SMPGD 2017:
Statistical Methods for Postgenomic Data

Imperial College London
12-13 January 2017

https://smpgd2017.wordpress.com

Conference Chairs:
Korbinian Strimmer and Marc Chadeau-Hyam
k.strimmer@imperial.ac.uk and m.chadeau@imperial.ac.uk

Epidemiology and Biostatistics, School of Public Health,
Imperial College London, Norfolk Place, London W2 1PG

Abstract Book and Programme

52 presentations:

4 Keynote talks,
9 Invited talks,
12 Contributed talks, and
27 Poster presentations

Document version: 9 January 2017
Day 1
Imperial College London - St. Mary's campus, Paddington (north of Hyde Park)

Theme 1: Big Data in Biology, Medicine and Health, and Precision Medicine

9:00 Opening: Korhbran Stimmer (ICL), Marc Chackaou-Hyam (ICL) Welcome
9:10 Keynote 1: Neil Lawrence (Amazon Cambridge, Sheffield) Bayesian inference in high-dimensional seemingly unrelated regressions, Applied to metabolomics data
10:00 Invited 1: Alex Lewin (Brunel) Stratification of patient trajectories using covariate latent variable models
10:30 Contributed: Kieran Campbell (Oxford) Target-averaged linear shrinkage: high-dimensional covariance matrix estimation In functional genomics
10:50 Coffee Break
11:10 Invited 2: Mark van de Weel (VU and VUMC Amsterdam) Fast computation of genome-metagenome interaction effect
11:40 Contributed: Verena Zuber (EBI Hinxton) Two-dimensional fine-mapping of global and molecular traits to identify gene-disease linkages
12:00 Contributed: Christophe Ambrozio (Evry) Fast computation of genome-metagenome interaction effect
12:15 Lunch Break

Theme 2: Bayesian Biostatistics and Machine Learning in Systems

13:40 Keynote 2: Pierre Alquier (ESIENS Paris Saclay) On the properties of variational approximations of posteriors and pseudo-posteriors
14:30 Invited 3: Maria De Oriol (UCL) Bayesian inference for multiple Gaussian graphical models with application to metabolic association networks
15:00 Contributed: Harry Gray (MRC BSU Cambridge) Target-averaged linear shrinkage: high-dimensional covariance matrix estimation In functional genomics
15:20 Coffee Break
15:40 Invited 4: Paul Kirk (MRC BSU Cambridge) MDI: Multiple dataset integration using Bayesian correlated clustering
16:10 Contributed: Dennis de Beer (VUMC Amsterdam) Using co-data to improve the predictive accuracy of a random forest
16:30 Invited 5: Silvia Liberani (Brunel) Modeling colinear and spatially correlated data
17:00 Break – walk to poster presentations

17:10 Poster Presentations - All Topics (with refreshments)

1. Mikael Andén (AstraZeneca Cambridge) Estimation of copy number variation calling uncertainty in NGS data
2. Lisa Almheim (Heinrichstr Munich) On the properties of variational approximations of posteriors and pseudo-posteriors
4. Alessandra Cobassi (MRC BSU Cambridge) target-averaged linear shrinkage: high-dimensional covariance matrix estimation in functional genomics
5. Ali Delahaye-Duriez (ICL, Paris Diderot, Sorbonne) Mining big data for environmental epidemiological analyses
6. Mathieu Emily (Rennes) Providing opportunities for novel antiepileptic drug discovery
7. Daniele Fecht (ICL) Rare and common epilepsies converge on a shared genetic network
8. Anna Frei (ICL) Providing opportunities for novel antiepileptic drug discovery
9. Daniel Grema (MRC BSU Cambridge) Single nucleotides the right scale for genome-wide association studies
10. Florlent Guinet (Evry) Single nucleotides the right scale for genome-wide association studies
11. Mathew Hall (Oxford) Novel SNV as variable selection method for classification: an application to SNP data
12. Nazatulshima Hasan (MRC BSU Cambridge) Fast computation of genome-metagenome interaction effect
13. Boris Hejblub (Bordeaux) Representation learning for complex network analysis
14. Eoin Holmes (Tomek, Paris Descartes) Predicting breast cancer metastasis from blood gene expression using time-to-diagnosis data
15. Ayssa Imbert (INRA Toulouse) To derive signature estimates
16. Antonella Iuliano (Naples) Estimation of copy number variation calling uncertainty in NGS data
17. Holger Kirsten (Leipzig) From low pass and low tumour purity samples
18. Erk Krause (ICL) Metabolic effects of environmental pollutants in pregnant women - an exposome approach
19. Norbert Krautenbacher (Helmholtz Munich) Multiple hot-deck imputation for network inference from RNA-sequencing data
20. Magnus Leiden (Leiden, VUMC Amsterdam) Network-based dimension reduction methods for the integrative analysis
21. Miguel Pereira (ICL) Assessing susceptibility to breast cancer in young adult women
22. Morgane Pierre-Just (Evry) Of multi-omics data in cancer
24. Olivier Robin (ICL) From low pass and low tumour purity samples
25. Tammo Ruedi (Oxford) With interpretable latent representations
26. Benjamin Sadacca (Curie Paris) Metabolic effects of environmental pollutants in pregnant women - an exposome approach
27. Virginie Stanislas (Evry) The OrMachine, a model for dimensionally reduction of binary data
18:40 End of post poster session
19:00 Conference dinner (various restaurants nearby)

