D-GRIP: DNA Genetic Risk Information Profile

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

Siddhartha Srivastava

Dr. Wyeth Wasserman
The Center for Molecular Medicine and Therapeutics
University of British Columbia

Terminology

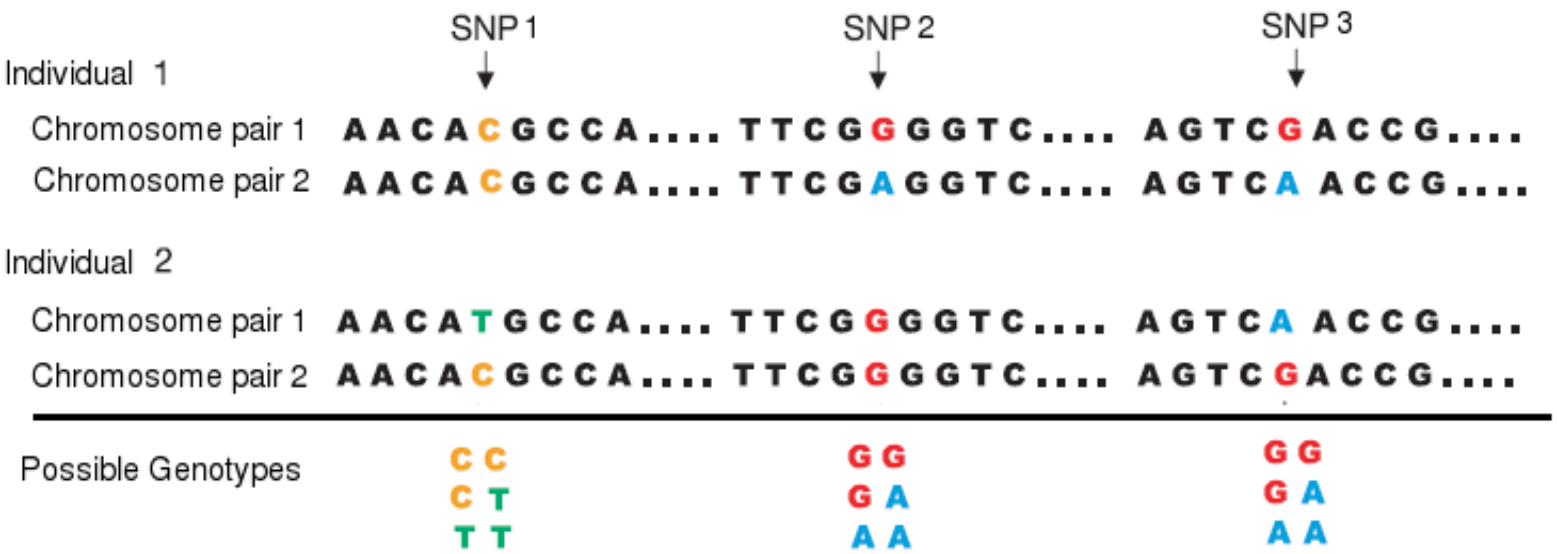
- Genes
- Genetic Variations
- DNA markers
  - SNP are the most common
  - Usually have 2 alleles
  - Genotype

Terminology

- Genes
- Genetic Variations
- DNA markers
  - SNP are the most common
  - Usually have 2 alleles
  - Genotype

Terminology

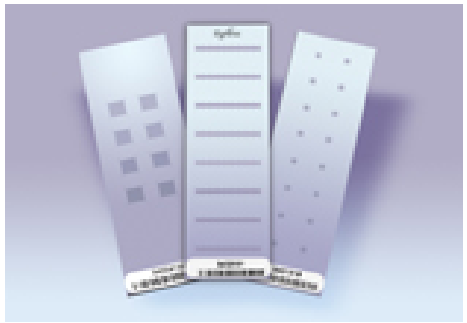
- Genes
- Genetic Variations
- DNA markers
  - SNP are the most common
  - Usually have 2 alleles
  - Genotype



