johan Rung, Christian Dina, Lishuang Shen, David Serre, Philippe Boutin, Daniel Vincent, Alexandre Belisle, Samy Hadjadj, Beverley Balkau, Barbara Heude, Guillanume Charpentier, Thomas J. Hudson, Alexandre Montpetit, Alexey V. Pshezhetsky, Marc Prentki, Barry I. Posner, David J. Balding, David Meyre, Constantin Polychronakos, and Philippe Froguel. A genome-wide association study identifies novel risk loci for type 2 diabetes. *Nature*, 445:881–885, February 2007.

[38] E M Smigielski, K Sirotkin, M Ward, and S T Sherry. dbsnp: a database of single nucleotide polymorphisms. *Nucleic Acids Res*, 28(1):352–5, January 2000.

[39] Valgerdur Steinthorsdottir, Gudmar Thorleifsson, Inga Reynisdottir, Rafn Benediktsson, Thorbjorg Jonsdottir, G Bragi Walters, Unnur Styrkarsdottir, Solveig Gretarsdottir, Valur Emilsson, Shyamali Ghosh, Adam Baker, Steinunn Snorradottir, Hjordis Bjarnason, Maggie C Y Ng, Torben Hansen, Yu Bagger, Robert L Wilensky, Muredach P Reilly,

Adebowale Adeyemo, Yuanxiu Chen, Jie Zhou, Vilmundur Gudnason, Guanjie Chen, Hanxia Huang, Kerrie Lashley, Ayo Doumatey, Wing-Yee So, Ronald C Y Ma, Gitte Andersen, Knut Borch-Johnsen, Torben Jorgensen, Jana V van Vliet-Ostaptchouk, Marten H Hofker, Cisca Wijmenga, Claus Christiansen, Daniel J Rader, Charles Rotimi, Mark Gurney, Juliana C N Chan, Oluf Pedersen, Gunnar Sigurdsson, Jeffrey R Gulcher, Unnur Thorsteinsdottir, Augustine Kong, and Kari Stefansson. A variant in cdkal1 influences insulin response and risk of type 2 diabetes. *Nat Genet*, April 2007.

[40] Gudmundur A Thorisson, Albert V Smith, Lalitha Krishnan, and Lincoln D Stein. The international hapmap project web site. *Genome Res*, 15(11):1592–3, November 2005.

[41] Wenyi Wang, Sining Chen, Kieran A Brune, Ralph H Hruban, Giovanni Parmigiani, and Alison P Klein. Pancpro: risk assessment for individuals with a family history of pancreatic cancer. *J Clin Oncol*, 25(11):1417–22, April 2007.

[42] A J Wright, J Weinman, and T M Marteau. The impact of learning of a genetic predisposition to nicotine dependence: an analogue study. *Tob Control*, 12(2):227–30, June 2003.

[43] Nan Yang, Hongzhe Li, Lindsey A Criswell, Peter K Gregersen, Marta E Alarcon-Riquelme, Rick Kittles, Russell Shigeta, Gabriel Silva, Pragna I Patel, John W Belmont, and Michael F Seldin. Examination of ancestry and ethnic affiliation using highly informative diallelic dna markers: application to diverse and admixed populations and implications for clinical epidemiology and forensic medicine. *Hum Genet*, 118(3-4):382–92, December 2005.

[44] Quanhe Yang, Muin J. Khoury, Lorenzo Botto, J.M. Friedman, and Dana Flanders. Improving the prediction of complex diseases by testing for multiple disease-susceptibility genes. *American Journal of Human Genetics*, 72:636–649, 2003.

[45] Lan-Juan Zhao, Miao-Xin Li, Yan-Fang Guo, Fu-Hua Xu, Jin-Long Li, and Hong-Wen Deng. Snpp: automating large-scale snp genotype data management. *Bioinformatics*, 21(2):266–8, January 2005.

Chapter 3

Conclusions and Future Directions

3.1 Further Observations

One of the most important observations noted during the D-GRIP development and testing was the lack of computationally efficient organization of existing and new discoveries in the genetics field [5, 10]. There has been an explosion of data from the recent progress in disease genetics field, and even though currently there are many types of mutation databases, the progress towards creation of new databases has been slow. The challenges involved are often technical in nature, such as, gathering, exchanging, integrating and interpreting the disease-related information. However, arguably the lack of targeted funding and the inherent bias towards making new discoveries rather than managing existing data are one of the main underlying problems [10].

In order to overcome the technical limitations of creating a comprehensive, computationally exploitable genotype-phenotype database, a few goals must be met. For easy computational access, complex phenotype data models that extensively utilize phenotype ontologies will be required. By using ontologies, a standard vocabulary can be established for use of terms, which will help integrate various types of data and make analysis computationally easier. Initially, the DNA changes related to phenotypes can be represented in a structured and standardized way. Then, a basic framework for gathering, integrating, analyzing and updating the stored information will be required. Given the enormous amounts of data being generated, a systematic and standardized way to manage phenotype data will be a necessity, which will require international cooperation and open access to anonymous data. Ultimately, an ideal genotype-phenotype database will provide a systems biology approach where all information, such as that derived from the genome, transcriptome, proteome and metabolome, pertaining to the connection between genotypic differences and phenotypic consequences will be recorded.

The second important observation that resulted from my work on D-GRIP was the limited number of variants that are known to be associated with complex diseases. Even though individual genome wide association studies(GWAs) are publishing results for many diseases [12, 11, 1, 2, 4, 9], most of the studies report only a few disease associated variants [3, 8]. In addition, the reported effects of individual genetic variants associated to common diseases are small (risk ratios $\leqslant 2.0$). Although, it has been shown that the combined effects of a moderate number (fewer than 20) of common genetic variants (with relative ratios $\leqslant 2.0$) could explain 50% of the burden of disease in a population [13]; there are numerous challenges with genome-wide association studies. These challenges include, for example, significance chasing bias (including publication bias, selective analysis and reporting bias), population stratification (due to heterogenous populations mixtures), misclassification of exposures and outcome, and the inherent problems that include, failure to detect gene-gene and gene-environment interactions, limited sample size, statistical power and false positive associations. All these issues can lead to difficulty in finding biologically meaningful genetic associations and thus slow the progress of understanding complex diseases.

In order to alleviate and infer true disease-associated variants from numerous GWAs, standards should be established for presenting and interpreting the accumulated evidence. Efforts by the Human Genome Epidemiology Network (HuGENet) are ongoing in developing systematic approaches for assessing combined evidence of disease associated variants. The approaches include criteria such as biological plausibility, experimental evidence, sound methods for conduct and analysis, and appropriate replication [8]. The opportunity to develop methods and standards for measuring, validating and interpreting genetic associations will be high in the next few years and will ultimately lead to benefit for individuals and population health.

3.2 Future Considerations

The goal of shifting the current medical paradigm from a reactive to preventative approach through personalized risk profiles appears within reach long-term. The generation of genetic risk profiles is intended to improve disease prevention by prompting at-risk individuals to take specific preventative actions that usually involve environmental exposures, diet or other lifestyle changes. However, before genetic risk assessment tools can be used in a clinical setting, an evaluation of the clinical utility of such tools needs to be conducted [7]

Clinical utility of a test refers to the likelihood a diagnostic test will lead to improved health outcomes [7]. For individuals with positive test results, the clinical utility depends on the availability, safety and effectiveness of therapeutic measures. The recommendation for ensuring clinical utility for any genetic test is to consider the clinical and social outcomes of the test. Clinical outcomes depend on effective changes in lifestyle due to positive test result. The social outcomes depend on the psychosocial, ethical, legal and social issues related to receiving a positive or negative outcome. Both clinical and social outcomes are important because they both contribute to the net balance between benefits and harms of genetic testing [6]. Thus, future evaluation of genomic profiles should encompass and clearly address validity of the test, clinical utility and social utility of the test.