Day 2
Imperial College London – main campus, South Kensington (south of Hyde Park)

Theme 3: Computational Epidemiology and Evolutionary Models

9:30 Keynote 3: Christiane Thieme (MRC BSU Cambridge) Linkage-free inference and predictions for computational epidemiology
10:00 Invited 6: Philippe Lemey (Laurens) Connecting sequence and trait evolution in a Bayesian phylogenetic framework
10:30 Contributed: Erik Völl (ICL) Phylogenetic inference across epidemic scales
10:50 Coffee Break
11:10 Invited 7: Tim Ebbels (ICL) Power and sample size determination in metabolomics
11:40 Contributed: Issa Davies (Glasgow) Sparse hierarchical Bayesian models for joint phylogenetic inference
12:00 Contributed: Florian Robert (Rennes) Sparse hierarchical Bayesian models for joint phylogenetic inference
12:20 Lunch Break

Theme 4: Systems Biology and Networks

13:40 Keynote 4: Benoît Schachow (Pasteur Paris) LEAN discovery of hotspots in networks
14:30 Invited 8: Mark van de Weel (VU and VUMC Amsterdam) Bayesian nonparametric approaches to quantifying dependence between random variables
15:00 Contributed: Séverine Affie (Curie Paris, Sorbonne, Descartes) Learning causal networks with latent variables from multivariate data in genomics
15:20 Coffee Break
15:40 Contributed: Wessel van der Wijden (VU and VUMC Amsterdam) A tale of two networks
16:00 Contributed: Ewen Crawford (MRC BSU Cambridge) Gene network reconstruction using global/local shrinkage priors
16:20 Contributed: Genevieve Robin (Ecole Polytechnique Paris) Graphical model inference with unobserved variables via latent tree aggregation
16:40 Invited 9: Lorenzo Wernisch (MRC BSU Cambridge) Gene regulatory networks from single cell data
17:00 End of meeting

Further (co)authors and committee members present at workshop:
Marco Barterle (Brunel)
Magalie Berland (INRA, Jouy-en-Josas)
Etienne Birmeke (Paris Descartes)
Marta Biancardi (ICL)
Isabel Brito (Curie Paris)
David Causeur (Rennes)
Emma Ehlers (Bruge)
Sylvie Huett (INRA, Jouy-en-Josas)
Corinna Hine (ICL)
Franck Picard (Lyon)
Stéphane Robin (INRA, AgroParisTech, Paris)
Michael Stumpf (ICL)
Ernest Turo (CRUK Cambridge)
Nathalie Villa-Vivenex (INRA Toulouse)
Christopher Yiu (Birmingham, Oxford)

https://smpgd2017.wordpress.com
Estimation of copy number variation calling uncertainty in NGS data from low pass and low tumour purity samples

Copy number variation (CNV) estimation is one of the important goals of high throughput sequencing (HTS) of tumour samples. Given the variation in tumour purity and depth of sequencing, it is imperative to understand the uncertainty in the estimated copy numbers. In this presentation we look at the variation in CNV estimates and visualise how it's a function of depth, tumour purity, width of the event and capture type. Results from cell lines with known CNVs are shown for both whole genome sequencing as well as targeted capture, with good resolution down to 30% tumour purity.
Learning causal networks with latent variables from multivariate information in genomic data

Constraint-based network reconstruction methods can in principle uncover causality from purely observational data, without assuming causal or non-causal a priori models. Advanced constraint-based methods reconstruct a broad class of "ancestral graphs", that include undirected (-), directed (-->), and bidirected (<-->) edges, the latter corresponding to latent common causes, L, unobserved in the available data (i.e. <--L-->). However, classical constraint-based methods are not robust on small datasets and have algorithmic complexity issues, which have so far limited their applicability in biology.

