

What my team can offer

The International HapMap Project was designed to create a genome-wide database of patterns of human genetic variation, with the expectation that these patterns would be useful for genetic association studies of common diseases. This expectation has already been fulfilled with just the initial output of genome-wide association studies, identifying nearly 100 loci for nearly 40 common diseases and traits (Manolio et al 2008). Genetic association studies can therefore be considered central to efforts to identify and characterize genomic variants that underly susceptibility to complex disease (Van Steen et al 2005).

We offer guidance and support to increase the success rate of your genetic association studies, in several areas including:

- Study design (e.g. subject selection: family- or population-based, variant or marker selection)
- Study implementation (e.g. sample-recruitment strategy)
- Power calculations
- Quality control of the data used for association analysis:
 - Detection of genotyping error (e.g. Hardy-Weinberg equilibrium, Mendelian inconsistency checks)
 - Reduction / Minimization of genotyping error
- Relevant data analysis (main effects analysis, interaction analysis, multi-locus analysis or multivariate analysis) using a variety of techniques that appropriately account for multiple testing and confounding factors:
 - Hypothesis testing techniques
 - Effect size estimation techniques
 - Data mining techniques
 - Descriptive data visualization techniques
 - Pattern searching techniques
 - Dimensionality reduction techniques
- Valid interpretation of association findings
- Meta-analysis of effect sizes (e.g. odds ratios), acknowledging any heterogeneity and the potential for bias within included studies (e.g., poor validity, selective reporting) and in the data set as a whole (e.g., publication bias).

Despite the successes of Genome Wide Association studies (GWAs), it has become clear that usually only a small percentage of total genetic heritability can be explained by the identified loci. For instance for inflammatory bowel disease (IBD), 32 loci significantly impact disease but they explain only 10% of disease risk and 20% of genetic risk (Barrett et al 2008). This may be attributed to the fact that reality shows multiple small associations (in contrast to statistical techniques that can only detect moderate to large associations), dominance or over-dominance, and involves non-SNP polymorphisms, as well as epigenetic effects and gene-gene interactions (Dixon et al 2000).

In response to these concerns we develop, evaluate and implement powerful methodologies to detect gene-gene interactions, in small-scale studies, but also at the genome wide level (GWAIs = genome wide association interaction studies). The newly developed strategies aim to have improved power to simultaneously detect multiple genetic loci associated with any clinical endpoint of interest (measured, dichotomous, survival type, longitudinal, etc). Special attention is given to integrating



information from different –omics data sources, incorporating important confounding information, adequately dealing with missing data (Van Steen et al 2007), adjusting for error sources such as population stratification and substructure in the data, and incorporating multi-staging into the analysis process (cfr multi-stage approaches in family-based association testing: Van Steen et al 2011a). While developing novel approaches, we particularly address the “curse of dimensionality” and the necessity to implement feasible, yet sound, multiple testing strategies. An “in-house” method to detect gene-gene interactions is MB-MDR (Calle et al 2008a,b, 2010; Cattaert et al 2011, Mahachie John et al 2011) and FAM-MDR (Cattaert et al 2010). Its roots lie in Vic (Spain), which initiated a long-term collaboration on the “MB-MDR” project. This multifactor dimensionality reduction method exhibits sufficient power in the presence of genetic heterogeneity (Cattaert et al 2011) and generates low false positive rates when screening at a genome wide scale (unpublished results). Future challenges include combining evidence from different modeling approaches, and visualizing results from epistasis screens (Figure 1: toy example), while adequately adjusting for lower-order effects or possible confounders (Van Steen et al 2011b).

And of course ... genetics is merely one of the playgrounds that we tend to empty our biostatistical tool box on.

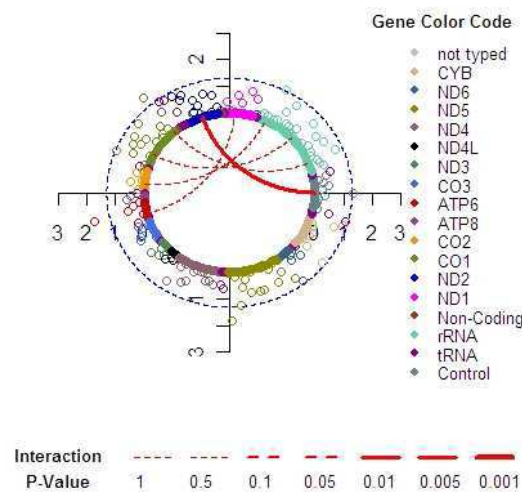
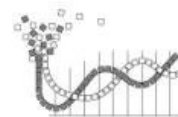


Figure 1: Epistasis results (red lines in the hyperbolic plane) and main effects signals (bullets radiating out in strength from the center of the plot). The dashed blue line demarcates the 0.05 overall significance level for adjusted p-values after an MB-MDR 1D run. The red dashed hyperbolic lines indicate pairs of markers that have a $-\log_{10}$ adjusted p-value above 0.05. As evidence amounts for epistatic pairs, the thickness of the line increases.