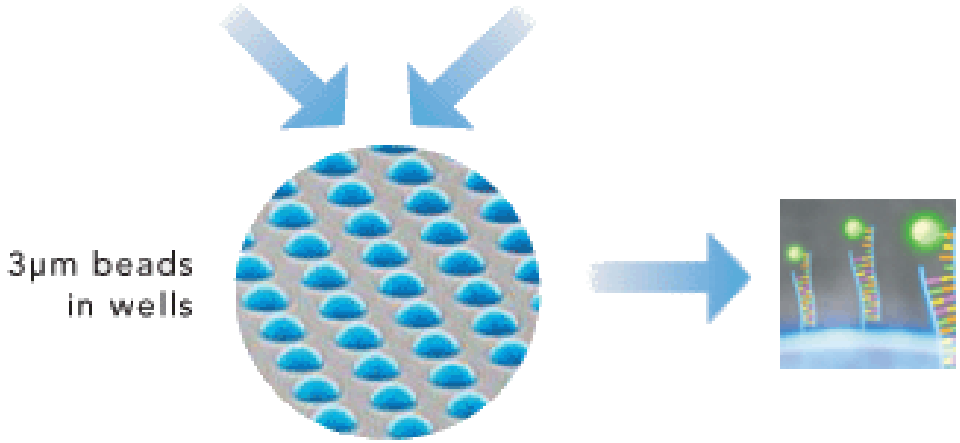
High throughput genotyping

Illumina BeadArray technology

Sentrix 96 Arrav Matrix



Sentrix BeadChips



1

[1] source: Illumina Inc

Genotyping Application

- Illumina technology
  - 500K genotypes
  - Genotyping has become inexpensive
- Genotyping can be used in genetic studies e.g. linkage studies, population diversity

Genotyping Application

- Illumina technology
  - 500K genotypes
  - Genotyping has become inexpensive
- Genotyping can be used in genetic studies e.g. linkage studies, population diversity

Genetics in the News

Main Idea

1. High throughput genotyping technology
   - What can we do with our genotype?

Main Idea

1. High throughput genotyping technology
   - What can we do with our genotype?

Main Idea

1. High throughput genotyping technology
   - What can we do with our genotype?
2. Disease associated genes/ markers
   - What are my risks?

Main Idea

1. High throughput genotyping technology
   - What can we do with our genotype?
2. Disease associated genes/ markers
   - What are my risks?

New Application

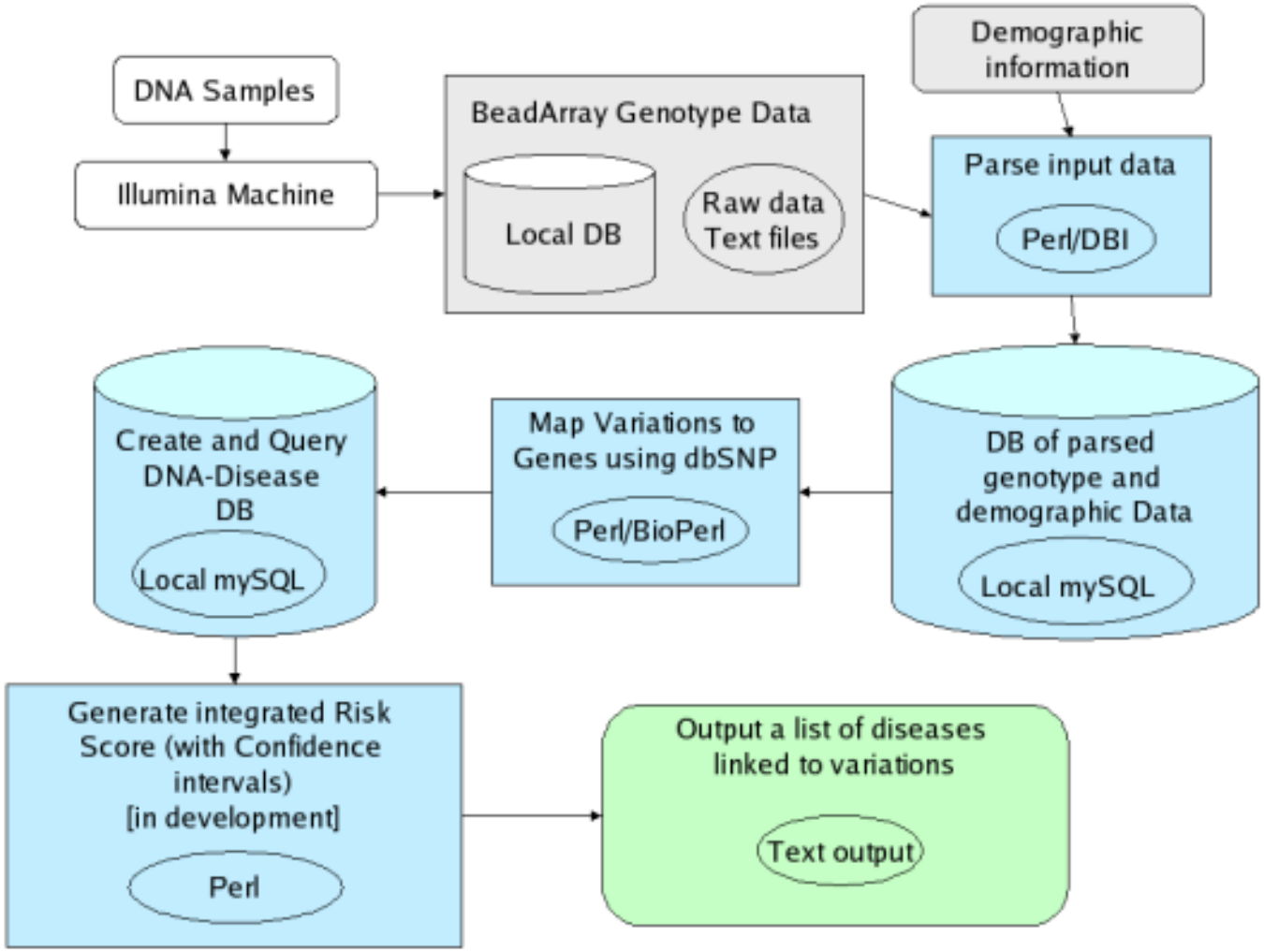
Create a tool to apply genotype data for prediction of diseases.