We report a novel information-theoretic method which learns a large class of graphical models from purely observational data by simultaneously circumventing the complexity and robustness issues of constraint-based algorithms. Our approach unifies causal and non-causal network learning frameworks while including the effects of unobserved latent variables. Starting from a complete graph, it iteratively removes dispensable edges, by uncovering significant multivariate information contributions from indirect paths, and assesses edge specific confidences from randomization of available data. The remaining edges are then oriented based on the signature of causality in observational data.

The approach outperforms traditional constraint-based methods on a broad range of real-life and simulated benchmark networks. It achieves significantly better results with much fewer samples and is typically ten to hundred times faster than the most advanced constraint-based methods. The complexity of our algorithm exhibits in fact optimal scalings, linear in terms of sample size and quadratic in terms of network size for sparse graphs. In addition, this novel approach does not need to make an a priori choice on the type of causal or non-causal graphical model. The most appropriate model is directly learned from the unified framework based on the available data. In particular, our method is found to accurately discard spurious causality even from relatively small datasets, as can be encountered with experimental results.

We have applied this novel reconstruction method to learn causal networks that include latent variables from a variety of genomic datasets at different biological size and time scales, from gene regulation in single cells to whole genome duplication in tumor development as well as long term evolution of vertebrates.

Coauthors: Louis Verny, Nadir Sella and Hervé Isambert
Keynote

Pierre Alquier
ENSAE Paristech
University Paris Saclay
Pierre.Alquier@ensae.fr

On the properties of variational approximations of posteriors and pseudo-posteriors

Bayesian estimators, and their analog in machine learning called EWA (Exponentially Weighted Aggregate) are usually computed through Monte Carlo methods. However, in many practical applications, the computational cost of Monte Carlo methods is prohibitive. It is thus tempting to replace these by (faster) optimization algorithms that aim at approximating the posterior: we will refer to these methods as variational Bayes (VB) methods.

In this talk I will show, thanks to a PAC-Bayesian theorem, that VB approximations are well founded, in the sense that the loss incurred in terms of prevision risk is negligible in some classical settings such as linear classification, ranking... These approximations are implemented in the R package pac-vb (written by James Ridgway) that I will briefly introduce, and apply to real datasets. I will discuss deeply the meaning of the PAC-Bayesian theorem in order to explain how this result can be extended to other settings.

Coauthors: Nicolas Chopin and James Ridgway
Fast computation of genome-metagenome interaction effect

Association studies usually search for association between common genetic variants in different individuals and a given phenotype. In this work we do consider two types of biological markers: genetic and metagenomic markers. The genotypic markers allow to characterise the individual by its inherited genetic information whereas the metagenomic markers are related to the environment. Both types of markers are available by millions and represent a unique signature characterizing each individual.

We focus on the detection of interaction between groups of metagenomic and genetic markers in order to better understand the complex relationship existing between environment and genome in the expression of a given phenotype. The proposed method first reduces the dimension of the search space by selecting a subset of super variables in both complementary data sets. The dimension reduction step combines hierachical clustering, which defines super variables, and Lasso, which selects relevant super variables. Relevant interactions are then explored via linear model testing.

The proposed approach is illustrated by means of both synthetic and real data.

Coauthors: Julien Chiquet, Florent Guinot and Marie Szafranski
Acute Myeloid Leukemia (AML) is a type of blood cancer affecting the myeloid lineage. The incidence of AML increases with age, and it is the most frequent type of acute leukemia among adults. Although approximately 70% of patients achieve complete remission, very small numbers of leukemic cells remain and cannot be detected with current diagnostic techniques. Nearly everybody with AML will relapse in the end if no further postremission or consolidation therapy is given, and this relapse is almost always lethal.

