Searching for missing heritability using univariate and multivariate approaches on both genotyping and sequencing data

Edith LE FLOCH[1] lefloch@cng.fr
Vincent Meyer[2], Lilia Mesrob[2], François Artiguenave[2], Jean-François Deleuze[1]
[1]CEA/DSV/IG/CNG, [2]CEA/DSV/IG/CNG/LBI

To investigate the missing heritability of Alzheimer's Disease, we used public data of 809 individuals from the Alzheimer's Disease Neuroimaging Initiative with both genotyping and whole genome sequencing data. We first performed a GWAS on both types of data, testing the association between each SNP/rare variant and the phenotype. We found significant associations for 16 common SNPs in Linkage Disequilibrium in the region of the well-known APOE gene, after Bonferroni correction. Gene-based approaches applied on sequencing data (SKAT, burden test) could only confirm the association of APOE driven by common SNPs. This suggests that such approaches have low power with very rare causal variants. Finally, multivariate methods (sparse linear regression, decision tree, SNP heritability) failed to predict the case/control status from genotyping data but these results could be greatly improved on a much larger dataset.

Context

Alzheimer's Disease:
- A neurodegenerative disease associated with cognitive disorders and memory loss
- The first type of dementia (60-70% of cases)
- Prevalence: 20% in people over 80 and 40% in people over 95

Some genetic origins:
- Polygenic disease assumed to be explained at 75% by genetic factors
- But the known causal genes account only for 8% (main gene APOE accounts for 6%) → **Huge missing heritability!**
- **Many possible reasons** for missing heritability:
  - rare variants
  - interaction effects between variants or variant × environment
  - many small effects that cannot be detected with current sample sizes

Dataset and objectives

Dataset:
- **Public data on 809 unrelated individuals:**
  - 188 patients with Alzheimer's Disease (AD)
  - 393 patients with Mild Cognitive Impairment (MCI)
  - 228 controls
- **Genotyping data (2.5M SNPs) and Whole Genome Sequencing Data** available
- **Brain imaging data** (MRI) also available

Objectives:
- Look for **missing heritability with rare variants** (sequencing data)
- Look for **missing heritability with multivariate approaches**

1) The univariate approach

On genotyping data:
- Comparison of Alzheimer patients (188) versus others (621) for the 2.5 million SNPs

- **3 variants with a significant p-value** after Bonferroni correction ($p<2*10^{-8}$) with χ^2 test or Fisher exact test
  - 1 in APOE intron and in regulatory region ($p=1.4*10^{-12}$)
  - 2 in intergenic regions near APOE ($p=3*10^{-13}$ and $p=6*10^{-14}$)

- **Associated variants are frequent** (Minor Allele Freq. 20-40%)

On Whole Genome Sequencing data:
- ~4 millions SNVs per individual w.r.t. the reference genome
  → ~**60 millions variants** for all individuals

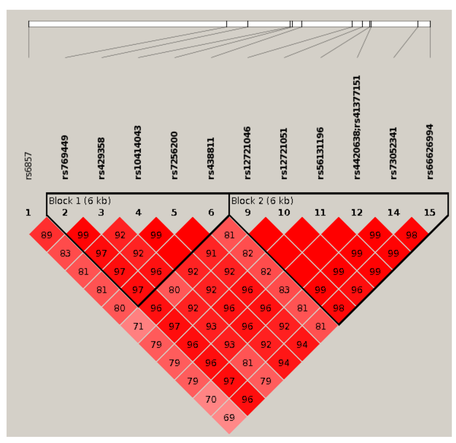
- **63% of the SNVs with good quality** → ~40 millions variants

- **46% of good quality variants are specific to only 1 individual**: Each individual carries between 20000 and 25000 unique variants that are not present in other individuals (<1% of the variants of the individual)

- **16 frequent SNPs with a significant** p-value after Bonferroni correction (χ^2 test/Fisher's exact test) : $10^{-18} < p < 10^{-9}$ **in the APOE region (36kb)**

- **Top associated SNP:**
  rs429358 (non-synonymous) : $p=5*10^{-18}$
  (one of the 2 SNPs of **APOε4 allele**)

- Associated SNPs are in **high linkage disequilibrium**
  → probably one single causal association



2) The multivariate approach

On sequencing data:
- **Multiple linear regression model per gene:**
  - p variants in **a given gene**
  - Genotypes of individual i : $\boldsymbol{X}_i\ (1 \times p)$, coded 0, 1 or 2
  - Covariates of individual i : $\boldsymbol{Z}_i\ (1 \times k)$ such as age, sex, pop. structure
  - Phenotype of individual i : Y_i

  For a **case/control** (1/0) phenotype :
  $$logit(p_i) = ln\left(\frac{p_i}{1-p_i}\right) = \alpha_0 + \boldsymbol{Z}_i\boldsymbol{\alpha} + \boldsymbol{X}_i\boldsymbol{\beta} = \alpha_0 + \sum_{j=1}^{k} Z_{ij}\alpha_j + \sum_{j=1}^{p} X_{ij}\beta_j$$
  with $p_i = P(Y_i = 1 | \boldsymbol{X}_i, \boldsymbol{Z}_i)$
  Test of no additive effect of the region: H_0: $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^T = \boldsymbol{0}$

- Two ways to improve power when p is large:
  - **Burden test [1]**: Suppose $\beta_1 = \cdots = \beta_p = \beta \rightarrow logit(p_i) = \alpha_0 + \boldsymbol{Z}_i\boldsymbol{\alpha} + C_i\beta$
    Where $C_i = \sum_{j=1}^{p} w_j X_{ij}$: **genetic burden ("meta-variant")**
    and w_j is an optional weight for variant j (higher for rare variants)

  - **Sequence Kernel Association Test [2] assumes random effects:**
    $\beta_j \sim distribution\ (0, w_j^2\,\tau)$ where w_j^2 is an optional weight for variant j

- **Global SKAT results:**
  - **On exons only: 1 gene almost significant** after correction ($p=2*10^{-5}$): **SORBS3** (already associated with AD)
  - **With all candidate genes** (exons) all together**: no significant results** (p=0.29)

- **Specific results on APOE gene:**
  - SKAT with weights: $p=10^{-3}$
  - SKAT without weights: **$p=1.2*10^{-19}$ → driven by common SNPs**
  - Burden test with weights: p=0.035
  - Burden test without weights : **$p=7.9*10^{-12}$ → Burden tests less adapted to few associated SNPs**

On genotyping data:
- **Similar model to SKAT** applied on 2.5M SNPs to estimate heritability: Heritability of 10% but high variance
- **Multivariate approaches** (Sparse logistic regression, SVM, Adaboost, Random Forests) failed to predict in a reproducible way

Accuracy AD	Accuracy Controls	Global accuracy
6%	100%	63%

Conclusions and Perspectives

- At the gene/SNP level, **a few significant associations and mainly on common SNPs** (no great improvement with sequencing data and rare variants yet)
- At the whole genome level, on-going tests show **promising results with multivariate algorithms on genotyping data,** suggesting cummulative effects of many SNPs
- We **need much larger samples!!** A lot are coming…

1. Morris, A.P. and Zeggini, E. (2010) **An evaluation of statistical approaches to rare variant analysis in genetic association studies**. *Genetic Epidemiology* doi:10.1002/gepi.20450.
2. Wu, M. C.*, Lee, S.*, Cai, T., Li, Y., Boehnke, M., and Lin, X. (2011) **Rare Variant Association Testing for Sequencing Data Using the Sequence Kernel Association Test (SKAT).** *American Journal of Human Genetics* , 89, 82-93.