Project

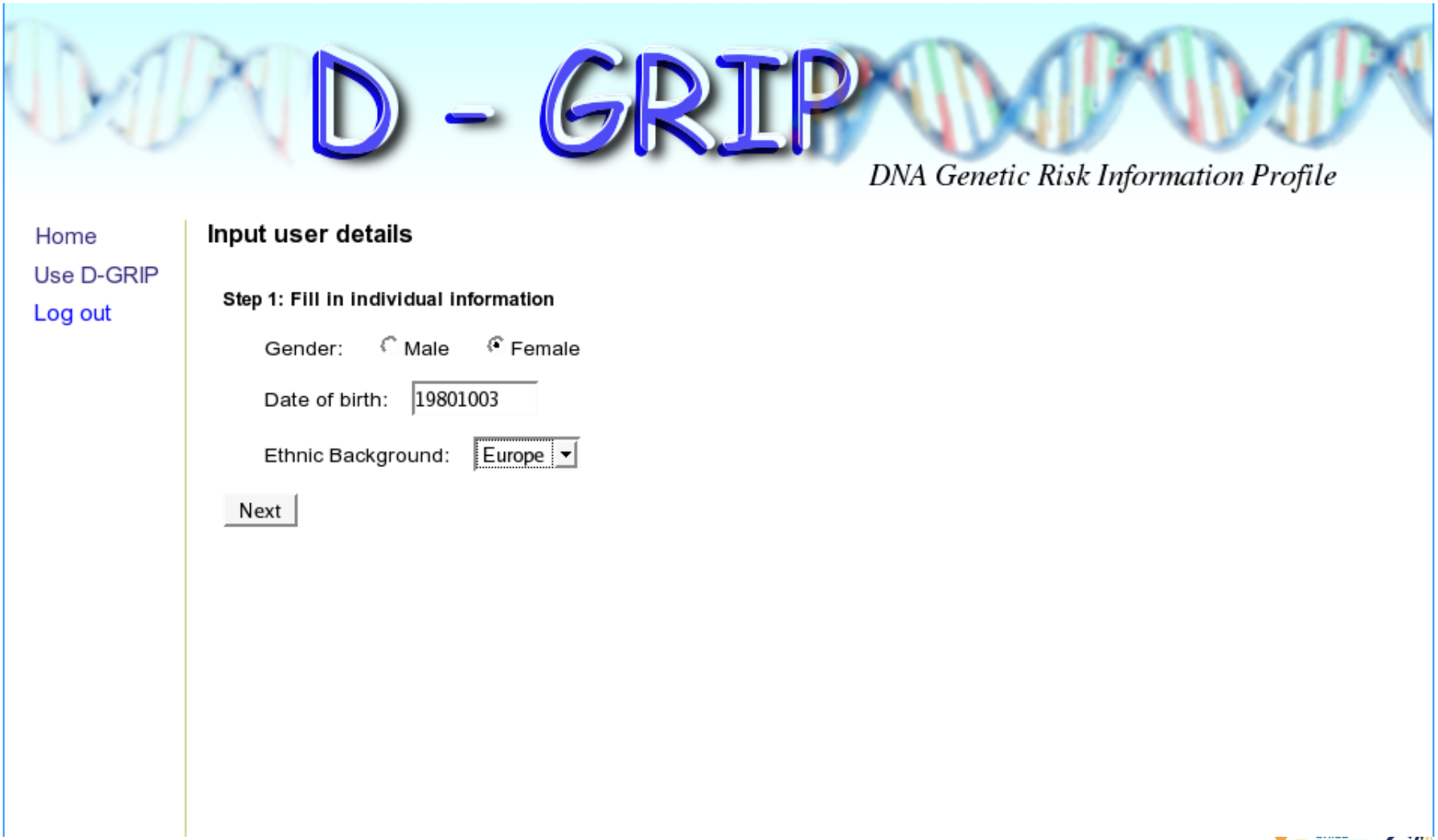
- Create a prototype web-tool that utilizes the genotyping data to predict a risk profiles for individuals
  - Input: Genotype data
  - Output: Prediction of potential diseases with associated risk for an individual
- Access: restricted
- Users: genetic counselors
- Implications