Regardless of the intended audience for a genetic risk profiling software, two crucial criteria are necessary for providing a genetic profile test. First, due to the still limited knowledge about clinical implications of such tests, the benefits and limitations of the tests should be clearly explained. Such limitations should be explicitly addressed, and individuals who provide tests should disclose what is known and not known about the test. Second, the tests should be offered in a controlled environment such that individual test takers are counseled about the results and implications of the tests. By having transparency when providing the genetic profile test and counseling the individual test taker, informed decisions can be made by health professionals, patients and general pubic.

Lastly, consensus needs to be achieved on when genomic profiling has achieved an acceptable standard in a clinical setting. In the future, genomic profiling will likely become common and thus the level of evidence that justifies clinical use of genomic profiling requires careful thought. It is recommended to develop an accepted process that incorporates defined procedures for evaluating evidence and reaching conclusions that include input from clinicians, health care payers and consumers.

3.3 Conclusion

Given the advent of new genotyping technologies and the rapid new discovery of new disease associated variants, experts have predicted that future medical care will become more personalized and geared towards disease prevention. We created a prototype web tool, called, DNA Genetic Risk Information Profile (D-GRIP), which predicts disease risk profiles based on an individual's genotype. The project outlined the current bioinformatic and scientific limitations involved in creating a genetic risk assessment software and addressed the main issues involved in the creation, evaluation and utility of such a tool in a clinical setting. By overcoming the major limitations and addressing the important issues, a viable and useful genetic risk profiling software is plausible in the future and thus will lead to a change in the way medicine is currently practiced.

Bibliography

[1] A E Baum, N Akula, M Cabanero, I Cardona, W Corona, B Klemens, T G Schulze, S Cichon, M Rietschel, M M Nöthen, A Georgi, J Schumacher, M Schwarz, R Abou Jamra, S Höfels, P Propping, J Satagopan, S D Detera-Wadleigh, J Hardy, and F J McMahon. A genome-wide association study implicates diacylglycerol kinase eta (dgkh) and several other genes in the etiology of bipolar disorder. *Mol Psychiatry*, May 2007.

[2] Keith D Coon, Amanda J Myers, David W Craig, Jennifer A Webster, John V Pearson, Diane Hu Lince, Victoria L Zismann, Thomas G Beach, Doris Leung, Leslie Bryden, Rebecca F Halperin, Lauren Marlowe, Mona Kaleem, Douglas G Walker, Rivka Ravid, Christopher B Heward, Joseph Rogers, Andreas Papassotiropoulos, Eric M Reiman, John Hardy, and Dietrich A Stephan. A high-density whole-genome association study reveals that apoe is the major susceptibility gene for sporadic late-onset alzheimer's disease. *J Clin Psychiatry*, 68(4):613–8, April 2007.

[3] Jennifer Couzin and Jocelyn Kaiser. Genome-wide association. closing the net on common disease genes. *Science*, 316(5826):820–2, May 2007.

[4] J R Fraser Cummings, Rachel Cooney, Saad Pathan, Carl A Anderson, Jeffrey C Barrett, John Beckly, Alessandra Geremia, Laura Hancock, Changcun Guo, Tariq Ahmad, Lon R Cardon, and Derek P Jewell. Confirmation of the role of atg16l1 as a crohn's disease susceptibility gene. *Inflamm Bowel Dis*, April 2007.

[5] Angela Frodsham and Julian Higgins. Online genetic databases informing human genome epidemiology. *BMC Med Res Methodol*, 7(1):31, July 2007.

[6] Scott D Grosse and Muin J Khoury. What is the clinical utility of genetic testing? *Genet Med*, 8(7):448–50, July 2006.

[7] Susanne B Haga, Muin J Khoury, and Wylie Burke. Genomic profiling to promote a healthy lifestyle: not ready for prime time. *Nat Genet*, 34(4):347–50, August 2003.

[8] Muin J Khoury, Julian Little, Marta Gwinn, and John Pa Ioannidis. On the synthesis and interpretation of consistent but weak gene-disease associations in the era of genome-wide association studies. *Int J Epidemiol*, 36(2):439–45, April 2007.

[9] Ruth McPherson, Alexander Pertsemlidis, Nihan Kavaslar, Alexandre Stewart, Robert Roberts, David R Cox, David A Hinds, Len A Pennacchio, Anne Tybjaerg-Hansen, Aaron R Folsom, Eric Boerwinkle, Helen H Hobbs, and Jonathan C Cohen. A common allele on chromosome 9 associated with coronary heart disease. *Science*, May 2007.

[10] George P Patrinos and Anthony J Brookes. Dna, diseases and databases: disastrously deficient. *Trends Genet*, 21(6):333–8, June 2005.

[11] Laura J Scott, Karen L Mohlke, Lori L Bonnycastle, Cristen J Willer, Yun Li, William L Duren, Michael R Erdos, Heather M Stringham, Peter S Chines, Anne U Jackson, Ludmila Prokunina-Olsson, Chia-Jen Ding, Amy J Swift, Narisu Narisu, Tianle Hu, Randall Pruim, Rui Xiao, Xiao-Yi Li, Karen N Conneely, Nancy L Riebow, Andrew G Sprau, Maurine Tong, Peggy P White, Kurt N Hetrick, Michael W Barnhart, Craig W Bark, Janet L Goldstein, Lee Watkins, Fang Xiang, Jouko Saramies, Thomas A Buchanan, Richard M Watanabe, Timo T Valle, Leena Kinnunen, Goncalo R Abecasis, Elizabeth W Pugh, Kimberly F Doheny, Richard N Bergman, Jaakko Tuomilehto, Francis .S Collins, and Michael Boehnke. A genome-wide association study of type 2 diabetes in finns detects multiple susceptibility variants. *Science*, April 2007.

[12] Robert Sladek, Ghislain Rocheleau, Johan Rung, Christian Dina, Lishuang Shen, David Serre, Philippe Boutin, Daniel Vincent, Alexandre Belisle, Samy Hadjadj, Beverley Balkau, Barbara Heude, Guillanume Charpentier, Thomas J. Hudson, Alexandre Montpetit, Alexey V. Pshezhetsky, Marc Prentki, Barry I. Posner, David J. Balding, David Meyre, Constantin Polychronakos, and Philippe Froguel. A genome-wide association study identifies novel risk loci for type 2 diabetes. *Nature*, 445:881–885, February 2007.

[13] Quanhe Yang, Muin J Khoury, Jm Friedman, Julian Little, and W Dana Flanders. How many genes underlie the occurrence of common complex diseases in the population? *Int J Epidemiol*, 34(5):1129–37, October 2005.

Appendix A

Feedback from Experts

A.1 Questions

The set of questions asked to each type of expert (clinical geneticist, molecular geneticist, genetic counselors and biostatisticians) are listed below.

1. Any comments on the user-interface of D-GRIP?
   - The input page?
   - The output page?
2. Any comments or references to available risk models that predict risk based on genotype data?
   - How to include age specific risk prediction without raw data?
3. How should an ideal system handle various complex diseases? Treat each separately with disease-specific risk model?
4. The system shows a very fatalistic view. Do you think we should include more positive news?
5. Who could be a potential user of D-GRIP?
   - Genetic Counselors?
   - Family Physicians?

- Insurance companies?
- Lay public?
- Yourself?

6. How many years down the line can you see this being used (respectively for each of the potential users from previous question)?

7. Do you think we should store people's genotype data? What about family doctor's storing their patient's genotype data?

