

GENETIC EPISTASIS ANALYSIS USING TAXONOMY3

Rémi Lebreton¹, Christophe Biernacki¹, Serge Iovleff¹, Julien Jacques¹, Cristian Preda¹, Alun McCarthy², Olivier Delrieu^{2*}

1: Université Lille 1, UFR de Mathématiques, UMR CNRS 8524, 59655 Villeneuve d'Ascq Cedex, France & INRIA Lille-Nord Europe.

2: Pharmacogenomic Innovative Solutions Ltd, Aston Court, Kingsmead Business Park, High Wycombe, Bucks, HP11 1LA, United Kingdom.

*Correspondence to: olivier.delrieu@pgxis.com

ABSTRACT

'Taxonomy3' is a novel mathematical method for the multivariate analysis of complex datasets. It is based on correlations of individualized divergences named Log Bayes Factors (LBFs), and their Eigen decomposition. LBFs represent a continuous measure of information each subject and variable provide pertaining to the overall case/control distinction. This framework offers a unique opportunity to conduct large scale interaction analyses and measure the epistasis level between markers. We applied this method to 51 flucloxacillin induced liver injury cases contributed by DILIGEN (UK) and 282 country- and gender-matched controls from the Population Reference Sample resource (POPRES). Subjects were genotyped with the Illumina 1M-Duo BeadChip. Results show significant novel markers that were not revealed by the published univariate analysis of main effects. This confirms the benefits of this new method, and the importance of assessing genetic epistasis in complex disorders.

INTRODUCTION

Tax3 is a multivariate analysis method based on individualized divergences, the natural metric for the comparison of individuals or groups in the co-analysis of multiple variable types. Tax3 is the only rigorous method allowing proper use of linear algebra tools with SNP data.

The method has the following characteristics:

- Greater power than conventional univariate analysis
- Identification of patient sub-groups /sub-phenotypes
- Elucidation of interactions between variables
- Patient selection based on whole genome prediction

OBJECTIVES

Our objective was to identify novel gene variants associated with flucloxacillin-induced liver injury (flucloxacillin DILI), using Tax3 incorporating a large scale interaction analysis.

SUBJECTS AND MATERIALS

Data were obtained from the iSAEC. The dataset consisted of genotype & phenotype data for 51 flucloxacillin DILI cases contributed by DILIGEN (UK) and 282 country- and gender-matched controls from the Population Reference Sample resource (POPRES). The samples were genotyped using the Illumina Human1M-Duo chip: after quality control, 904,158 SNPs were taken forward for analysis.

METHODS

LBFs, PCA, Interactions & Residuals

LBF calculation and Principal Component Analysis (PCA) were carried out as previously described¹. Interaction LBF values were regressed on the per-person LBF values of both the original constituent markers within that interaction. The residual of this regression is a measure of the epistasis between the two original markers².