Project

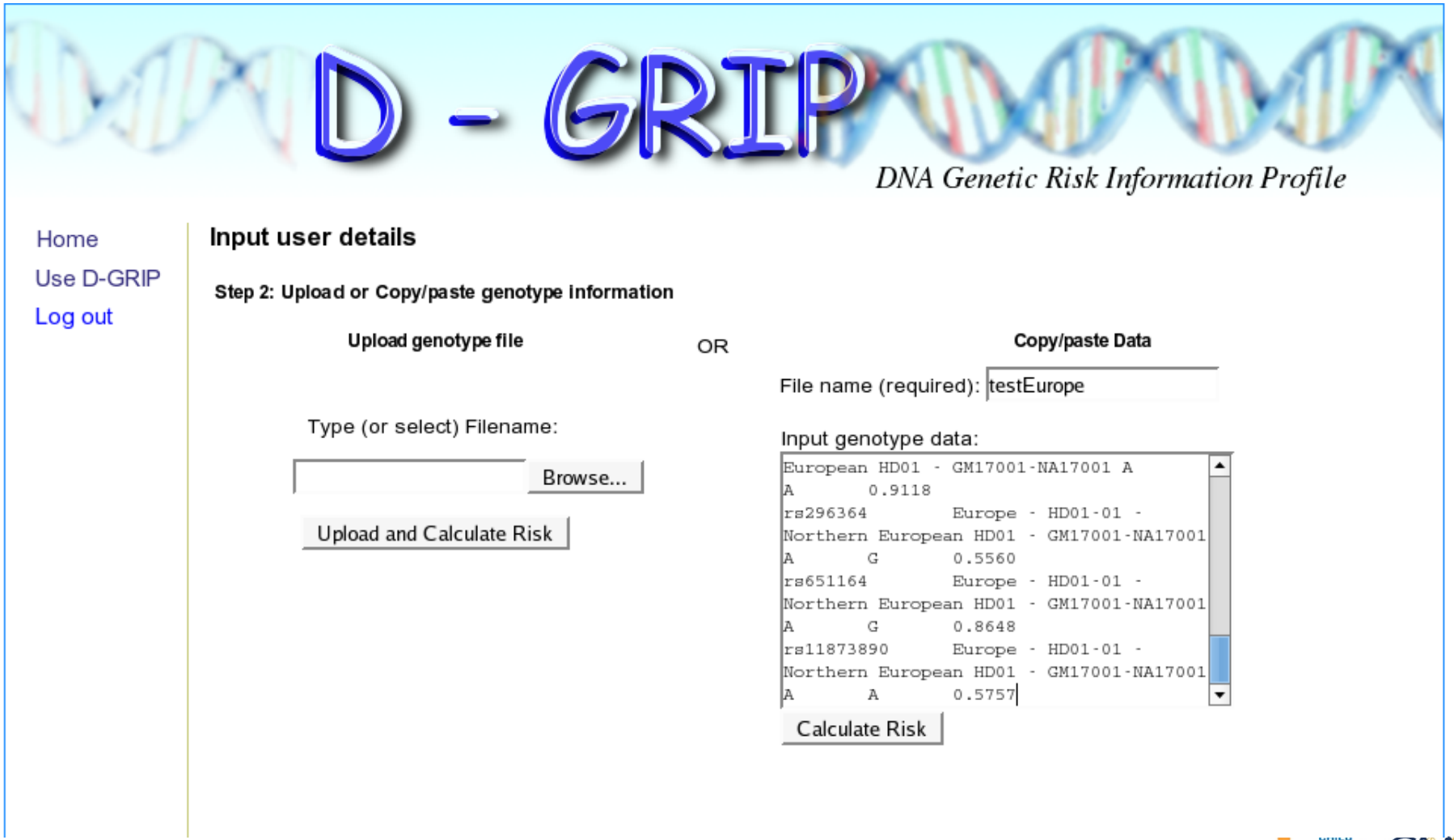
- Create a prototype web-tool that utilizes the genotyping data to predict a risk profiles for individuals
  - Input: Genotype data
  - Output: Prediction of potential diseases with associated risk for an individual
- Access: restricted
- Users: genetic counselors
- Implications


