# association studies

**Mathieu Emily, Leif Schauser, Thomas Mailund and Mikkel H. Schierup**
*Bioinformatics Research Center (BiRC) – Aarhus University – Denmark*
*memily@daimi.au.dk*

## Introduction

- Over the last few years, technological improvements have made possible the genotyping of hundreds of thousands of SNPs, enabling whole-genome association studies. Although increasing evidence suggests that interaction between loci should be considered, most of these studies proceed by considering each SNP independently. One reason for this choice comes from the dramatic number of tests (~ 50 billions of tests), requiring strong multiple testing correction.
- In this work, a feasible and powerful approach is proposed to drive search by biological knowledge. We focus on SNPs that belong to genes or proteins known to interact in some biological network. Although some interactions might be missed, these pairs are good candidates for epistasis.

## Interaction network

### Method

- We consider pairs of proteins known to interact. The interactions include direct (physical) and indirect (functional) associations; they are derived from different sources (Genomic context, High-throughput experiments, etc…) (See Figure 1).
- Each pair of SNPs within a protein-protein interaction is tested for association with the disease (See Figure 2).

### Statistical procedures

#### SNPs-Association test

- χ2 test with 8 degrees of freedom (9 possible genotypes and 2 possible phenotypes)
- p-value is denoted by $p_{SNP-SNP}$.

#### Proteins-Association test

- A Simes correction is applied to account for the correlation between SNP pairs in a single protein pair.
- p-value is denoted by $p_{Prot-Prot}$.

$$p_{prot-prot} = \min_{1 \le i \le N} \left( p^{(i)}_{SNP-SNP} * N/i \right) \text{ where } p^{(i)}_{SNP-SNP} \text{ are the sorted p - values}$$

#### Test for interaction against marginal effects

- A F-test is performed to detect SNP marginal effect. We consider two models:
  - Mod_Marg: Logistic regression with 2 covariates (SNP1 and SNP2)
  - Mod_Inter: Logistic regression with 3 covariates (SNP1, SNP2 and SNP1:SNP2)

### Data

- Parkinson dataset [2] is composed of 271 cases and 270 controls. 396,613 unique SNPs were used from the Illumina Infinium I and HumanHap300 assays (More than 78 billions of SNP pairs).
- Two networks have been studied: the STRING database [1] and the Epidermial Growth Factor Receptor pathway [3].

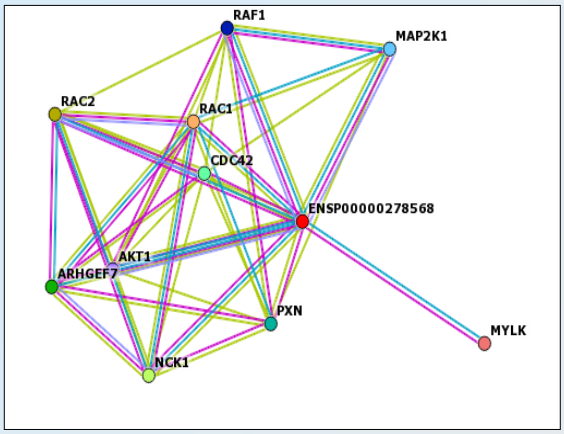


**Fig1.** ENSP00000278568 - Kinase PAK's interaction network in the STRING database [1]

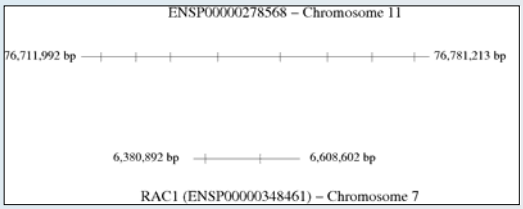


**Fig2.** Example of a Protein-Protein interaction between ENSP0000278568 and RAC1 (ENSP00000348461). The first protein includes 8 SNPs of the Illumina Chip and the second one includes 2 SNPs. 16 SNP pairs are tested for this interaction

