Predicting a phenotype from DNA using statistics

VINCENT GUILLEMOT

BUT 99 % OF THE COURSE MATERIAL WAS WRITTEN BY EDITH LE FLOCH
Single Nucleotide Polymorphisms (SNPs) 1/3

**SNP**: position on the genome where a single nucleotide varies in the population (>1% of individuals) → due to an ancestral mutation

**Main form of DNA variability** in the population

About **30 millions** SNPs
Single Nucleotide Polymorphisms (SNPs) 2/3

Usually only **two possible alleles for a SNP**: one major (e.g. A) and one minor (e.g. G)

The **genotype** of an individual is defined by considering the pair of homologous chromosomes: 3 possibilities (e.g. AA, AG or GG)

→ often coded as the number of minor alleles:

0 (AA), 1 (AG) or 2 (GG)
Single Nucleotide Polymorphisms (SNPs) 3/3

SNPs may be located:
- inside a gene:
  - in an exon (coding for the protein)
  - in an intron (non-coding)
- outside a gene
Linkage Disequilibrium (LD) 1/2

**LD:** non-random association of alleles between two nearby SNPs

→ due to physical linkage (ie SNPs on the same chromosome)

→ The 2 SNPs transmitted together through generations

Recombination between homologous chromosomes during meiosis

→ Probability of recombination increases (and thus LD decreases) with the distance between the 2 SNPs
Linkage Disequilibrium (LD) 2/2

Non-homogeneous recombination between homologous chromosomes during meiosis: hot spots of recombination

→ LD blocks
Characteristics of Genome Wide Association Studies

A lot of data

Sometimes very rare events

High *correlation:*
- between variables,
- between individuals

→ A lot of caution
The univariate approach

$Y = F(G)$
Classical statistical analysis of genotyping data

Univariate approach:

Test for the association of each SNP (with the phenotype) independently

If the phenotype is disease (case) / not disease (control):
Is the distribution of the 3 genotypes the same for cases and controls?

For example for a SNP with two possible alleles A and T:

<table>
<thead>
<tr>
<th></th>
<th>Cases</th>
<th>Controls</th>
</tr>
</thead>
<tbody>
<tr>
<td>AA</td>
<td>20</td>
<td>10</td>
</tr>
<tr>
<td>AT</td>
<td>40</td>
<td>66</td>
</tr>
<tr>
<td>TT</td>
<td>75</td>
<td>163</td>
</tr>
</tbody>
</table>

→ p-value of a **Chi-squared test** = 0.007 but many tests (= nb SNPs) !!

→ Need to correct for **multiple comparisons**
Case/control study 1/2

Genotypic model:

<table>
<thead>
<tr>
<th></th>
<th>Cases</th>
<th>Controls</th>
</tr>
</thead>
<tbody>
<tr>
<td>AA</td>
<td>6</td>
<td>10</td>
</tr>
<tr>
<td>AT</td>
<td>8</td>
<td>66</td>
</tr>
<tr>
<td>TT</td>
<td>75</td>
<td>163</td>
</tr>
</tbody>
</table>

Dominant model (only 1 allele needed for the disease):

<table>
<thead>
<tr>
<th></th>
<th>Cases</th>
<th>Controls</th>
</tr>
</thead>
<tbody>
<tr>
<td>AA+AT</td>
<td>140</td>
<td>10</td>
</tr>
<tr>
<td>TT</td>
<td>20</td>
<td>163</td>
</tr>
</tbody>
</table>

Recessive model (2 alleles needed for the disease):

<table>
<thead>
<tr>
<th></th>
<th>Cases</th>
<th>Controls</th>
</tr>
</thead>
<tbody>
<tr>
<td>AA</td>
<td>60</td>
<td>10</td>
</tr>
<tr>
<td>AT + TT</td>
<td>10</td>
<td>229</td>
</tr>
</tbody>
</table>
Case/control study 2/2

Genotypic test:
- The most general
- Less powerful on average to detect moderate associations

<table>
<thead>
<tr>
<th></th>
<th>Cases</th>
<th>Controls</th>
</tr>
</thead>
<tbody>
<tr>
<td>AA</td>
<td>20</td>
<td>10</td>
</tr>
<tr>
<td>AT</td>
<td>40</td>
<td>66</td>
</tr>
<tr>
<td>TT</td>
<td>75</td>
<td>163</td>
</tr>
</tbody>
</table>