DNA Samples
Illumina Machine
BeadArray Genotype Data
Local DB
Raw data
Text files
Demographic information
Parse input data
Perl/DBI
DB of parsed genotype and demographic Data
Local mySQL
Map Variations to Genes using dbSNP
Perl/BioPerl
Create and Query DNA-Disease DB
Local mySQL
Generate integrated Risk Score (with Confidence intervals) [in development]
Perl
Output a list of diseases linked to variations
Text output

D-GRIP Input


D - GRIP
DNA Genetic Risk Information Profile
Home
Use D-GRIP
Log out
Input user details
Step 1: Fill in individual information
Gender: Male Female
Date of birth: 19801003
Ethnic Background: Europe
Next

D-GRIP Input


D - GRIP
DNA Genetic Risk Information Profile
Home
Use D-GRIP
Log out
Input user details
Step 2: Upload or Copy/paste genotype information
Upload genotype file
OR
Copy/paste Data
Type (or select) Filename:
Browse...
Upload and Calculate Risk
File name (required): testEurope
Input genotype data:
European HD01 - GM17001-NA17001 A
A 0.9118
rs296364 Europe - HD01-01 -
Northern European HD01 - GM17001-NA17001
A G 0.5560
rs651164 Europe - HD01-01 -
Northern European HD01 - GM17001-NA17001
A G 0.8648
rs11873890 Europe - HD01-01 -
Northern European HD01 - GM17001-NA17001
A A 0.5757
Calculate Risk

Structure

1. Parse input data and store in a database
2. Query a DNA-disease database and compare input to known disease associated genotypes
3. Calculate a risk score for each disease