8. What are some of the implications you see from using such a system?

    - Personal implications?
    - Effect on patients?
    - Societal implications?

9. In what journal can you see this type of paper being published?

A.2 Feedback

A summary of the feedback provided by several experts is detailed below. The experts consisted of two biostatisticians, two molecular geneticists, five clinical geneticists and 12 MSc genetic counseling students. The comments and recommendations are categorized into various aspects of D-GRIP, for example, user interface issues regarding input and output features, core of D-GRIP dealing with DNA-Disease database and risk prediction model, issues pertaining to the users and any ethical, legal and social implications.

A.2.1 User Interface

Input and general usability

- Allow option for users to provide family history along with genotype data.
- Ethnicity classification is currently biased. Provide two options, one user-specified ethnicity and two, calculate ethnicity based on a verified and reliable predetermined-determined markers from genotype data provided. Consensus was to calculate the ethnicity but only when calculations can be done reliably.
- When more data is available, allow input for copy number variantions data.
- Provide a disclaimer that explicitly informs the user of all the limitations and assumptions of the software.

- As it is currently, keep the interface simple and easy to use.

Risk Profile Report

- Tailor the final risk report towards the intended user. Currently, the view is more geared towards genetic researchers and counselors. In contrast, for a family physician or a consumer, provide a 'Patient view' where communication of probabilities and risk is done visually, links to prevention and therapeutic options and any relevant links for lifestyle and behavior changes are provided.
- Provide the option of restricting analysis to specific diseases, for instance, diseases where prevention is an option versus where currently no preventative options are available.

A.2.2 D-GRIP Core

Diseases, DNA-Disease database

- Implement a meta-analysis engine for each disease so that whenever new studies are published, the entire database is updated. In addition, whenever such updates are performed, create a notification system for users to inform them.
- Store gene-gene and gene-environment and epigenetic information in DNA-Disease database. Data on gender and age related to diseases is very important, especially for age-dependent diseases.

- When information regarding copy number variations related to diseases is available, store this into the DNA-Disease database.

- Also store intermediate phenotypes associated with markers in addition to disease associated markers.

Risk prediction issues

- Implement disease specific risk models so that each disease is treated separately. Also, allow advanced users to choose multiple risk prediction models for each disease.

- When data are available, incorporate gene-gene and gene-environment effects into the respective disease risk models.

- Perform rigours validation of each predictive model and prediction. Show the results of the tests performed, such as sensitivity, specificity, positive predictive values. Ensure validations of the prediction models is performed with genotype data that is not part of the case-control population data in the DNA-Disease database. Currently, such volume of data for testing is not available so future versions will require this feature. Also, provide links to studies supporting the risk predictions models for respective diseases.

A.2.3 Potential Users

- Genetic counselors are a good initial user for the software. During initial deployment of D-GRIP, user training will be required so that all limitations and proper interpretation of results is performed.

- Family physicians (or in a primary care setting) can be other potential users. But training for family physicians on how to use and interpret results from such a tool will be a necessity.

- Potentially, general public could act as consumers of such a software. But all implications will need to be addressed by health professionals, governments and industry before such a software is released to the general public.

- Insurance companies could also be potential users but the many social, legal and ethical implications will need to be addressed and a supporting framework will need to be implemented so handle third party use of genetic data.

- As mentioned, user interface of software should be tailored towards the user.

- The consensus was that currently, D-GRIP is ahead of its time. But a similar software can be seen used in the next 5-10years time. However, better understanding of disease associated variants and reliable predictions will be a necessity.

- Until proper standards and procedures are developed to handle all the ethical, legal and social implications, such a software should always be used under a guided setting where the counseled individual is explained all the limitations and provide guidance in understanding the results from such a software.

A.2.4 Implications

- As it is currently, there should be no user identifiable storing of genotype data. User genotype data can be stored only when the family physician is the user and storing the patient's genotype data. However, in the future, proper framework will be required to handle genetic data management, to support privacy, confidentiality and anonymity.
- The level of care required in helping the general public interpret and understand the results is enormous and should be done appropriately.
- At the current rate, not enough genetic counselors to support the future demand for counseling of individuals wanting a genetic risk profile.
- All necessary ethical, social and legal implications will need to be addressed by the providers of such a tool.

Appendix B

D-GRIP User Manual

B.1 Introduction to D-GRIP

This user Guide assumes you have access to D-GRIP since D-GRIP is a closed and secure web tool. The guide explains the various features of D-GRIP and provides a brief walk through. This guide is not intended to explain the results of D-GRIP or how to interpret them.

The guide explains:

- The overall processes.
- Basic features that are available.

B.1.1 D-GRIP System

DNA Genetic Risk Information Profile (D-GRIP) is a genotype analysis system that predicts an individual's genetic risk profile based on the genotype. The system can take as input, observed genotypes of up to one million positions of known single nucleotide polymorphisms (SNPs) in human populations.

The flow of information in D-GRIP begins from the input of user data. The user is asked to fill in demographic information (ethnic background, age and gender) and a genotype file which is parsed and temporarily stored. Next, The system compares the genotyping results to an internal DNA-DISEASE risk database and for each disease, calculates a risk score for developing the disease. Finally, a tabular output of potential diseases with the relevant disease risk for the individual is displayed.

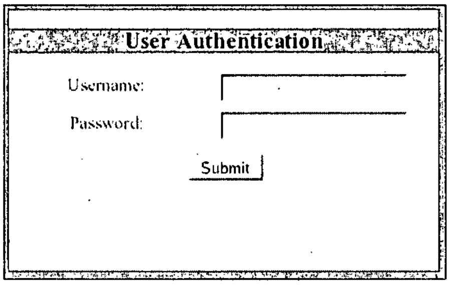
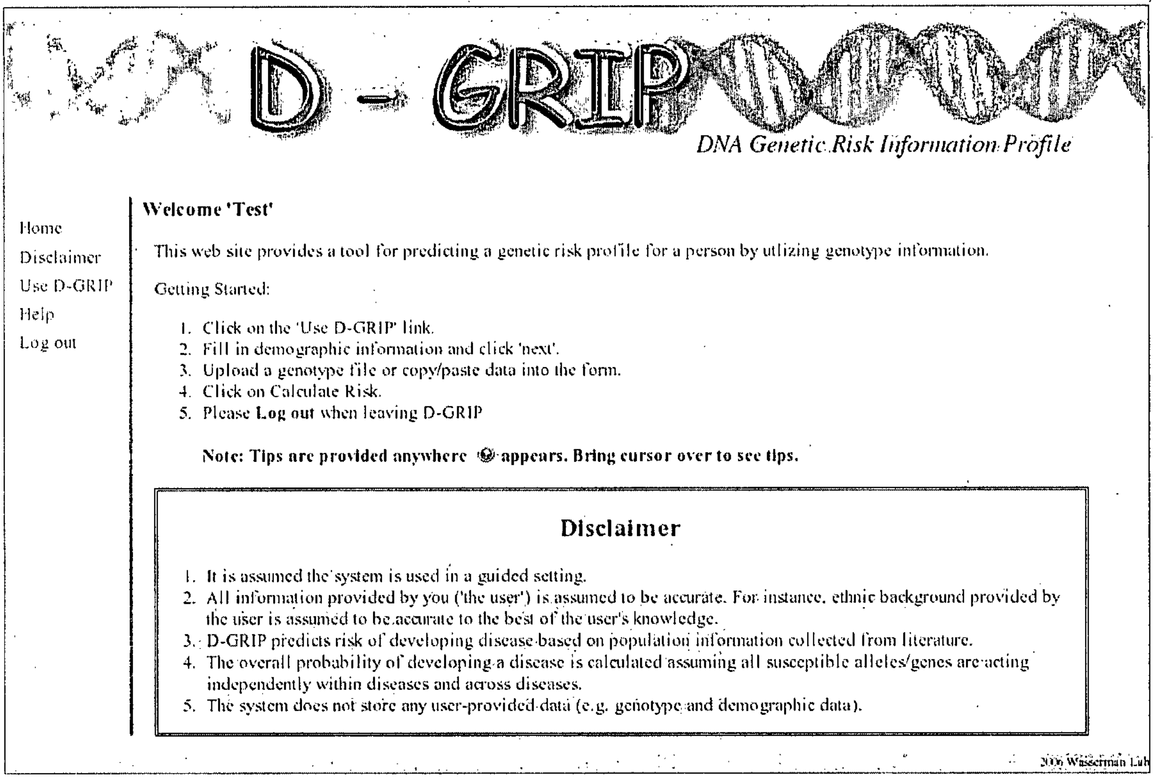


Figure B.1: The entry into D-GRIP occurs with user authentication. A valid username and password is required to access D-GRIP.