Allelic test:
- Assumes the 2 alleles of an individual are independent (Hardy-Weinberg)
- Intermediate between dominant and recessive
- Works in most cases

<table>
<thead>
<tr>
<th></th>
<th>Cases</th>
<th>Controls</th>
</tr>
</thead>
<tbody>
<tr>
<td>Minor allele A</td>
<td>20*2+40=80</td>
<td>10*2+66=86</td>
</tr>
<tr>
<td>Major allele T</td>
<td>75*2+40=190</td>
<td>163*2+66=392</td>
</tr>
</tbody>
</table>
If the phenotype \((y)\) is quantitative \(\rightarrow\) simple linear regression

\[
y = \beta_0 + \beta \times \#\text{minor alleles}
\]

\(\beta \neq 0\ ? \ (T\text{-test})\)

Like for case/control studies, an additive model is usually used
Quantitative phenotype 2/4

One may also use a dominant model → **simple linear regression with a binary variable**

\[ y = \beta_0 + \beta \times (G \neq AA) \]

\[ \beta \neq 0 \quad (T\text{-test}) \]
Quantitative phenotype 3/4

Or a recessive model → simple linear regression with a binary variable

\[ y = \beta_0 + \beta \times (G = aa) \]

\[ \beta \neq 0 \ ? \ (T\text{-test}) \]
Quantitative phenotype 4/4

Or a genotypic model $\rightarrow$ **multiple linear regression with two binary variables**

$$y = \beta_0 + \beta_{Aa} \times (G == Aa) + \beta_{aa} \times (G == aa)$$

It’s actually bivariate

$\beta_{Aa} \neq 0$ and/or $\beta_{aa} \neq 0$? (F-test)
Multiple comparison with univariate tests

Such analyses performed on SNPs covering the whole genome are called **Genome Wide Association Studies (GWAS)**

Whatever the type of the test, one still needs to **correct for multiple comparisons**...

The test will be more powerful to detect an association:

- with **high sample size** (often 10s of 1000s of individuals)
- with **frequent polymorphisms**
- with **strong effects**
P-value correction in Genome-Wide Association Studies (GWAS)

“Manhattan plot” of the p-values of the SNPs along the genome:

A *Bonferroni-like correction* is commonly used to correct for multiple tests: → genome-wide significance threshold = 5*10^-8

Before correction: 0.05, After correction: 0.00000005
Limitations of univariate analysis of Whole Genome Sequencing data

Univariate GWAS methods (linear regression, chi-square) may be applied BUT:
- much more multiple comparisons (tens of millions)
- very low power to detect association for rare variants

More individuals needed but expensive (>1000$ per individual)

Statistical methods need to be adapted by collapsing nearby variants: Region-based analysis (multivariate approach) → stronger signals and fewer tests (20000-25000 genes)
The multivariate approach

\[ Y = F(G_1, G_2, X_1, \ldots) \]
Multivariate methods

Can be used to:

• use different genetic models,
• take into account the effect of confounding factors,
• uncover hidden effects,
• etc. (detect interactions, Machine Learning etc.)
Population structure within Europe

Principal component analysis on 197,146 SNPs (coded 0, 1 or 2) in 1387 individuals

→ when plotting the two first principal components, the map of Europe appears!

Problems with population structure in GWAS 1/2

case/control study: If the % of each population different in cases and controls

→ Alleles specific to Pop. 1 artificially associated with the disease!
Problems with population structure in GWAS 2/2

quantitative trait: If the 2 populations have different means (due to a few population-specific SNPs or to different lifestyles)

→ Alleles specific to 1 population artificially associated with the trait!
Correcting for population structure in GWAS

**Principal Component Analysis** on the matrix of genotypes: $X$ (nb of individuals * nb of SNPs) with genotypes coded 0, 1 or 2

**Top principal components used as covariates** in the linear regression model to predict phenotype $Y$ from SNP $X_j$

Other confounding variables (age, sex) often used as covariates too
Multivariate methods

In this example:

Impossible to separate the two classes with one variable (dimension) only

But possible by taking in account the two variables together → multivariate method

Two types of multivariate methods:

Linear model (Additive effects)
Non-linear approach
Other analyses
Heritability ($h^2$)

$\sigma_T^2$: Total Variance of the phenotype
$\sigma_g^2$: Variance due to additive genetics
$\sigma_r^2$: Unexplained residual variance

Then, heritability is defined as $h^2 = \frac{\sigma_g^2}{\sigma_T^2}$

Test
- $H_0$: $\sigma_T^2 = \sigma_r^2$ (Likelihood: $L_1$)
- $H_1$: $\sigma_T^2 = \sigma_r^2 + \sigma_g^2$ (Likelihood: $L_0$)
- Statistic $= 2 \times \log(L_1/L_0)$
  
  *which follows $\chi^2(k)$, with $k$ the difference in number of parameters between $L_0$ & $L_1$*
Global results of GWAS on diseases
Analyze genetic data with multiple Ys

Partial Least Squares Correspondence Analysis: A Framework to Simultaneously Analyze Behavioral and Genetic Data

Derek Beaton and Joseph Dunlop
The University of Texas at Dallas

Alzheimer’s Disease Neuroimaging Initiative

Hervé Abdi
The University of Texas at Dallas

For nearly a century, detecting the genetic contributions to cognitive and behavioral phenomena has been a core interest for psychological research. Recently, this interest has been reinvigorated by the availability of genotyping technologies (e.g., microarrays) that provide new genetic data, such as single nucleotide polymorphisms (SNPs). These SNPs—which represent pairs of nucleotide letters (e.g., AA, AG, or GG) found at specific positions on human chromosomes—are best considered as categorical variables, but this coding scheme can make difficult the multivariate analysis of their relationships with behavioral measurements, because most multivariate techniques developed for the analysis between sets of variables are designed for quantitative variables. To palliate this problem, we present a generalization of partial least squares—a technique used to extract the information common to 2 different data tables measured on the same observations—called partial least squares correspondence analysis—that is specifically tailored for the analysis of categorical and mixed (“heterogeneous”) data types. Here, we formally define and illustrate—in a tutorial format—how partial least squares correspondence analysis extends to various types of data and design problems that are particularly relevant for psychological research that include genetic data. We illustrate partial least squares correspondence analysis with genetic, behavioral, and neuroimaging data from the Alzheimer’s Disease Neuroimaging Initiative. R code is available on the Comprehensive R Archive Network and via the authors’ websites.
Machine Learning in Genome-Wide Association Studies

Silke Szymczak, Joanna M. Biernacka, Heather J. Cordell, Oscar González-Recio, Inke R. König, Heping Zhang, and Yan V. Sun

1Institut für Medizinische Biometrie und Statistik, Universität zu Lübeck, Lübeck, Germany
2Department of Health Sciences Research, Mayo Clinic, Rochester, Minnesota
3Institute of Human Genetics, International Centre for Life, Newcastle University, Central Parkway, Newcastle upon Tyne, United Kingdom
4Department of Dairy Science, University of Wisconsin-Madison, Madison, Wisconsin
5Department of Epidemiology and Public Health, Yale University School of Medicine, New Haven, Connecticut
6Department of Epidemiology, School of Public Health, University of Michigan, Ann Arbor, Michigan

Recently, genome-wide association studies have substantially expanded our knowledge about genetic variants that influence the susceptibility to complex diseases. Although standard statistical tests for each single-nucleotide polymorphism (SNP) separately are able to capture main genetic effects, different approaches are necessary to identify SNPs that influence disease risk jointly or in complex interactions. Experimental and simulated genome-wide SNP data provided by the Genetic Analysis Workshop 16 afforded an opportunity to analyze the applicability and benefit of several machine learning methods. Penalized regression, ensemble methods, and network analyses resulted in several new findings while known and simulated genetic risk variants were also identified. In conclusion, machine learning approaches are promising complements to standard single-and multi-SNP analysis methods for understanding the overall genetic architecture of complex human diseases. However, because they are not optimized for genome-wide SNP data, improved implementations and new variable selection procedures are required. Genet. Epidemiol. 33 (Suppl. 1):S51–S57, 2009. © 2009 Wiley-Liss, Inc.

Key words: Genetic Analysis Workshop; data mining; penalized regression; random forests; network analysis