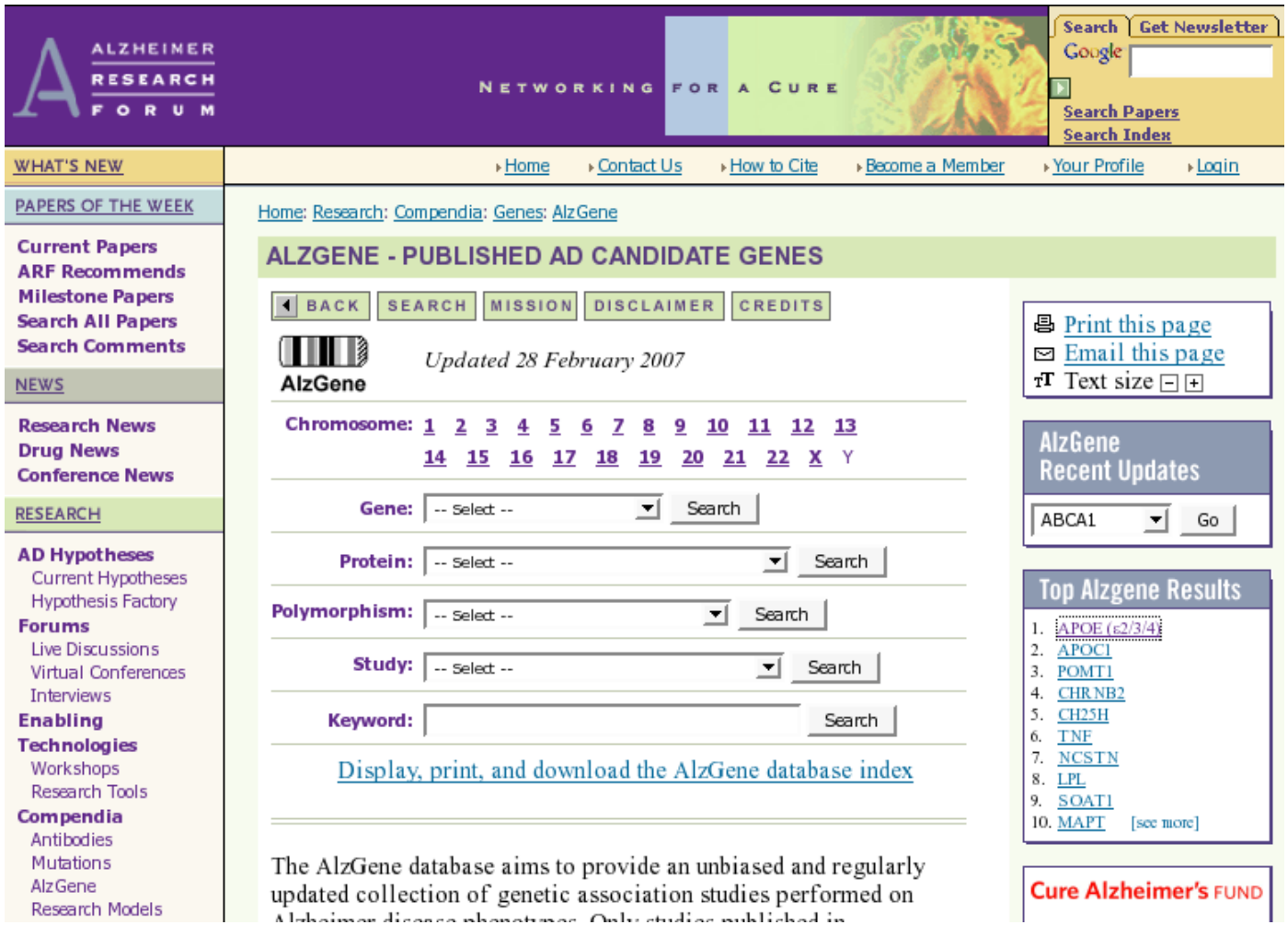
Structure

1. Parse input data and store in a database
2. Query a DNA-disease database and compare input to known disease associated genotypes
3. Calculate a risk score for each disease

What is needed?

- DNA-disease database
- Model for risk calculation

Focus on Alzheimer Disease for Prototype



[2]

[2] Bertram et. al 2007

Focus on Alzheimer Disease for Prototype

META-ANALYSIS OF ALL PUBLISHED AD ASSOCIATION STUDIES (CASE-CONTROL ONLY) apoe ε2/3/4

BACK | SEARCH | MISSION | DISCLAIMER | CREDITS

Polymorphism: apoe ε2/3/4

Gene: APOE (ε2/3/4) (AD2 (ε2/3/4), MGC1571 (ε2/3/4)) Entrez Gene View on PDGene View on SZGene

Protein: apolipoprotein E (ε2/3/4) (Alzheimer disease 2 (ε2/3/4)) Protein

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

Status: Updated 10 December 2006 ; ** See APOE Methods **

VIEW GENE OVERVIEW

Polymorphism: apoe ε2/3/4 Display

1) SUMMARY OF ALLELE AND GENOTYPE DISTRIBUTION OF ALL PUBLISHED ASSOCIATION STUDIES (BY ETHNIC GROUP)