B.2 D-GRIP Features

There are various features in D-GRIP and a detailed description of each with illustrations is provided below. The page is laid out with a menu on the left and all the relevant content on the right. The menu contains navigation links to Home page (Figure B.2), Disclaimer page, Use D-GRIP page, Help page and link to Log out of D-GRIP.



Welcome 'Test'

This web site provides a tool for predicting a genetic risk profile for a person by utlizing genotype information.

Getting Started:

1. Click on the 'Use D-GRIP' link.
2. Fill in demographic information and click 'next'.
3. Upload a genotype file or copy/paste data into the form.
4. Click on Calculate Risk.
5. Please **Log out** when leaving D-GRIP

Note: Tips are provided anywhere ☺ appears. Bring cursor over to see tips.

Disclaimer

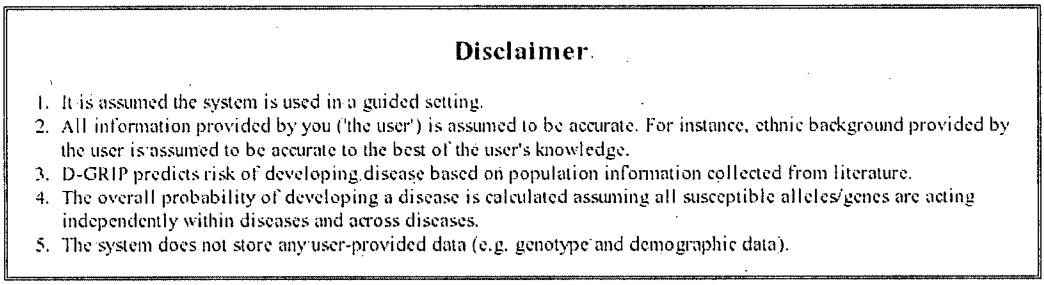
1. It is assumed the system is used in a guided setting.
2. All information provided by you ('the user') is assumed to be accurate. For instance, ethnic background provided by the user is assumed to be accurate to the best of the user's knowledge.
3. D-GRIP predicts risk of developing disease based on population information collected from literature.
4. The overall probability of developing a disease is calculated assuming all susceptible alleles/genes are acting independently within diseases and across diseases.
5. The system does not store any user-provided data (e.g. genotype and demographic data).

2006 Wasserman Lab

Figure B.2: A snapshot of D-GRIP's main page. The page describes instructions on how to use D-GRIP and outlines a disclaimer for the user to read.

B.2.1 Disclaimer

The disclaimer explicitly outlines the assumptions made by D-GRIP (Figure B.3). The disclaimer is shown on the first page, when the user accesses the site. Also, a separate link is provided to view the disclaimer.



Disclaimer

1. It is assumed the system is used in a guided setting.
2. All information provided by you ('the user') is assumed to be accurate. For instance, ethnic background provided by the user is assumed to be accurate to the best of the user's knowledge.
3. D-GRIP predicts risk of developing disease based on population information collected from literature.
4. The overall probability of developing a disease is calculated assuming all susceptible alleles/genes are acting independently within diseases and across diseases.
5. The system does not store any user-provided data (e.g. genotype and demographic data).

Figure B.3: The assumptions made by D-GRIP are listed as a disclaimer and shown here

B.2.2 Input

The input page can be accessed by clicking on the 'Use D-GRIP' link in the menu on the left. The input for D-GRIP occurs in two steps. First, demographic information and configuration options are presented. Next, genotype data is requested from the user.

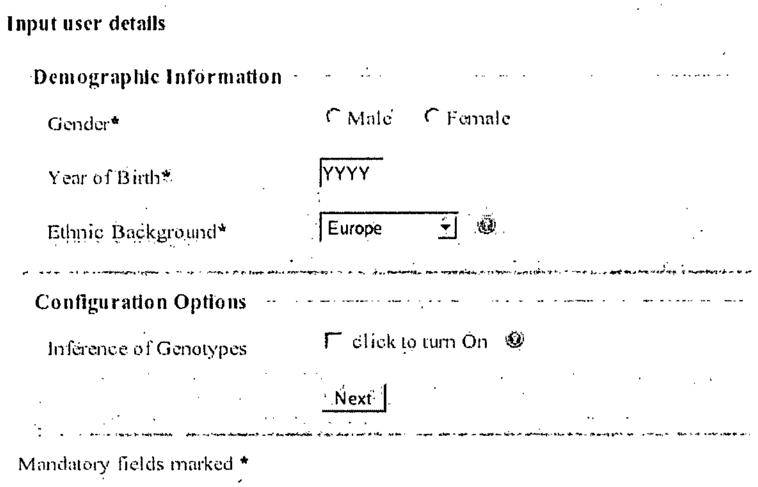
Demographic Information

Figure B.4 shows the first stage of the input. The mandatory information requested from the user is Gender, Age and Ethnic background.

For the Age, the user enters the year of birth. For the Ethnic background, the user should select the most appropriate option based on the geographic ancestry of the user. The options presented are: Africa, Asia, Europe, Pacific, First nations/Aboriginals and Mixed.

The configuration option currently has one checkbox for 'inference of genotypes'. The inference of genotypes utilizes the haplotype information from the Hapmap Project Website to infer disease-associated genotypes from the genotype data provided by the user. By default, the inference option is turned off (no tick in checkbox).

Once the user fills in the demographic information form, proceed to loading genotype data by clicking the 'Next' button.



Input user details

Demographic Information

Gender* ○ Male ○ Female

Year of Birth* YYYY

Ethnic Background* Europe

Configuration Options

Inference of Genotypes □ click to turn On

Mandatory fields marked *

Figure B.4: Demographic information and configuration options submitted to D-GRIP are shown here.

Genotype Data

Figure B.5 shows how the genotype data can be loaded into D-GRIP.

There are two ways to load the genotype data. The copy/paste option allows the user to copy the genotype data and paste into the text area provided. The mandatory fields for copy/paste form are file format, file name and genotype data. After filling in the form, click on 'Calculate Risk' button to generate the risk profile output.