## STRING database

- Number of proteins: 9,814
- Number of Interactions: 83,756
- Number of SNP Pairs tested: 8,726,558

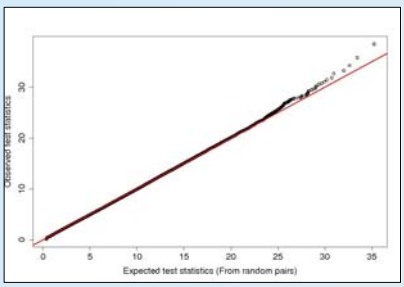


**Fig3.** Quantile-Quantile plots of SNPs-test statistics. The observed statistics are from the STRING database SNP pairs and the expected statistics are from pairs randomly choosen in the all Parkison dataset.

| Pairs of SNPs | SNP1 $\chi^2(2)$ test | SNP2 $\chi^2(2)$ test | SNPs-Association test | Interaction test |
|---|---|---|---|---|
| Chr4-Chr9<br>SNP1: rs2866413<br>SNP2: rs1009305 | 0.007 | 0.582 | **$6.18 \times 10^{-6}$** | **$1.27 \times 10^{-5}$** |
| Chr15-Chr15<br>SNP1: rs804282<br>SNP2: rs1603785 | 0.08 | 0.02 | **$9.22 \times 10^{-6}$** | **$6.5 \times 10^{-5}$** |

**Tab1.** Two most associated pairs of SNPs in the STRING database.

| Pairs of Protein | P1 Marginal test | P2 Marginal test | SNPs-Association test |
|---|---|---|---|
| Chr6-Chr5<br>P1: ENSP00000310144<br>P2: ENSP00000338208 | 0.07 | 0.004 | **$8.33 \times 10^{-5}$** |
| Chr11-Chr15<br>P1: ENSP00000310040<br>P2: ENSP00000311430 | 0.12 | 0.80 | **$8.832 \times 10^{-6}$** |

**Tab2.** Two most associated pairs of Genes in the STRING database.

## Epidermial Growth Factor Receptor (EGFR1)

- Number of molecules: 177
- Number of Interactions: 221
- Number of SNP Pairs tested: 44,090

| Pairs of SNPs | SNP1 $\chi^2(2)$ test | SNP2 $\chi^2(2)$ test | SNPs-Association test | Interaction test |
|---|---|---|---|---|
| Chr7-Chr16<br>SNP1: rs7809332<br>SNP2: rs3922849 | 0.104 | $2.06 \times 10^{-4}$ | **$5.61 \times 10^{-5}$** | **0.031** |
| Chr22-Chr12<br>SNP1: rs804282<br>SNP2: rs1603785 | 0.58 | $1.31 \times 10^{-4}$ | **$6.59 \times 10^{-5}$** | **0.009** |

**Tab3.** Two most associated pairs of SNPs in the EGFR1 pathway.

## Conclusion

- The proposed method is an alternative to techniques based on marginal effects:
  - SNPs association statistics deviate from random pair statistics (see Figure 3)
  - Most associated pairs show real interaction and not only marginal effect (see Tables 1-3)
- Perspectives:
  - Protein association test may be improved by using haplotype-based method (Blossoc [4])
  - The approach will gain by considering cohorts with 2,000 cases and 2,000 controls.

## References

[1]. Von Mering et al. STRING: known and predicted protein-protein associations, integrated and transferred across organisms, *Nucl. Ac. Res.*, 33 (2005)
[2]. Fung et al. Genome-wide genotyping in Parkinson's disease and neurologically normal controls: first stage analysis and public release of data, *Lancet Neurol*, 5 (2006)
[3]. Bader et al. PathGuide: a pathway resource list, *Nucl. Ac. Res.*, 34 (2006)
[4]. Mailund et al. Whole genome association mapping by incompatibilities and local perfect phylogenies, *BMC Bioinformatics*, 7 (2006)