			Alleles			Genotypes					
	# Case-Control Samples		2-Allele (frequency)	3-Allele (frequency)	4-Allele (frequency)	2/2 (frequency)	2/3 (frequency)	3/3 (frequency)	2/4 (frequency)	3/4 (frequency)	4/4 (frequency)
Caucasian	23	AD	0.04	0.58	0.37	7 (0.003)	123 (0.053)	804 (0.346)	65 (0.028)	986 (0.424)	340 (0.146)
		CTR	0.08	0.78	0.14	32 (0.009)	425 (0.119)	2175 (0.609)	66 (0.018)	810 (0.227)	66 (0.018)
Asian	4	AD	0.02	0.70	0.28	0 (0.000)	9 (0.044)	101 (0.490)	1 (0.005)	76 (0.369)	19 (0.092)
		CTR	0.05	0.86	0.09	1 (0.001)	113 (0.085)	986 (0.739)	16 (0.012)	210 (0.157)	9 (0.007)
African descent	2	AD	0.09	0.56	0.35	2 (0.028)	8 (0.111)	23 (0.319)	1 (0.014)	26 (0.361)	12 (0.167)
		CTR	0.05	0.76	0.19	1 (0.009)	10 (0.090)	62 (0.559)	0 (0.000)	34 (0.306)	4 (0.036)

[2]

[2] Bertram et. al 2007

Calculating Risk

- Paper by Yang et. al 2003, *Improving Prediction of Complex disease by testing for multiple Disease-susceptibility genes*
- Created a statistical model:
  1. For a given set of genetic tests G, compute likelihood ratios, $LR(G)$, using logistic regression
  2. Calculate posterior probability of developing disease, given set of genetic tests G and pretest of disease $P(D)$

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

Calculating Risk

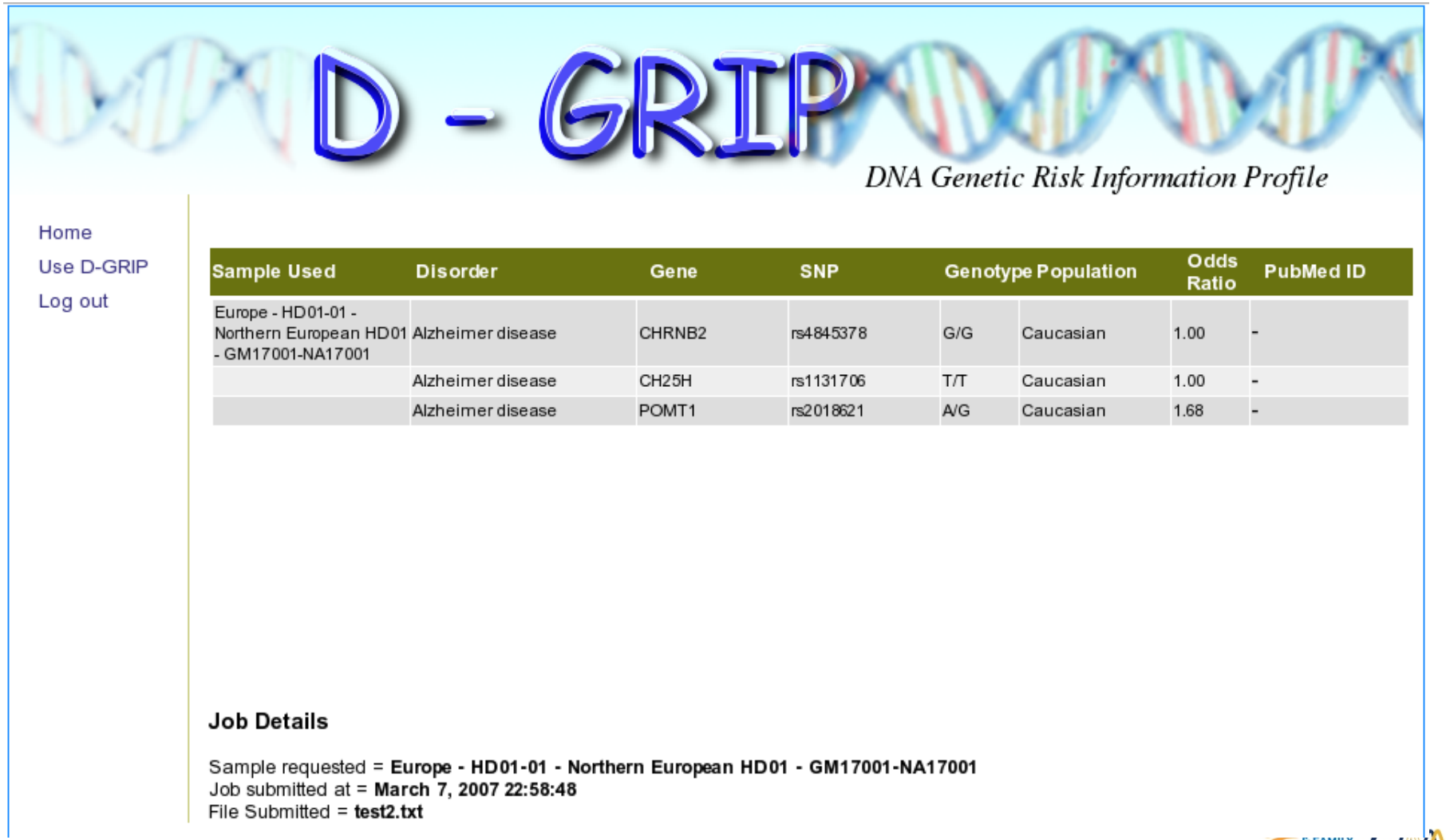
- Paper by Yang et. al 2003, *Improving Prediction of Complex disease by testing for multiple Disease-susceptibility genes*
- Created a statistical model:
  1. For a given set of genetic tests G, compute likelihood ratios, $LR(G)$, using logistic regression
  2. Calculate posterior probability of developing disease, given set of genetic tests G and pretest of disease $P(D)$

