Genome-wide environmental interaction analysis using multidimensional data reduction principles to identify asthma pharmacogenetic loci in relation to corticosteroid therapy

F. Van Lishout[1], K. Bessonov[1], Q.L. Duan[2], E.S. Gusareva[1], J.M. Mahachie John[1], K. Tantishira[2,3], and K. Van Steen[1]

[1]Montefiore Institute and GIGA-Research, University of Liège, Belgium
[2]Channing Division of Network Medicine, Brigham & Women's Hospital and Harvard Medical School, USA
[3]Pulmonary Division, Brigham & Women's Hospital and Harvard Medical School, USA

Problem Formulation

- How to efficiently discover the most significant SNP-environment interactions in search for asthma pharmacogenetic loci?
- We analyze the difference in pre-bronchodilator FEV1 in patients following or not ICS therapy for a period of 8 weeks (*prech_short*), for 550 pediatric Caucasian CAMP (ages 5-12) from the SHARE project
- The trait of interest is *prech_short* expressed on a continuous scale and represents a relative difference in preFEV1:

  $prech\_short=(prefev_{on\_ICS}-prefev_{off\_ICS})/prefev_{off\_ICS}$
- The environmental variable is dichotomous and refers to inhaled corticosteroids therapy (ICS) based on budesonide. If ICS is administrated it is coded 1 and 0 otherwise

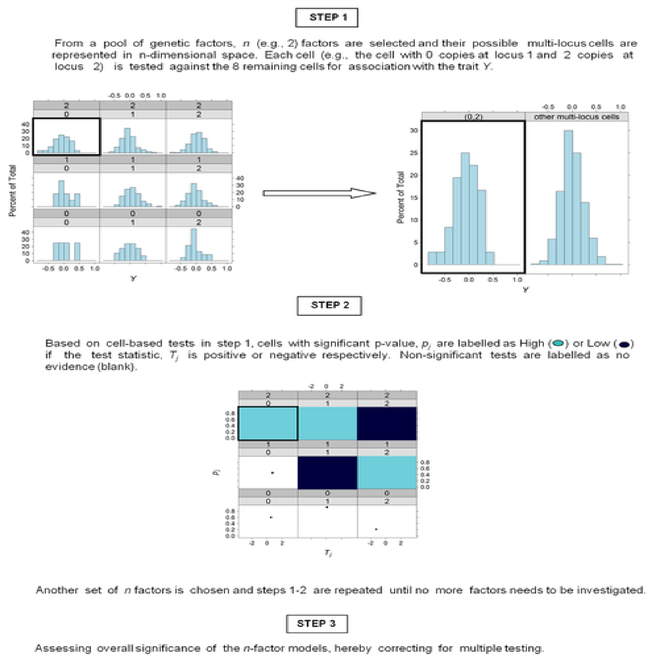
Data Preparation

- We analyze 550 samples containing no reported family structure
- Missing genotypes were MaCH-imputed using 1000 Genomes Project Reference Panels resulting in 8,221,073 SNPs
- Since the T-gene was found to be associated to asthma [6], a total of 5,793 SNPs found within 1Mb range from T-gene coding region start and end in both 5' and 3' direction where added to the marker panel
- Genotype data QC steps consisted of the following steps: LD pruning (*SNPRelate* library in R) with maximum between-marker r^2 of 0.2 (yielding 231,568 SNPs), removal of poorly annotated SNPs, removal of SNPs not present in the dbSNP database, removal of SNPs with MAF< 0.01, HWE at FDR maximum cutoff of 0.2. Samples and their genotypes passing QC were extracted with PLINK
- The final subset consisted of 69,171 markers with population inflation factor λ=1.001 (minimal population stratification effects).
- Genetic Idenity-by-State (IBS) kinship matrix was calculated using allelic frequency and applied as part of polygenic model
- Trait residuals were computed in two ways from trait ~ sex+age+BMI: 1) based on a polygenic regression model (POLY) using observed kindships (GenABEL 1.7.6); 2) based on linear regression (GLM) in R. These were taken as input to MB-MDR 4.0.1 either as such or Rank Transformed to Normality (RTN)