Case/control Prediction

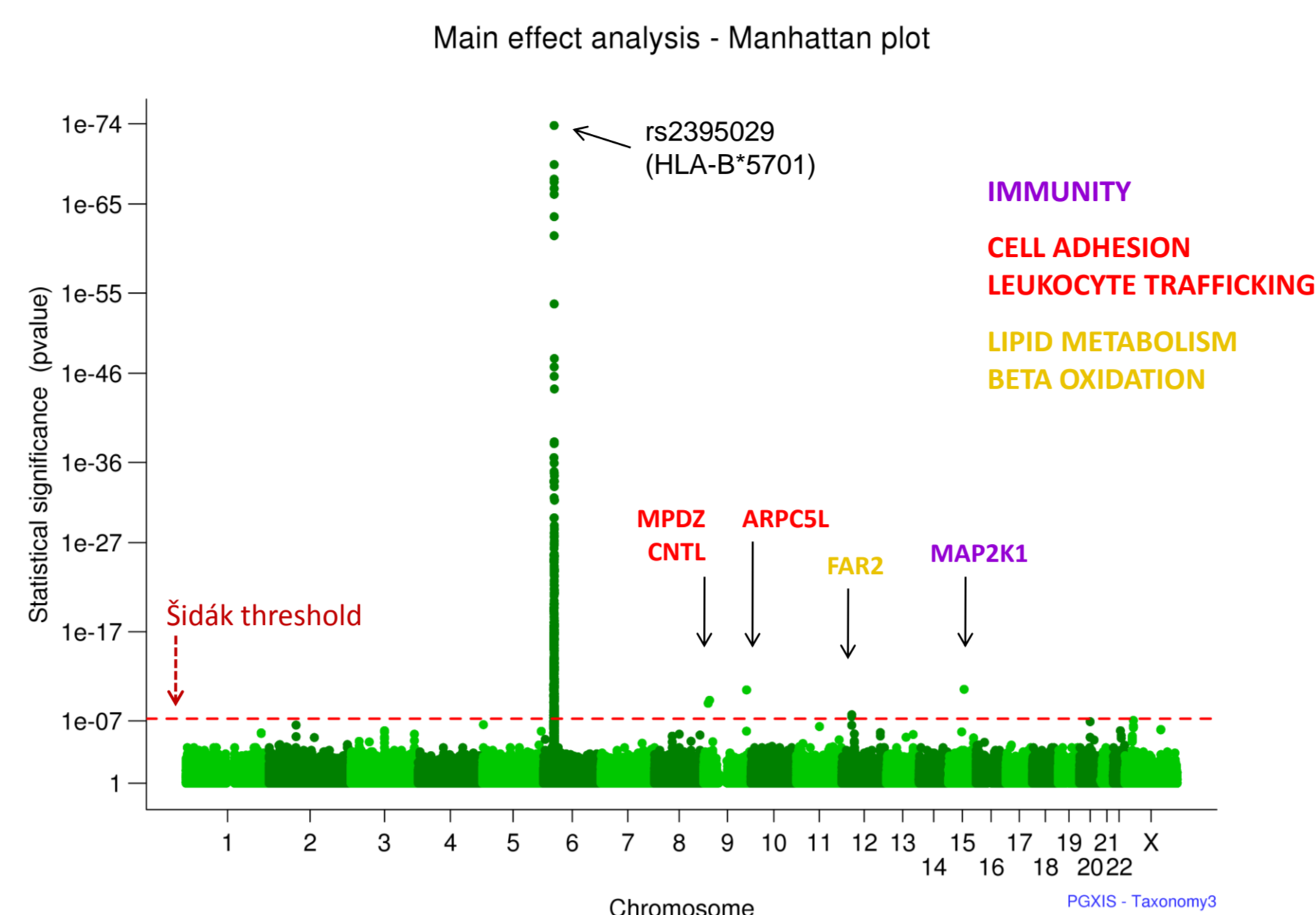
Leave-One-Out Cross-Validation was used to assess dataset predictive characteristics, with or without the interaction terms. Subjects of unknown status were attributed LBFs given their genotypes. Their position on PCA case/control axis allowed status inference.