References

- Barrett, J. C. et al (2008) Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. *Nat Genet* 40(8), 955-962.
- Calle, M. L.; Urrea, V.; Malats, N.; Van Steen, K. (2008a) MB-MDR: Model-Based Multifactor Dimensionality Reduction for detecting interactions in high-dimensional genomic data. Technical Report n.24. Department of Systems Biology. Universitat de Vic.
- Calle M.; Urrea V.; Vellalta G.; Malats N.; Van Steen K. (2008b) Improving strategies for detecting genetic patterns of disease susceptibility in association studies. *Statistics in Medicine* 27(30):6532-46
- Calle ML, Urrea V, Van Steen K (2010) mbmdr: an R package for exploring gene-gene interactions associated with binary or quantitative traits. *Bioinformatics Applications Note* 26 (17): 2198-2199
- Cattaert T, Urrea V, Naj AC, De Lobel L, De Wit V, Fu M, Mahachie John JM, Shen H, Calle ML, Ritchie MD, Edwards T, Van Steen K. (2010) FAM-MDR: a flexible family-based multifactor dimensionality reduction technique to detect epistasis using related individuals, *PLoS One* 5 (4).
- Cattaert T, Calle ML, Dudek SM, Mahachie John JM, Van Lishout F, Urrea V, Ritchie MD, Van Steen K (2011) A detailed view on Model-Based Multifactor Dimensionality Reduction for detecting gene-gene interactions in case-control data in the absence and presence of noise . (invited paper of K Van Steen as expert entry to special issue on epistasis). *Ann Hum Genet.* 2011 Jan;75(1):78-89
- Chanda P.; Zhang A.; Brazeau D.; Sucheston L.; Freudenheim J.; Ambrosone C.; Ramanathan M. (2008), Information-Theoretic Metrics for Visualizing Gene-Environment Interactions, *Am J Hum Genet* 81(5): 939-963
- Dixon, M. S.; Golstein, C.; Thomas, C. M.; van Der Biezen, E. A. & Jones, J. D. (2000), Genetic complexity of pathogen perception by plants: the example of Rcr3, a tomato gene required specifically by Cf-2. *Proc Natl Acad Sci U S A* 97(16), 8807-8814.
- Mahachie John JM, Cattaert T, Van Lishout F, Van Steen K (2011) Model-Based Multifactor Dimensionality Reduction to detect epistasis for quantitative traits in the presence of error-free and noisy data – accepted in *EJHG*.
- Manolio, T. A.; Brooks, L. D. & Collins, F. S. A (2008) HapMap harvest of insights into the genetics of common disease. *J Clin Invest* 118(5), 1590-1605
- Ritchie, M. D. et al. Multifactor-dimensionality reduction reveals high-order interactions amongestrogen-metabolism genes in sporadic breast cancer (2001). *Am. J. Hum. Genet.* 69, 138–147.
- Schwartz, D. F.; Ziegler, A. & König, I. R. (2008) Beyond the results of genome-wide association studies. *Genet. Epidemiol.* 32, 671.
- Van Steen K.; Laird N.M.; Markel P.; Molenberghs G. (2007), Approaches to handle incomplete data in family-based association testing. *Ann Hum Genet.* Mar;71(Pt 2):141-51.
- Van Steen K.; Mcqueen M.B.; Herbert A.; Raby B.; Lyon H.; Demeo D.L.; Murphy A.; Su J.; Datta S.; Rosenow C.; Christman M.; Silverman E.K.; Laird N.M.; Weiss S.T.; Lange C. (2005), Genomic screening and replication using the same data set in family-based association testing, *Nature Genetics* 37 (7), 683 – 691.
- Van Steen K (2011a) Perspectives on Large-Scale Multi-Stage Family-Based Association Studies – invited submission to *Stats in Med* - accepted
- Van Steen K (2011b) Travelling the world of gene-gene interactions – invited submission to *Briefings in Bioinformatics* – accepted

Disclosure

The MB-MDR team consists of two research groups with PI's K Van Steen (Montefiore, Belgium) and M Calle (UVic, Spain).

In general, new methodologies are developed under the umbrella of the University of Liège (Montefiore Institute / Systems and Modelling Research Unit).