**Linkage analysis on high density SNP arrays**

In large and complex pedigrees

Aude Saint-Pierre(1), Yuri D’Elia (2), Peter P Pramstaller(1,2,3), Cristian Pattaro(1)

(1) Center for Biomedicine, European Academy of Bozen/Bolzano (EURAC), Bolzano, Italy · Affiliated Institute of the University of Lübeck, Germany; (2) Department of Neurology, University of Lübeck, Germany; (3) Department of Neurology, Central Hospital of Bolzano, Bolzano, Italy.

**Workflow for linkage analysis using SNPs**

- **SNPs data selection (300K)**
  - Minimize LD between SNPs based on r2 (MASEL)
  - Clean for Mendelian inconsistencies (PEDCHECK)
- **Pedigree splitting**
  - MERLIN
- **Approximate IBD calculation** (LOKI)
- **Exact IBD calculation**
- **Linkage analysis on multiple pedigree configurations (MERLIN)**
- **Linkage analysis in extended pedigrees (SOLAR)**

**Table 3: Empirical type 1 error at 5%**

<table>
<thead>
<tr>
<th>Method</th>
<th>Full pedigrees</th>
<th>5%</th>
<th>1%</th>
<th>0.1%</th>
</tr>
</thead>
<tbody>
<tr>
<td>SOLAR</td>
<td>4.76%</td>
<td>0.90%</td>
<td>0.11%</td>
<td></td>
</tr>
<tr>
<td>MER3</td>
<td>4.39%</td>
<td>0.71%</td>
<td>0.04%</td>
<td></td>
</tr>
<tr>
<td>MER11</td>
<td>4.68%</td>
<td>0.81%</td>
<td>0.05%</td>
<td></td>
</tr>
<tr>
<td>MER21</td>
<td>4.44%</td>
<td>0.73%</td>
<td>0.06%</td>
<td></td>
</tr>
</tbody>
</table>

**Materials and Methods**

Pedigree reconstruction and SNPs data selection

- We considered 598 individuals from the MICROS dataset [4]. 322 were genotyped with the Illumina 300K SNP chip.
- Pedigree reconstruction with Buildped identified 1 informative family, but with the full pedigree.
- For simulation, we considered all SNPs in chromosome 22. 5381 SNPs were available for analyses after QC.
- Minimal LD SNP selection for linkage was performed with MASEL [5] (r2=0.01) and Mendelian inconsistency check with PEDCHECK [6].
- 133 SNPs remained for linkage analysis (Table 1).

Linkage analysis

- To analyse the full pedigree, the multipoint IBD matrix was estimated with LOKI [7] and VC linkage analysis was performed with SOLAR [8].
- For the multiple-splitting approach, a set of pedigree configurations was generated using the multiple pedigree splitting method [9]. Splitting was based on the kinship coefficient (Kin) and the Min-Max size of genotyped individuals within a family (Min-Max) (Table 2). VC analysis and exact multipoint IBD estimation were performed with MERLIN [10].

Empirical distribution of the VC linkage test

- Family information (family size, missing genotype/phenotype data) was kept as observed in the full pedigrees.
- Genotypes were simulated with MORGAN [11] for segments of 5 consecutive SNPs, randomly sampled on chr. 22.
- Null hypothesis of no linkage. Simulated values in the full pedigree assigned according to each pedigree configuration (Expl. var.=35%).
- Alternative hypothesis. For each replicate, a QTL (Expl. var=10%, MAF=0.1) was drawn in the middle of the map. For each workflow, both genotypes and phenotypes were simulated.

Summary statistic controlling for multiple splitting: maximum (MAX) and median (MED) LOD score across configurations for each replicate were calculated.

Empirical assessment: We assessed empirically the ability of the two pipelines to detect a known QTL by performing a linkage analysis of serum cystatin C [12]. The analysis was focused on chr. 20, with two SNPs placed on the cystatin gene locus (20p11.21). In the whole chromosome, a total of 315 independent SNPs were selected for linkage analysis.

**Conclusions**

The multiple splitting approach apparently has lower power than a full-pedigree analysis. However, for large number of SNPs analyzed, the multiple splitting approach is more efficient in terms of calculation time. Ideally, the two pipelines could be combined together by using the multiple-splitting approach as a fast screening tool on lower density maps and the full-pedigree analysis as a fine assessment tool on small, selected regions. Further assessment is ongoing.

**References**

[10] Abecasis et al., 2002

**Figure 1: Empirical power at 5%**

Results based on 1,000 replicates.

**Figure 2: Computation time**

<table>
<thead>
<tr>
<th>Time calculation (s)</th>
<th>SOLAR</th>
<th>MERLIN</th>
<th>1 Proc</th>
<th>4 Proc</th>
<th>32 Proc</th>
</tr>
</thead>
<tbody>
<tr>
<td>SOLAR</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MERLIN</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 Proc</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4 Proc</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>32 Proc</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Horizontal lines shows the empirical thresholds at 5% according to the analysis.

**Figure 3: LOD score distribution on chromosome 20 for Cystatin C**

The LOD score distribution is shown for each of the workflows, with the empirical thresholds at 5% according to the analysis.