Significance Assessment

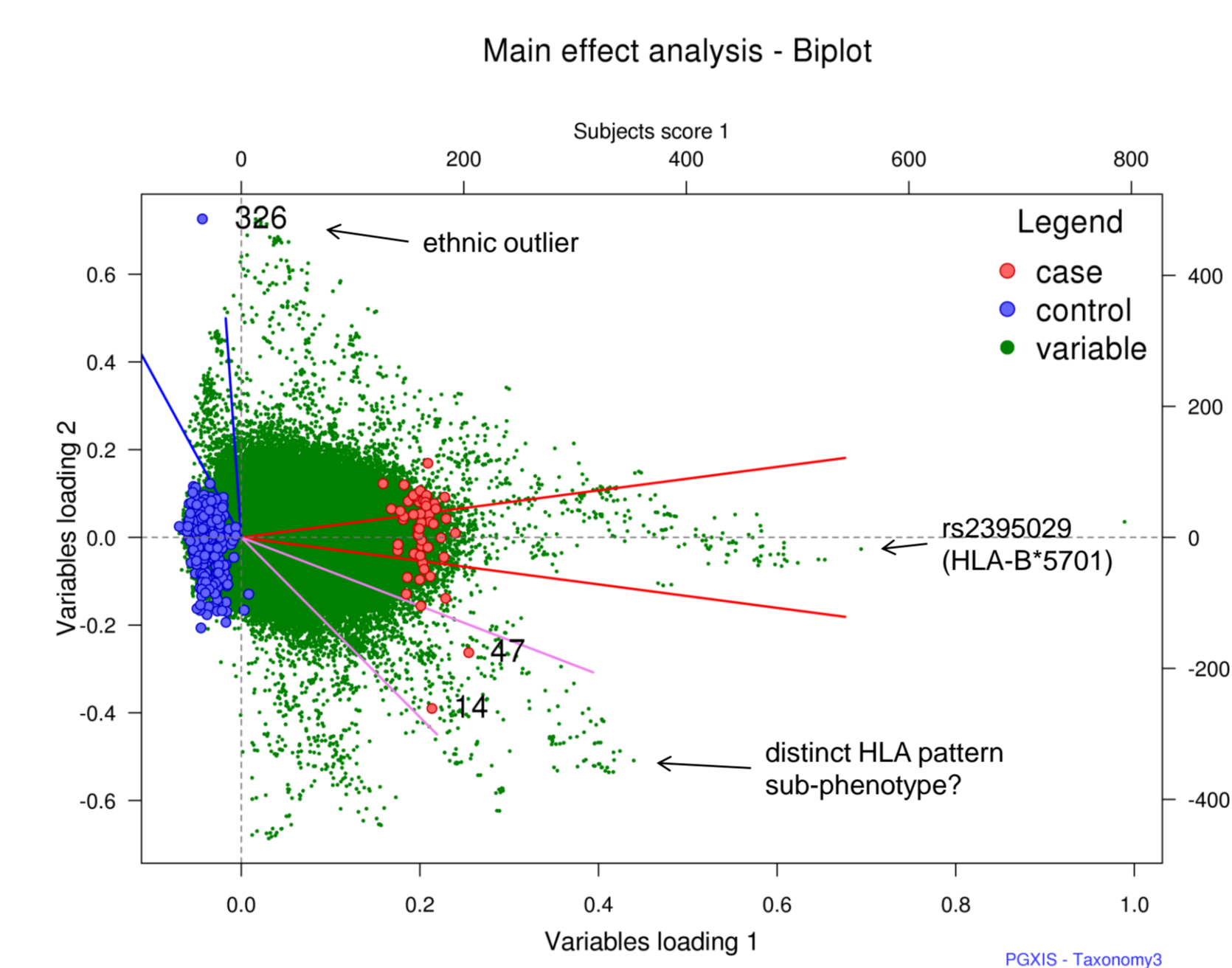
Case/control status permutation resampling (5,000 cycles) was used to assess statistical significance. For each variable, precise assessment of significance was carried out with a density estimation of resampled loadings, using a Gaussian mixture model approach based on expectation-maximization algorithm (MIXMOD³). The Family Wise Error Rate Šidák correction was used to define the genome-wide threshold for significant variables.

CASE/CONTROL ANALYSIS SHOWS GREATER POWER

In addition to confirming the published⁴ association with HLA-B*5701, further significant loci were discovered which mapped to a highly plausible set of genes.



The main case/control analysis also revealed genetic heterogeneity among the cases and potential sub-phenotypes, suggesting that the method can distinguish distinct genetic patterns related to distinct forms of DILI susceptibility.