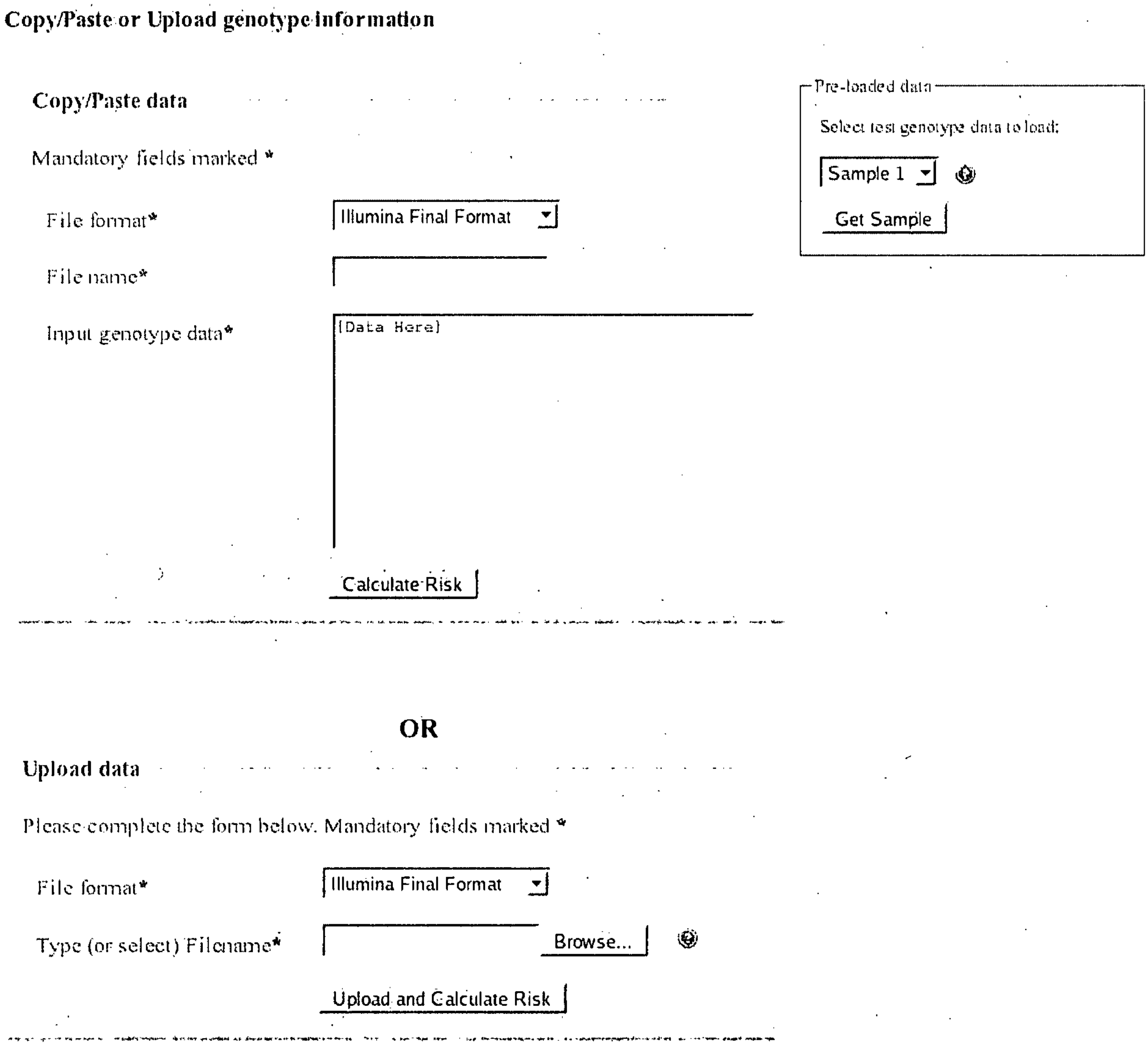
For the uploading of genotype file, the mandatory fields are file format and address where the file is stored. The user may use the 'Browse' button to find the genotype file on the hard drive. Note, the maximum allowed size for the genotype file to be uploaded is 10Mb. This size limit can contain genotypes for more than 1 million SNPs in the file. After filling in the form, click on 'Upload File and Calculate Risk' button to generate the risk profile output.

Currently, D-GRIP accepts two file formats: Illumina Final format and Affymetrix Text Output. An example of the respective genotype file formats are shown in Figure B.6.

The Illumina Final format can be obtained by generating a tab delimited 'Final Report' when using the Illumina platform's BeadStudio Genotyping Module software. The only fields necessary are: SNP Name, Allele 1 and Allele 2. The sample Id and GC score are not necessary for D-GRIP.

The Affymetrix text output can be obtained by using the SNP Export feature in the Affymetrix GeneChip Genotyping Analysis Software and generating a tab delimited output file. Again, the only fields necessary are SNP identifier and SNP genotype (two alleles).

In Figure B.5, next to the copy/paste form is a box with 'Pre-loaded' data. To illustrate D-GRIP, sample genotype files have been created and can be loaded using this 'Pre-loaded' data box. Simply select the particular



Copy/Paste or Upload genotype information

Copy/Paste data

Mandatory fields marked *

File format* Illumina Final Format

File name*

Input genotype data* [Data Here]

Pre-loaded data

Select test genotype data to load:

Sample 1

OR

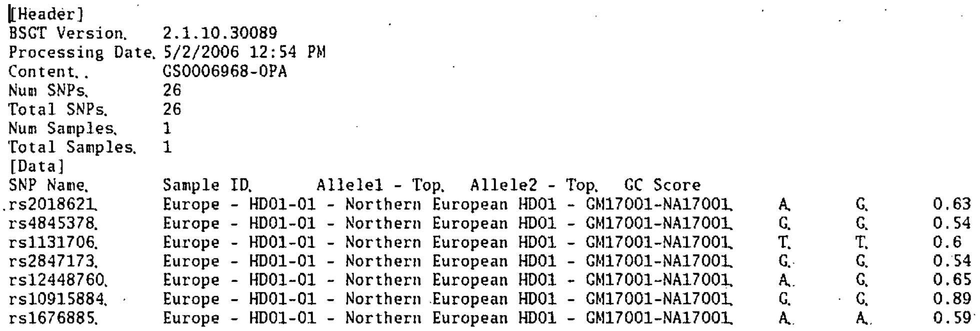
Upload data

Please complete the form below. Mandatory fields marked *

File format* Illumina Final Format

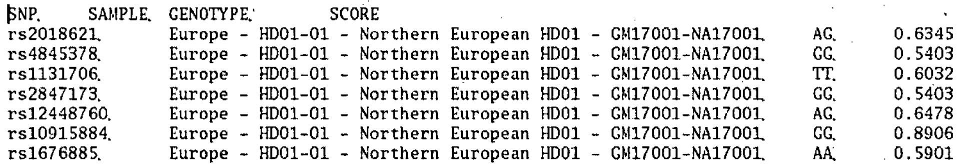
Type (or select) Filename*

Figure B.5: Form for submitting the genotype data is shown here. The user can either copy/paste the genotype data or upload a genotype file. A set of sample genotypes are provided and can be loaded into the copy/paste form by clicking on 'Get Sample'.



```
[Header]
BSGT Version.    2.1.10.30089
Processing Date. 5/2/2006 12:54 PM
Content..        GS0006968-OPA
Num SNPs.        26
Total SNPs.      26
Num Samples.     1
Total Samples.   1
[Data]
SNP Name.        Sample ID.      Allele1 - Top.  Allele2 - Top.  GC Score
rs2018621.       Europe - HD01-01 - Northern European HD01 - GM17001-NA17001.    A.       G.        0.63
rs4845378.       Europe - HD01-01 - Northern European HD01 - GM17001-NA17001.    G.       G.        0.54
rs1131706.       Europe - HD01-01 - Northern European HD01 - GM17001-NA17001.    T.       T.        0.6
rs2847173.       Europe - HD01-01 - Northern European HD01 - GM17001-NA17001.    G.       G.        0.54
rs12448760.      Europe - HD01-01 - Northern European HD01 - GM17001-NA17001.    A.       G.        0.65
rs10915884.      Europe - HD01-01 - Northern European HD01 - GM17001-NA17001.    G.       G.        0.89
rs1676885.       Europe - HD01-01 - Northern European HD01 - GM17001-NA17001.    A.       A.        0.59
```