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

Advantage

Method permits use of external information.

Output



Sample Used	Disorder	Gene	SNP	Genotype	Population	Odds Ratio	PubMed ID
Europe - HD01-01 - Northern European HD01 - GM17001-NA17001	Alzheimer disease	CHRNB2	rs4845378	G/G	Caucasian	1.00	-
	Alzheimer disease	CH25H	rs1131706	T/T	Caucasian	1.00	-
	Alzheimer disease	POMT1	rs2018621	A/G	Caucasian	1.68	-

Job Details

Sample requested = **Europe - HD01-01 - Northern European HD01 - GM17001-NA17001**
Job submitted at = **March 7, 2007 22:58:48**
File Submitted = **test2.txt**

Issues and Observations

- Number of disease entries
- Manual mining of Literature for other diseases

Issues and Observations

- Number of disease entries
- Manual mining of Literature for other diseases

Observation

Lack of organized information on genotype to phenotype

More work

- Refine risk model
- Incorporate haplotype data
- More entries into disease database
- Features to web interface
- Testing and feedback from users

Acknowledgments

- Wyeth Wasserman
- Wasserman Lab
- CMMT
- BITP

References I

- Bertram L, McQueen MB, Mullin K, Blacker D, Tanzi RE. (2007) *Systematic meta-analyses of Alzheimer disease genetic association studies: the AlzGene database.* Nat Genet 39(1): 17-23.
- Fiddy S, Cattermol D, Xie D, Duan X.Y and Mott R. 2006. *An integrated system for genetic analysis*. BMC Bioinformatics 7:210.
- GeneSage. `http://www.genesage.com`. Website, accessed July 2006.
- GeneTrack. `http://www.genetrack.bc.ca`. Website, accessed July 2006.
- International Hapmap Consortium. 2003. *The International HapMap Project*. Nature. USA. 426, 789-796.

References II

- Patrinos G.P. and Brookes A.J, 2005. *DNA, diseases and databases: disastrously deficient*. Science Direct, Trends in Genetics. 21(6): 333-338.
- Shah N, Teplitsky M.V, Minovitsky S, Pennacchio L.A, Hugenholtz P, Hanmann B and Dubchak I.L. 2005. *SNP-VISTA: An interactive SNP visualization tool*. BMC Bioinformatics. 6:292.
- Zhao L, Li M, Guo Y, Xu F, Li J and Deng H. 2004. *SNPP: automating large-scale SNP genotype data management*. Bioinformatics 21(2): 266-68.
- Yang Q, Khoury M.J., Botto L, Friedman J.M. and Flanders D.W. 2003. *Improving the Prediction of Complex Diseases by Testing for Multiple Disease-Susceptibility Genes*