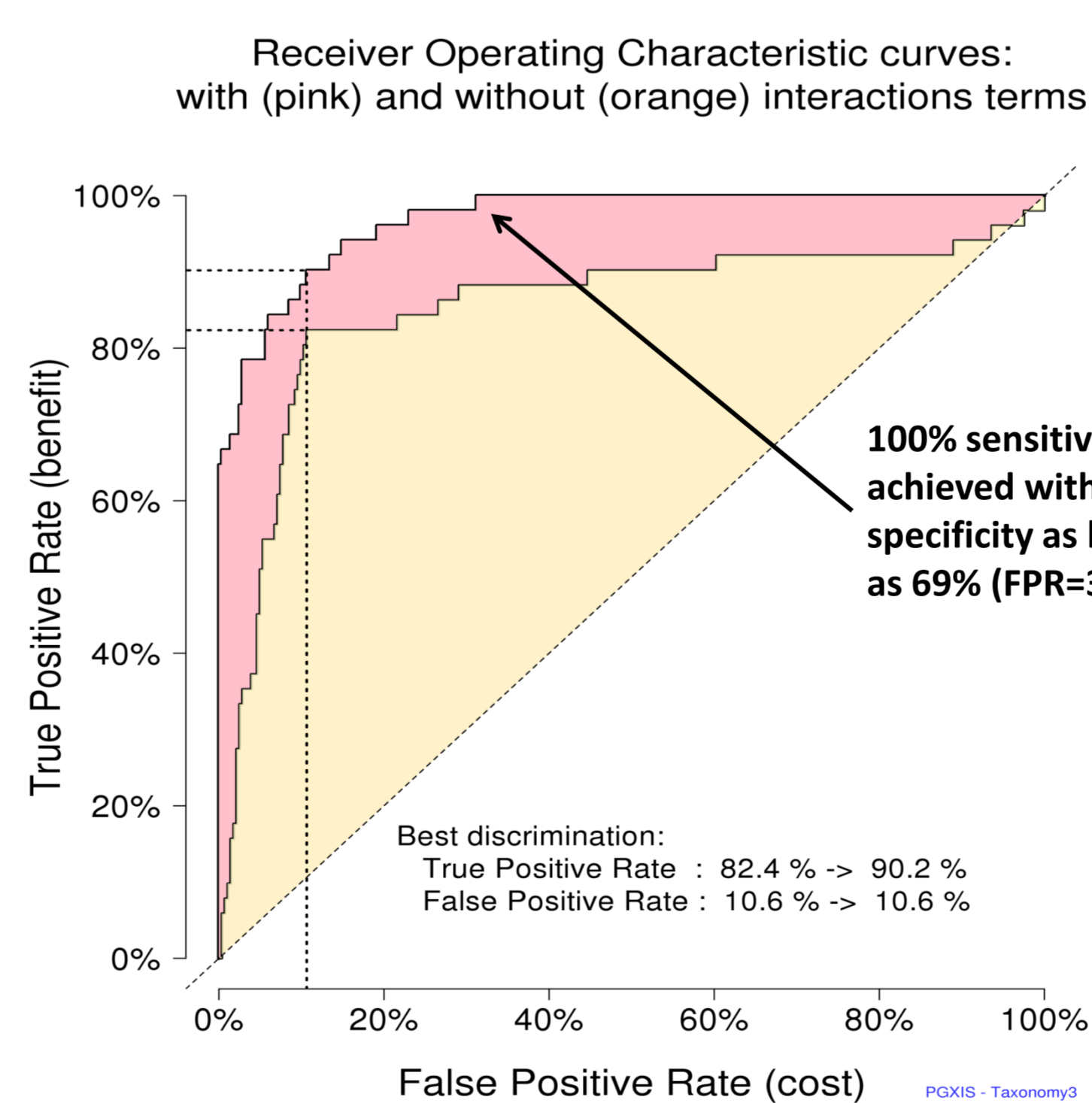


INTERACTIONS IMPROVE PREDICTION

An interaction analysis was performed with 2757 candidate SNPs having elevated main effect.

A case/control inference model was put in place using all available data (*whole genome predictor*).

Adding all interaction terms improved significantly the characteristics of the model. The best discriminator had a sensitivity of 90% and a specificity of 89%.

