

On the value of family data in genome-wide association studies for quantitative traits

INTRODUCTION

Two different designs are used in genetic association studies : those that use families and population designs that use unrelated individuals. These two designs have different strengths and weaknesses. They are thus complementary rather than competitive. Here, our main aims are to investigate the statistical properties of a new two-stage GWAS strategy that uses these two types of sample data and these two samples are not statistically independent. The practical relevance of this methodology is related to the genome-wide association study for Bone Mineral Density we are conducting in the NEMO data using the Infinium 370K panel.

MATERIALS AND METHODS

NEMO SAMPLES : (1) **NEMO-family sample** : 105 pedigrees ascertained through a low (Z-score ≤ -2) BMD values at the Lumbar Spine or at the Femoral Neck. The mean pedigree size is ~ 8 subjects and, out of a total of 835 subjects ~ 600 have DNA available and known phenotype. (2) **NEMO-population sample** : 240 unrelated individuals (all men) with BMD values selected from the tails : Z-score ≤ -2 at LS-BMD (« High » subset) or Z-score > 0.5 at both LS and FN-BMD (« Low » subset). All NEMO subjects and families were collected in France or in Belgium.

TWO-STAGE DESIGN : In the first step (**scan**) we combined the NEMO-population and the probands from the NEMO-family sample (scan sample : ~ 190 and 155 subjects in the « High » and « Low » subset respectively), and tested for association the full set of SNPs. The most significant SNPs (i.e., nominal $P \leq 5\%$) are tested in the second step (**replication**) that uses the **NEMO-family sample with or without the phenotypes of the probands**. In the former replication design, probands are used twice (in the scan and in the replication step). Thus the scan and replication samples are not statistically independent. Under the second design (« No-Proband »), the replication testing uses data on the relatives of probands only. Figure 1 shows the distribution of LS-BMD values in the NEMO data.

ASSOCIATION TESTS : The scan step was carried out using linear regression tests. In the replication step, we used two family-based association tests : (1) the Quantitative Transmission Disequilibrium Test and the Measured Genotype test. We also computed the MG test accounting for hidden stratification (MG|Strat). All association tests were computed assuming SNPs with additive effects.

SIMULATION STUDY : BMD values and family information (family sizes, missing genotype/phenotype data) were kept as observed in NEMO data. Genotypes of K independent SNPs were first generated under the null hypothesis of no association to LS-BMD. Each generated data was analyzed as follows : a) All K SNPs were tested for association in the « scan » replicate -b) the most significant SNPs ($P \leq 5\%$) were tested for replication in the « replication » replicate with the QTDT and MG tests using the full or « No-Proband » NEMO-family sample -c) the number of SNPs replicated at a given nominal P-value under QTDT or MG (MG|Strat) was recorded. The type I error of our two-stage design was obtained as the mean number of replicated SNPs out of a total number of R replicates. We also generated SNPs genotypes under the alternative hypothesis of association. Power estimates were derived for common ($MAF = 0.20$ to 0.50) causative variants explaining from 0.5 to 5% of the trait variance.

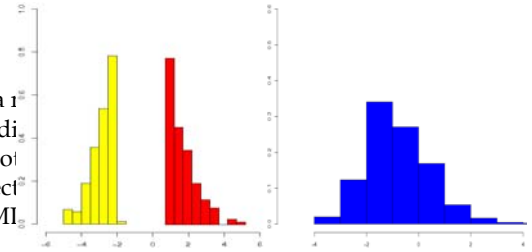


Figure 1 : Distribution of LS-BMD values in the scan (« Low » and « High » subsets) and in the relatives of probands (« No-Proband » family subset) samples.

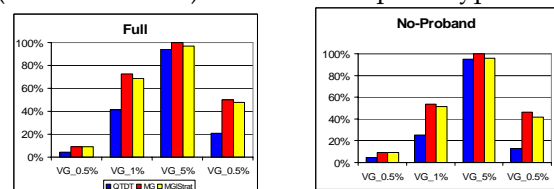
RESULTS & CONCLUSIONS

Preliminary results shown : genomic screening of a panel of 1,500 SNPs only – empirical significance rates based on 500 replicates only

1- Empirical Type I Error rates of our two-stage design at a nominal $P=5\%$, for QTDT and MG tests and with or without the Proband's phenotypes.

Replication sample	MG	QTDT	STRAT	MG Strat
Full	8.6%	5.8%	4.5%	8.2%
No-Proband	5.5%	3.2%	4.3%	5.3%

2- Power of our two-stage design at a nominal $P=5\%$, for QTDT and MG tests and with (« Full ») or without (« No-Proband ») the Proband's phenotypes.



• When conducting the replication step using the full NEMO-family sample as expected, rates of false positive association were increased as compared to the nominal values, especially with MG test. Empirical type I error rates were closer to the nominal ones when the replication step does not use phenotypic data of the probands.

• Not surprisingly, our study design lacks power to detect variants with weak effects (explaining less than 1% of the total variance), especially when using the QTDT test. Otherwise, our proposed two-stage design appears to be a powerful GWAS design to detect common variants for complex quantitative traits.

More work needed : increase the number of SNPs tested at the scan step, and evaluate significance rates under a larger number of replicates.