State-of-the-art

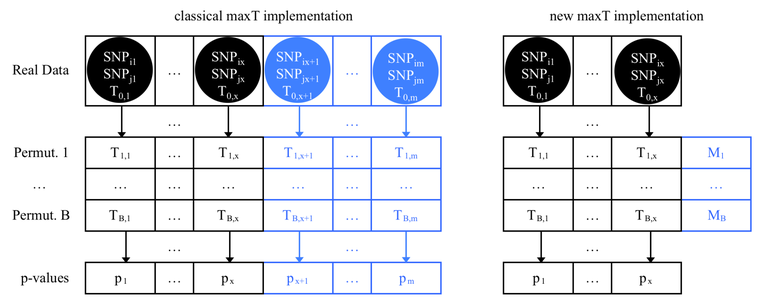
- Genome-wide gene-environment (GWEI) and gene-gene (GWAI) interaction studies share a lot of challenges due to high-dimensionality concerns. GWEI studies may benefit from methodologically resolved issues in the context of GWAIs
- Model-Based Multifactor Dimensionality Reduction (MB-MDR), initially built for epistasis detection is also useful to discover gene-drug interactions. It does not make any assumption about the genetic inheritance model and involves reducing a high-dimensional GxE space to a GxE summary variable with factor levels that either exhibit high, low or no evidence for their association to disease outcome. In contrast to logistic regression and random forests, MB-MDR can be used to detect GxE interactions in the absence of any main effects.
- The nature and the effect of population stratification in genome-wide interaction context has not rigorously been studied

Graphical Workflow of MB-MDR Methodology [1]



Multiple Testing

- Especially in the context of high-order genome-wide interaction studies one of the challenges is to handle the severe multiple-testing problem associated with them, while adequately controlling the number of false positives and acknowledging intrinsic complexities dependencies between tests
- We have developed a new implementation of the maxT algorithm of Westfall & Young [4], requiring an amount of memory independent from the number of genetic effects to be investigated
- A graphical explanation of the differences between the classical and new implementation of the maxT algorithm is given below [5]



- In the classical maxT implementation all $T_{i,j}$ values are in memory. If the 1000 best MB-MDR *p*-values are envisaged, then only the maximum $M_1,...,M_B$ of the $[T_{1,1000+1},...,T_{1,m}],...,[T_{B,1000+1},...,T_{B,m}]$ together with $[T_{1,1},...,T_{1,1000}],...,[T_{B,1},...,T_{B,1000}]$ are retained

Population Stratification Correction

- We propose two strategies to correct for population stratification, hereafter referred to as STRAT1 and STRAT2, avoiding the use of principal components (PCs)
- In both cases, we first compute the median M_1 of all observed MB-MDR test-statistics. Second, we use the new implementation of MAXT on re-scaled MB-MDR test values. In particular,
  - in STRAT1 we divide all observed MB-MDR test values by M_1/M_2, where M_2 is the median of all permutation-based MB-MDR test values the statistics computed on the permuted data
  - in STRAT2, we divide each observed MB-MDR test value for interaction i by $M_1/M_{2,i}$, where $M_{2,i}$ is the median of the permutation-based MB-MDR test values for the i^{th} interaction

Simulation Study: Epistasis=NO - Pop Strat.=YES/NO

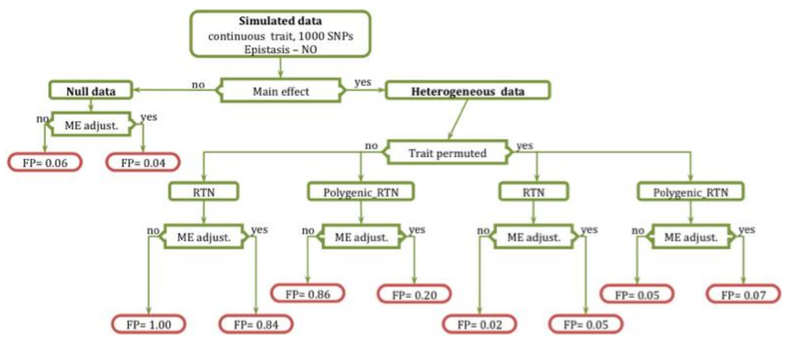


Figure 1: MB-MDR false positive rates (FP) under a variety of scenarios

Results

- We performed 12 different analyses according to
  - algorithm: MAXT, STRAT1 or STRAT2
  - polygenic regression: yes (POLY) or no (GLM)
  - rank transformation to normality: yes (RTN) or no (NONE)