(a) Illumina final format sample file

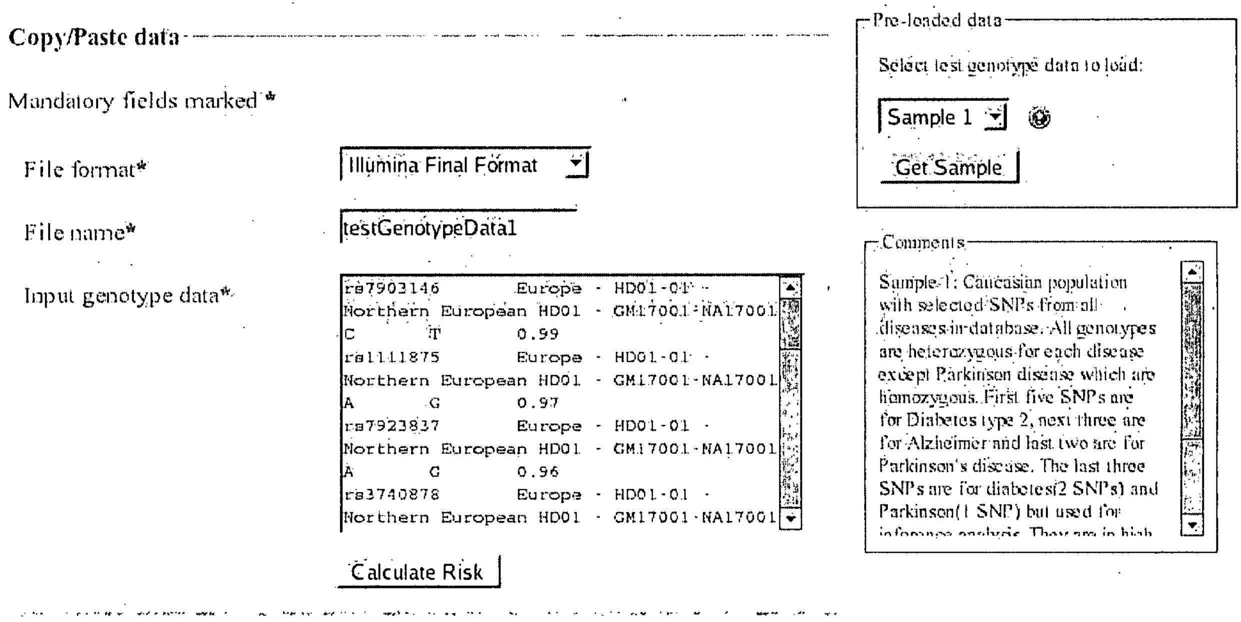


```
SNP.      SAMPLE.  GENOTYPE.           SCORE
rs2018621.         Europe - HD01-01 - Northern European HD01 - GM17001-NA17001.     AG.        0.6345
rs4845378.         Europe - HD01-01 - Northern European HD01 - GM17001-NA17001.     GG.        0.5403
rs1131706.         Europe - HD01-01 - Northern European HD01 - GM17001-NA17001.     TT.        0.6032
rs2847173.         Europe - HD01-01 - Northern European HD01 - GM17001-NA17001.     GG.        0.5403
rs12448760.        Europe - HD01-01 - Northern European HD01 - GM17001-NA17001.     AG.        0.6478
rs10915884.        Europe - HD01-01 - Northern European HD01 - GM17001-NA17001.     GG.        0.8906
rs1676885.         Europe - HD01-01 - Northern European HD01 - GM17001-NA17001.     AA.        0.5901
```

(b) Affymetrix text output sample file

Figure B.6: The Illumina and Affymetrix tab-delimited file formats for D-GRIP. The respective column names are shown at the top.

sample and click on 'Get Sample'. A 'Comments' box appears describing the sample file and the sample file appears in the copy/paste text area. Genotype Sample 1 is shown in Figure B.7.



Copy/Paste data

Mandatory fields marked *

File format* Illumina Final Format

File name* testGenotypeData1

Input genotype data*

```
rs7903146          Europe  -  HD01-01  -
Northern European HD01  -  GM17001-NA17001
C        T         0.99
rs1111875          Europe  -  HD01-01  -
Northern European HD01  -  GM17001-NA17001
A        G         0.97
rs7923837          Europe  -  HD01-01  -
Northern European HD01  -  GM17001-NA17001
A        G         0.96
rs3740878          Europe  -  HD01-01  -
Northern European HD01  -  GM17001-NA17001
```

Pre-loaded data

Select test genotype data to load: Sample 1

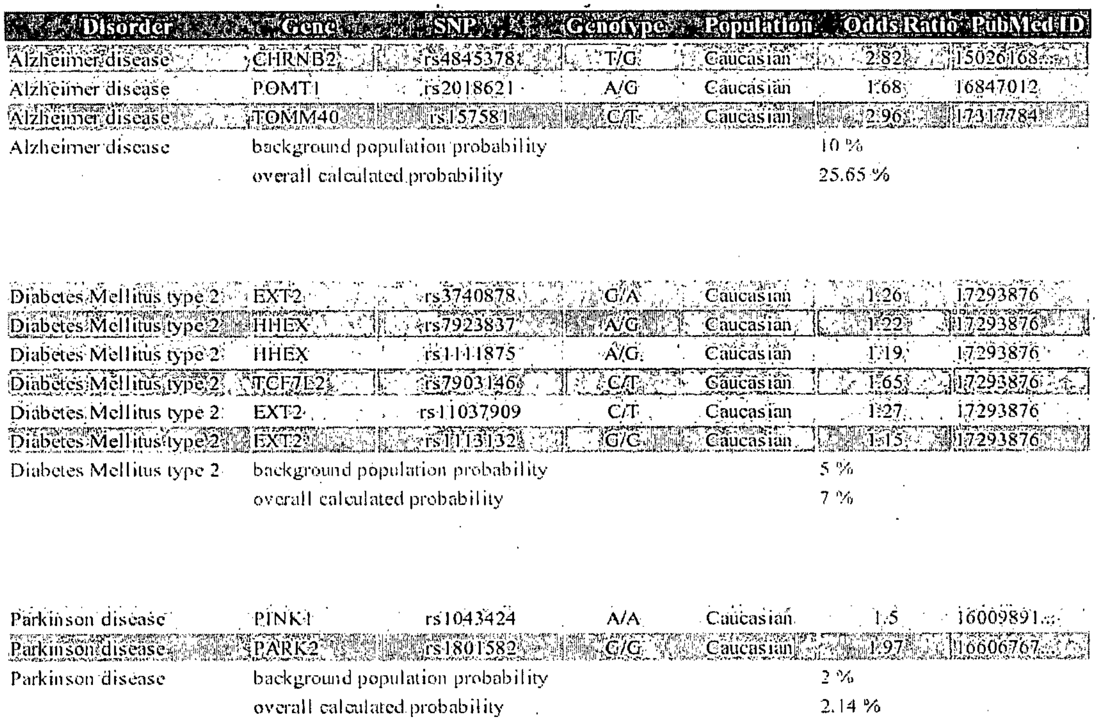
Comments

Sample 1: Caucasian population with selected SNPs from all diseases in database. All genotypes are heterozygous for each disease except Parkinson disease which are homozygous. First five SNPs are for Diabetes type 2, next three are for Alzheimer and last two are for Parkinson's disease. The last three SNPs are for diabetes(2 SNPs) and Parkinson(1 SNP) but used for inference analysis. They are in high

Figure B.7: Genotype sample 1 is loaded into the copy/paste form by clicking on 'Get Sample'. A description of the sample genotype file are illustrated in the 'Comments' box.