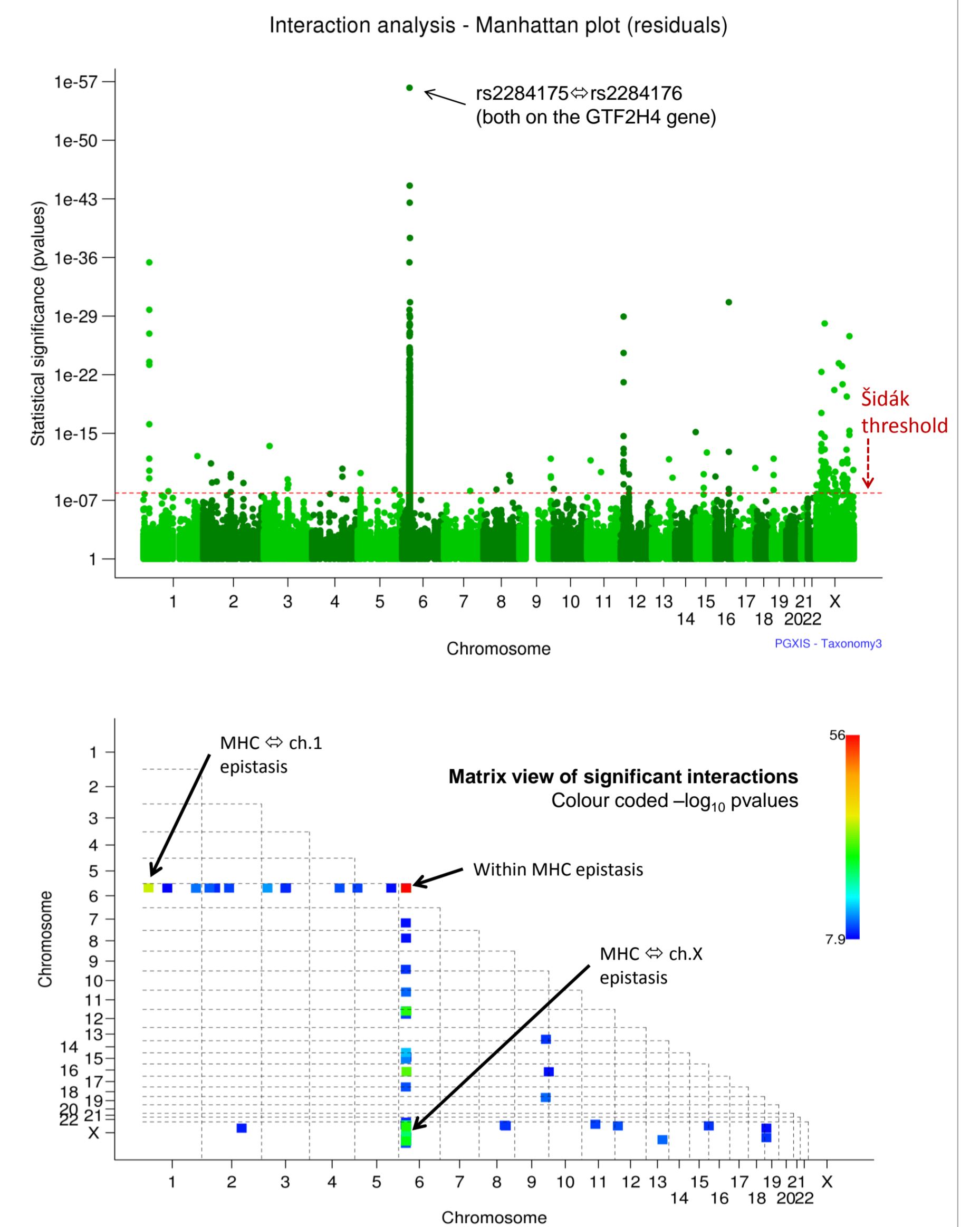


SOFTWARE

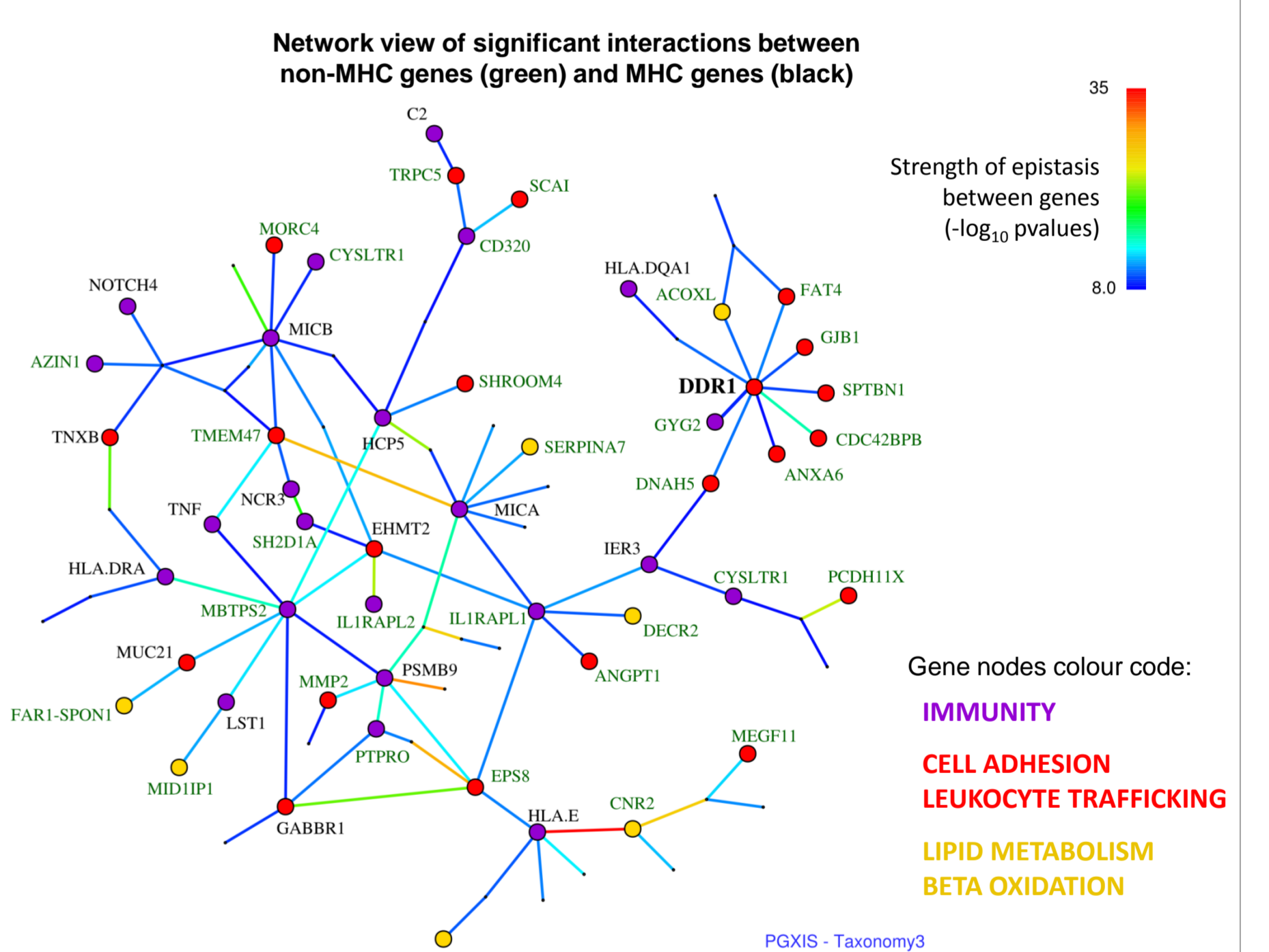
Our proprietary software allows rapid and efficient analyses. It is written in C++, multi-threaded, makes use of High Performance Message Passing Library (openMPI), is cluster-ready and encodes SNPs at the binary level.

INTERACTION ANALYSIS REVEALS NEW SIGNIFICANT PATHWAYS

Significant epistasis was found mainly within the MHC region, and between the MHC and non MHC regions, especially on chromosome 1, 12 and X.



These non MHC loci form highly plausible sets of genes, mapping to three distinct pathways – immunity; leucocyte adhesion/trafficking; lipid metabolism. They have significant epistasis and delimit a network with key 'entry points' MHC genes, such as DDR1.



CONCLUSION

This analysis using Tax3 firstly confirmed the published association of HLA-B*5701 with Flucloxacillin DILI and in addition, uncovered additional significant associations showing the greater power of the multivariate analysis. Interaction analysis – an important capability of Tax3 analysis – yielded a number of associations that flag three biological pathways, providing novel biological insights into the liver injury. Finally the whole genome data and interaction variables can be used to significantly enhance case/control prediction.

REFERENCES

1. Visualizing gene determinants of disease in drug discovery. Delrieu O and Bowman C. Pharmacogenomics. 2006 Apr;7(3):311-29.
2. Correlation laplacians, haplotype networks and residual pharmacogenetics. Bowman, C.E., and Delrieu, O. (2009). In A. Gusnanto, K.V. Mardia, & C.J. Fallaize (eds), Statistical Tools for Challenges in Bioinformatics, pp.25-31. Leeds, Leeds University Press.
3. Model-based cluster and discriminant analysis with the MIXMOD software. Biernacki, C et al. Computational Statistics & Data Analysis 51, 587-600 (2006).
4. HLA-B*5701 genotype is a major determinant of drug-induced liver injury due to flucloxacillin. Daly, A.K. et al. Nat. Genet 41, 816-819 (2009).

ACKNOWLEDGEMENTS

International Serious Adverse Event Consortium (iSAEC) and DILIGEN for providing the dataset.