Table 1: MB-MDR *p*-values

	POLY + RTN		POLY + NONE		GLM + RTN		GLM + NONE	
MAXT	rs62396388	0.702	rs7228304	0.311	rs12154075	0.663	rs7228304	0.352
	rs60857913	0.798	rs61680817	0.365	rs12194567	0.663	rs61680817	0.358
	rs12154075	0.972	rs10275131	0.37	rs60880137	0.663	rs10275131	0.381
	rs12194567	0.972	rs28705331	0.547	rs2277094	0.663	rs6591575	0.487
	rs60880137	0.972	rs2354224	0.555	rs3816304	0.663	rs2354224	0.549
	rs2277094	0.972	rs6591575	0.592	rs35606910	0.663	rs28705331	0.571
	rs3816304	0.972	rs11048468	0.601	rs3127328	0.663	rs17729514	0.699
	rs35606910	0.972	rs1118406	0.652	rs60857913	0.673	rs1118406	0.712
	rs3127328	0.972	rs17729514	0.682	rs13191186	0.733	rs11048468	0.731
	rs2229604	0.98	rs2979120	0.706	rs35402370	0.733	rs2979120	0.731
STRAT1	rs62396388	0.683	rs7228304	0.009	rs12154075	0.068	rs7228304	0.115
	rs60857913	0.769	rs61680817	0.01	rs2277094	0.068	rs61680817	0.117
	rs2277094	0.946	rs10275131	0.01	rs60880137	0.068	rs10275131	0.128
	rs3816304	0.946	rs28705331	0.015	rs3816304	0.068	rs6591575	0.177
	rs35606910	0.946	rs2354224	0.015	rs35606910	0.068	rs2354224	0.206
	rs12154075	0.946	rs6591575	0.017	rs3127328	0.068	rs28705331	0.22
	rs12194567	0.946	rs11048468	0.018	rs12194567	0.068	rs17729514	0.298
	rs3127328	0.946	rs1118406	0.019	rs60857913	0.07	rs1118406	0.306
	rs60880137	0.946	rs17729514	0.021	rs12200529	0.09	rs11048468	0.314
	rs2229604	0.956	rs2979120	0.028	rs12207298	0.09	rs2979120	0.315
STRAT2	rs60857913	0.741	rs11782301	0.104	rs8097909	0.206	rs10275131	0.118
	rs12407302	0.971	rs12407302	0.284	rs3843762	0.21	rs2970667	0.511
	rs11782301	0.976	rs60857913	0.288	rs1335049	0.254	rs11782301	0.604
	rs1335049	0.977	rs41302377	0.313	rs12407302	0.333	rs10045785	0.716
	rs767412	0.98	rs13333892	0.37	rs12611982	0.535	rs3796058	0.744
	rs11055237	0.994	rs1335049	0.419	rs9534197	0.57	rs7924068	0.772
	rs9735778	0.997	rs10741458	0.469	rs2229604	0.643	rs1335049	0.807
	rs60880137	0.997	rs62159914	0.516	rs17447754	0.663	rs61943065	0.902
	rs10830629	1	rs62510045	0.519	rs12918211	0.689	rs4462275	0.927
	rs9631715	1	rs4936023	0.526	rs4448534	0.689	rs9872256	0.929