B.2.3 Output

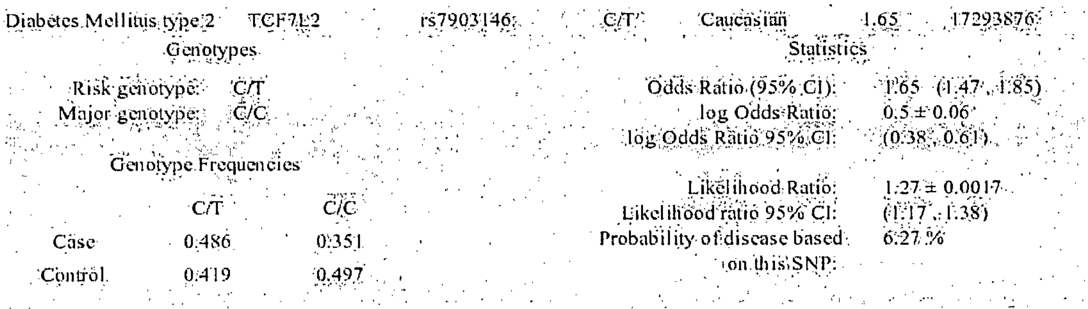
An example output of D-GRIP is shown in Figure B.8. The output of D-GRIP is table that shows user's SNPs that matched disease-associated SNPs. The table illustrates the disorder, gene, SNP and genotype associated with the disorder, population in which the SNP occurs, calculated odds ratio and link to Pubmed for literature articles supporting the association.



Disorder	Gene	SNP	Genotype	Population	Odds Ratio	PubMed ID
Alzheimer disease	CHRNB2	rs4845378	T/G	Caucasian	2.82	15026168...
Alzheimer disease	POMT1	rs2018621	A/G	Caucasian	1.68	16847012
Alzheimer disease	TOMM40	rs157581	C/T	Caucasian	2.96	17317784
Alzheimer disease	background population probability				10 %	
	overall calculated probability				25.65 %	
Diabetes Mellitus type 2	EXT2	rs3740878	G/A	Caucasian	1.26	17293876
Diabetes Mellitus type 2	HHEX	rs7923837	A/G	Caucasian	1.22	17293876
Diabetes Mellitus type 2	HHEX	rs1111875	A/G	Caucasian	1.19	17293876
Diabetes Mellitus type 2	TCF7L2	rs7903146	C/T	Caucasian	1.65	17293876
Diabetes Mellitus type 2	EXT2	rs11037909	C/T	Caucasian	1.27	17293876
Diabetes Mellitus type 2	EXT2	rs1113132	G/C	Caucasian	1.15	17293876
Diabetes Mellitus type 2	background population probability				5 %	
	overall calculated probability				7 %	
Parkinson disease	PINK1	rs1043424	A/A	Caucasian	1.5	16009891...
Parkinson disease	PARK2	rs1801582	G/G	Caucasian	1.97	16606767...
Parkinson disease	background population probability				2 %	
	overall calculated probability				2.14 %	

Figure B.8: D-GRIP risk profile sample output. The output illustrates 3 diseases, Alzheimer's, Diabetes type 2 and Parkinson's disease. The respective associated SNPs with each disease is shown. The background and overall calculated probability of developing the disease is also shown.

The user can click on the gene name, and disorder name for external links to genbank and OMIM respectively. In addition, by clicking on each SNP row, more details about the SNP can be seen (Figure B.9).



Diabetes Mellitus type 2	TCF7L2	rs7903146	C/T	Caucasian	1.65	17293876

Genotypes

Risk genotype:	C/T
Major genotype:	C/C

Genotype Frequencies

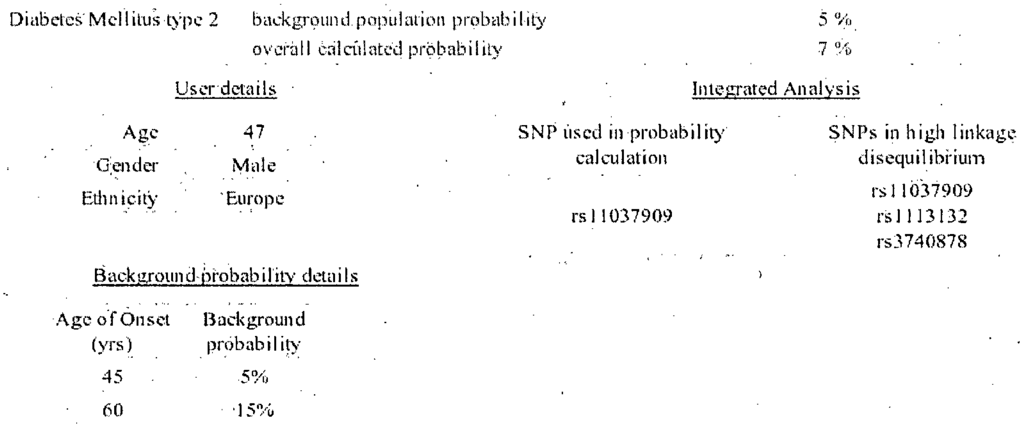
	C/T	C/C
Case	0.486	0.351
Control	0.419	0.497

Statistics

Odds Ratio (95% CI):	1.65 (1.47 , 1.85)
log Odds Ratio:	0.5 ± 0.06
log Odds Ratio 95% CI:	(0.38 , 0.61)
Likelihood Ratio:	1.27 ± 0.0017
Likelihood ratio 95% CI:	(1.17 , 1.38)
Probability of disease based on this SNP:	6.27 %

Figure B.9: Details about one SNP from Diabetes type II disease.

More details about the probability calculation for each disease can be seen by clicking on the probability row (Figure B.10). If there are SNPs found that are in high linkage disequilibrium ($r^2 > 0.8$) then integrated analysis is performed where only one SNP from the set of high LD SNPs is chosen to be in the overall calculated probability. This is illustrated on the right side of Figure B.10.



Diabetes Mellitus type 2	background population probability	5 %
	overall calculated probability	7 %

User details

Age	47
Gender	Male
Ethnicity	Europe

Integrated Analysis

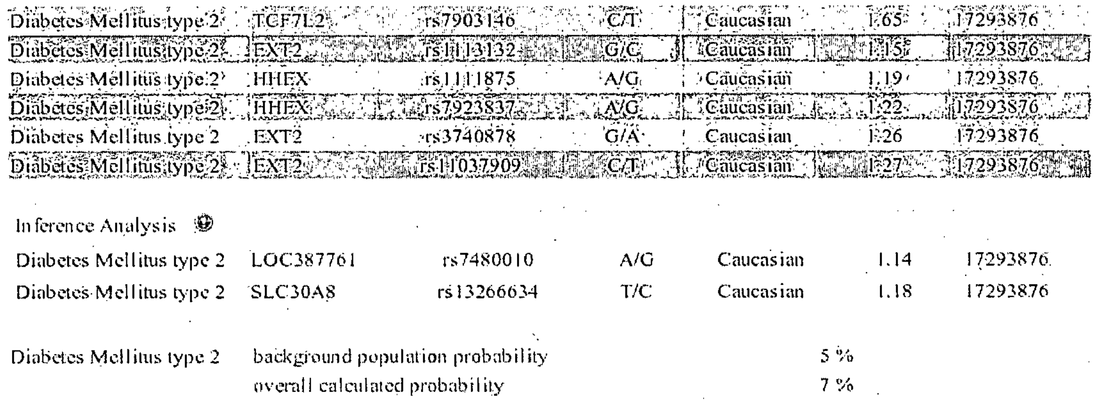
SNP used in probability calculation	SNPs in high linkage disequilibrium
rs11037909	rs11037909 rs1113132 rs3740878

Background probability details

Age of Onset (yrs)	Background probability
45	5%
60	15%

Figure B.10: Probability details for diabetes type 2 is shown here.

