Linkage analysis on high density SNP arrays in large and complex pedigrees

Aude Saint-Pierre[(1)], Yuri D'Elia [(1)], 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; contact: aude.saintpierre@eurac.edu;
(2) Department of Neurology, University of Lübeck, Germany; (3) Department of Neurology, Central Hospital of Bolzano, Bolzano, Italy.

Background

Motivation

- **Linkage analysis** is a complementary approach to association analysis in QTL mapping.
- If the linkage disequilibrium (LD) is appropriately handled, dense **SNPs** arrays can offer equal or superior power than **microsatellites** to detect **linkage [1]**.
- **Variance components** (VC) method is a popular approach to map QTL in large genealogies **[2]** but few studies evaluated its power on dense **SNPs arrays**.
- IBD estimation is the major drawback when using extended pedigrees. **Pedigree-splitting** has been advocated as a way to reduce the computational burden and to make linkage studies feasible. However, splitting may reduce power substantially **[3]**.

Objective

We assessed the performance of two alternative linkage analysis pipelines on dense SNP arrays: the full-pedigree approach was compared to the multiple-splitting approach by means of an extensive simulation. Finally, we assessed empirically the ability of the two pipelines to detect a known QTL.

Table 1: Characteristics of the final SNPs subset

Number of SNPs	133
Median MAF [lower quartile - upper quartile]	0.39 [0.21-0.47]
Mean MAF (sd)	0.33 (0.16)
Mean inter-SNP distance in Mb (sd)	0.26 (0.25)
Median r^2 [lower quartile - upper quartile]	0.005 [0.002-0.014]

Table 2: Characteristics of pedigree configurations extracted from the full pedigree

Minimum kinship	*0.0125*	*0.03125*	*0.0625*
Range size [minS-maxS]	*3-7*	*3-8*	*3-8*
Number of families	53	49	52
Number of phenotyped individuals	325	319	311
Mean kinship (sd)	0.145 (0.082)	0.152 (0.080)	0.171 (0.074)
Mean information content	0.72	0.80	0.84
Pedigree configuration number	MER3	MER11	MER21

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, bit no. = 764, 55% both genotyped and phenotyped.
- 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 pedigree**.**
- 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 pipeline 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.

Workflow for linkage analysis using SNPs

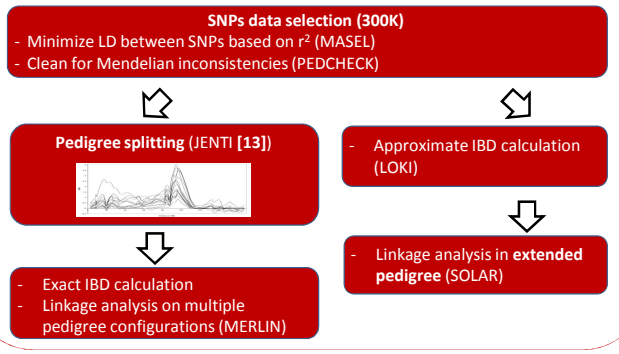
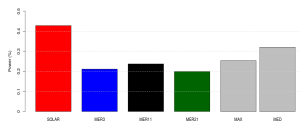


Table 3: Empirical type 1 error at 5%

	5%	1%	0.1%
SOLAR	4.76%	0.90%	0.11%
MER3	4.39%	0.71%	0.04%
MER11	4.68%	0.81%	0.05%
MER21	4.44%	0.73%	0.06%

+: Full pedigrees.
Empirical type 1 errors based on 20,000/30,000 replicates for SOLAR/MERLIN.

Figure 1: Empirical power at 5%



Results based on 1,000 replicates .

Figure 2: Computation time

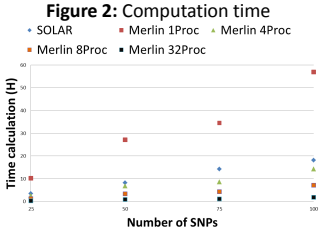
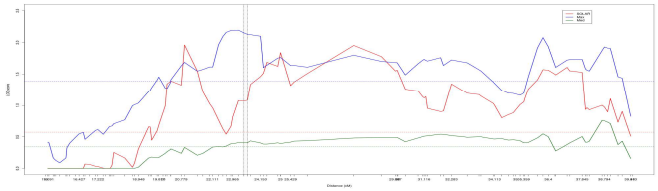


Figure 3: LOD score distribution on chromosome 20 for Cystatin C



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

Results

Type I error. When between-SNP LD is appropriately handled, both pipelines controlled the type I error appropriately (**Table 3**).

Power. Despite the approximated IBD estimation, the power of the full pedigree analysis was higher than the power of the multiple-splitting approach, which is based on exact IBD estimation, for any summary statistics (**Table 3**).

Computational time. The computational time is related to the density of the SNP map. At increasing SNP density, LOKI approximate IBD estimation is substantially slower than the MERLIN exact IBD estimation on small pedigrees (**Fig. 2**).

Application to cystatin C. Surprisingly, despite its higher power, we were unable to identify the cystatin C QTL with the full-pedigree pipeline (**Fig. 3**, red line). On the other hand, the QTL was clearly identified by the multiple-splitting approach when the maximum LOD score across splits was considered (**Fig. 3**, blue line). The QTL could not be identified by the multiple-splitting pipeline when the median LOD score across splits was chosen (**Fig. 3**, green line).

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

[1] Kruglyak, 1997
[2] Amos, 1994 .
[3] Dyer et al, 2001
[4] Pattaro et al, 2007
[5] Bellenguez et al, 2008
[6] O'Connell et al.,1998
[7] Heath, 1997
[8] Almasy and Blangero 1998
[9] Bellenguez et al, 2009
[10] Abecasis et al., 2002
[11] Thompson, 1995
[12] Köttgen et al, 2009
[13] Falchi and Fuchsberger, 2008