Discussion

- The smallest *p*-values are obtained for STRAT1-POLY-NONE
- In the presence of population stratification MAXT-POLY-RTN has lower false-positive rates (FP=0.20) as compared to MAXT-POLY-NONE. In the absence of population stratification these options keep FP under control (**Figure1**)
- By construction STRAT2 better accommodates allele-specific MB-MDR test distributions and is to be preferred (cfr., pair-specific "genomic" control; simulation results not shown) (**Table 1**)
- Whether polygenic control is appropriate for structured data in genome-wide interaction settings needs further investigation. Similarly, assessing the relative advantage of polygenic control over regression models not using kinship information (GLM), yet possibly corrected for ancestry-related confounding variables, is underway.
- Nevertheless, rs11782301 was one of the 94 SNPs that occurred in the top 1000 MB-MDR outputs for all 12 investigated scenarios. It had on average the lowest MB-MDR *p*-value. This SNP maps to the ZHX2 gene - transcription factor, a member of (zinc fingers and homeoboxes 2). ZHX2 was shown to be differentially expressed in airway smooth muscle cells and might be implicated in asthma [7]

About the Software: MBMDR-4.0.1

- MBMDR 4.0.1 is a flexible and efficient C++ implementation of the MB-MDR methodology [1]. The software can be downloaded from http://www.statgen.ulg.ac.be/ and is available for mac and linux.
- The C++ MB-MDR software can optionally convert PLINK formatted input data files into MBMDR-4.0.1's internal format.
- Traits can either be expressed on a binary or continuous scale. Censored traits are also accommodated
- The software can either perform a global test or an interaction-specific test that adjusts for main effects. Co-dominant main effect corrections are recommended [1]
- Apart from two-order interactions, three-order interactions such as: GxGxG, GxGxE, GxExE, are easily run in parallel mode

Conclusions

- We have designed a new implementation of the maxT algorithm [5], which makes genome-wide interaction studies with MB-MDR feasible
- We have developed new algorithms to correct for population stratification, that avoid making choices about the number of principal components to retain and how to compute them
- The STRAT corrections use ideas from genomic controlling in main effects GWAs. The recommended use of RTN for quantitative trait MB-MDR analysis needs to be revised in the context of population stratification

Acknowledgements

This work was partially funded by the Interuniversity Attraction Poles Programme (IAP P6/25 BIOMAGNET and IAP P7 DYSCO), initiated by the Belgian State, Science Policy Office and by the European Network of Excellence PASCAL2. The authors thank the GIGA Bioinformatics platform for providing computing resources. Tantishira K. and Duan QL acknowledges the source of funding U01 HL06589. In addition, Duan QL receives funding from the National Heart, Lung and Blood Institute: 1K99HL116651

References

[1] Mahachie JM, Cattaert T, Van Lishout F, Gusareva ES and Van Steen K.: **Lower-Order Effects Adjustment in Quantitative Traits Model-Based Multifactor Dimensionality Reduction.** *PlosOne* 2011.

[2] Ritchie MD, Hahn LW, Roodi N, Bailey LR, Dupont WD, Parl FF, Moore JH: **Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer**. *Am J Hum Genet* 2001.

[3] Van Steen K.: **Travelling the world of gene-gene interactions**. *Briefings in Bioinformatics* 2011.

[4] Westfall PH and Young SS: **Resampling-base multiple testing.** *New York, Wiley* 1993.

[5] Van Lishout F, Mahachie John JM, Gusareva ES, Urrea V, Cleynen I, Théâtre E, Charloteaux B, Calle ML, Wehenkel L, Van Steen K: **An efficient algorithm to perform multiple testing in epistasis screening.** *BMC Bioinformatics* 2013.

[6] Tantisira KG, Damask A, Szefler SJ, Schuemann B, Markezich A, Su J, Klanderman B, Sylvia J, Wu R, Martinez F, boushey HA, Chinchilli VM, Mauger D, Weiss ST, Israel E; SHARP Investigators: **Genome-wide Association Identifies the T Gene as a Novel Asthma Pharmacogenetic Locus.** *American Journal of Respiratory and critical care medicine* 2012.

[7] Misior A. PKA- and gluco corticoid-dependent regulation of gene expression in human airway smooth muscle cells. Am K Respir Cell Mol Biology 41(1) 2009

For further information contact f.vanlishout@ulg.ac.be