If the inference of genotypes configuration option was selected, the output will display SNPs from inference analysis. An example of inferred SNPs and their corresponding details is illustrated in Figures B.11 and B.12.



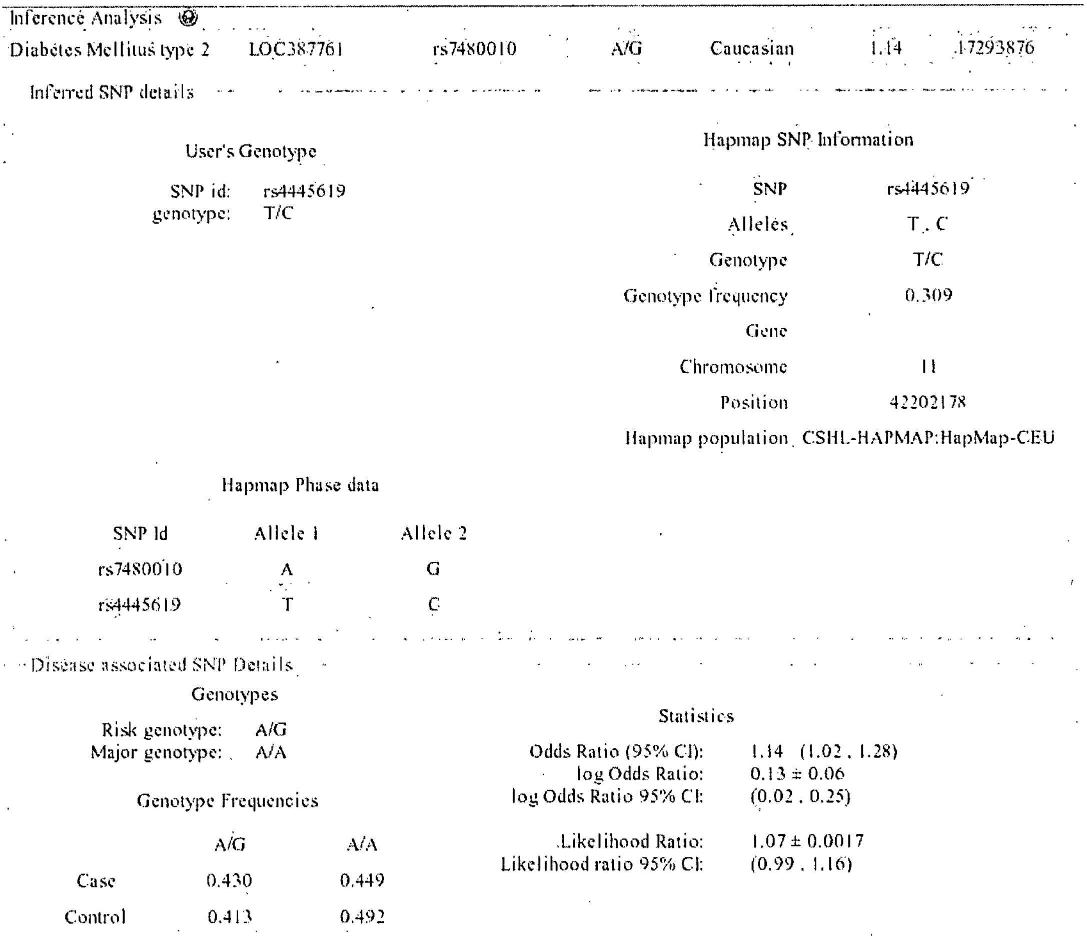
Diabetes Mellitus type 2	TCF7L2	rs7903146	C/T	Caucasian	1.65	17293876
Diabetes Mellitus type 2	EXT2	rs1113132	G/C	Caucasian	1.15	17293876
Diabetes Mellitus type 2	HHEX	rs1111875	A/G	Caucasian	1.19	17293876
Diabetes Mellitus type 2	HHEX	rs7923837	A/G	Caucasian	1.22	17293876
Diabetes Mellitus type 2	EXT2	rs3740878	G/A	Caucasian	1.26	17293876
Diabetes Mellitus type 2	EXT2	rs11037909	C/T	Caucasian	1.27	17293876

Inference Analysis ⊕

Diabetes Mellitus type 2	LOC387761	rs7480010	A/G	Caucasian	1.14	17293876
Diabetes Mellitus type 2	SLC30A8	rs13266634	T/C	Caucasian	1.18	17293876

Diabetes Mellitus type 2	background population probability	5 %
	overall calculated probability	7 %

Figure B.11: SNPs from Inference analysis for Diabetes type 2 are shown.



Inference Analysis

Diabetes Mellitus type 2	LOC387761	rs7480010	A/G	Caucasian	1.14	17293876

Inferred SNP details

User's Genotype

SNP id:	rs4445619
genotype:	T/C

Hapmap SNP Information

SNP	rs4445619
Alleles	T , C
Genotype	T/C
Genotype frequency	0.309
Gene	
Chromosome	11
Position	42202178
Hapmap population	CSHL-HAPMAP:HapMap-CEU

Hapmap Phase data

SNP Id	Allele 1	Allele 2
rs7480010	A	G
rs4445619	T	C

Disease associated SNP Details

Genotypes

Risk genotype:	A/G
Major genotype:	A/A

Genotype Frequencies

	A/G	A/A
Case	0.430	0.449
Control	0.413	0.492

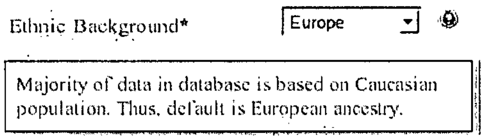
Statistics

Odds Ratio (95% CI):	1.14 (1.02 , 1.28)
log Odds Ratio:	0.13 ± 0.06
log Odds Ratio 95% CI:	(0.02 , 0.25)
Likelihood Ratio:	1.07 ± 0.0017
Likelihood ratio 95% CI:	(0.99 , 1.16)

Figure B.12: Details about the inferred SNPs is shown. The details include the user's genotype, Hapmap data from which inference was performed and the relevant statistics for the disease-associated SNP.

B.2.4 Help Tips

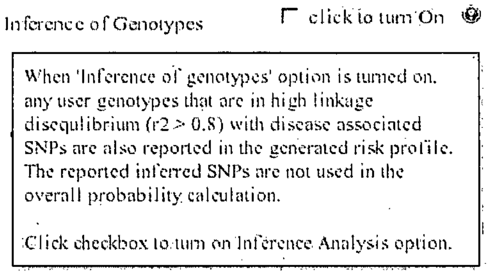
Help tips appear as pop-up on the top right of the page. Whenever a blue question mark icon is displayed, the user can bring the mouse over to the question mark to see the relevant tip. This is done to help guide the user when using D-GRIP. Examples are shown below.



Ethnic Background* Europe

Majority of data in database is based on Caucasian population. Thus, default is European ancestry.

Figure B.13: An example of a ethnic background help tip is shown.



Inference of Genotypes □ click to turn On

When 'Inference of genotypes' option is turned on, any user genotypes that are in high linkage disequlibrium (r2 > 0.8) with disease associated SNPs are also reported in the generated risk profile. The reported inferred SNPs are not used in the overall probability calculation.

Click checkbox to turn on Inference Analysis option.

Figure B.14: An example of inference of genotypes help tip is shown.