AML patients frequently carry a mixture of different cancer cell types, so-called subclones, which evolve over time, so that the mixture at relapse is different from the one at diagnosis. Understanding clonal evolution and identifying rare subclones, especially for those mutations causing relapse, is still an open challenge.

We aim to parameterize transcriptional heterogeneity from RNA-Seq counts taken from small groups of cells (e.g., 10-cells). To that end, we will extend our Stochastic Profiling Method previously proposed for microarray data. This technique infers single-cell regulatory states by mathematically deconvolving n-cell measurements. This averaging-and-deconvolution approach allows us to quantify single-cell regulatory heterogeneities while avoiding the technical measurement noise of single-cell techniques.

Coauthor: Christiane Fuchs
Variational inference for semi-supervised clustering of ’omics data

Clustering remains a fundamental analysis technique in genomics and precision medicine. However, with high-dimensional genomics datasets, there are typically many plausible ways to cluster the data, and the best way to identify the most relevant clustering structure(s) is still an open question. For this reason, it can be useful to take into account additional "side information" related to the problem of interest (e.g. GO terms when clustering genes, or survival data when clustering patients), which can help us to find the most relevant structure in the data. Traditional approaches only make use of this information after the clustering has been performed.

Here we extend traditional mixture modelling, incorporating ideas from “profile regression” (Molitor et al., 2010), in order to allow us to exploit "side information" to guide the clustering process. We fit our models using variational inference, which greatly improves scalability and allows large datasets to be considered. We apply these methods to genomic datasets, where we use phenotypic traits as side information and demonstrate that these can help to provide insight into the data structure.

Coauthor: Paul Kirk
Stratification of patient trajectories using covariate latent variable models

Standard models assign disease progression to discrete categories or stages based on well-characterized clinical markers. However, such a system is potentially at odds with our understanding of the underlying biology, which in highly complex systems may support a (near-)continuous evolution of disease from inception to terminal state. To learn such a continuous disease score one could infer a latent variable from dynamic omics data such as RNA-seq that correlates with an outcome of interest such as survival time.

However, such analyses may be confounded by extraneous factors measured by clinical covariates in electronic health records (EHRs). As a solution to this we introduce covariate latent variable models, a novel type of latent variable model that learns a low-dimensional data representation in the presence of two (asymmetric) views of the same data source. In this model the factor loading matrix of one set of observables (such as gene expression) is “perturbed” by a second set (such as mutation status). A hierarchical Bayesian prior structure identifies genes whose behaviour along such patient trajectories is dependent on the covariate under study, identifying novel molecular interactions.

We apply our model to TCGA colorectal cancer RNA-seq data and demonstrate how incorporating microsatellite-instability (MSI) and metastatic status as external covariates allows us to identify genes that stratify patients on an immune-response trajectory. Finally, we propose an extension termed Covariate Gaussian Process Latent Variable Models for learning nonparametric, nonlinear representations.

Coauthor: Christopher Yau
AriCode an R package to efficiently compute rand-index and other clustering comparison information measures

Measures for clustering comparison – e.g. the adjusted rand-index (ARI) or the normalized information distance (NID), see [1] – are essential either to assess the quality of a clustering solution according to some ground truth or more generally to find or to select a robust and stable clustering solution [2].

Consider two clusterings C1 and C2 in respectively c1 and c2 classes of n data-points. For most measures (including ARI and NID), one needs to count the number of samples that are simultaneously in class j of C1 and class k of C2 for all c1c2 possible (j, k)-pairs. A naive calculation of these measures is typically in order O(n + c1 c2) both in space and time. This is prohibitive when considering very large clustering problems in which c1c2 > n.

Here we describe a faster algorithm based on bucket sorting [3], which is in O(n) both in space and time. We implemented the code in an R package named ariCode, which is typically one order of magnitude faster than standard R implementations and easily copes with large data sets. We demonstrate the utility of our package in the context of metagenomics where n, c1 and c2 are typically very large. We also show how these measures can be used to select the number of clusters by assessing the robustness of the clustering thanks to resampling or random projection approaches.


Coauthors: **Valentien Dervieux** and **Guillem Rigaill**
Likelihood-free inference and predictions for computational epidemiology

Simulator-based models often allow inference and predictions under more realistic assumptions than those employed in standard statistical models. For example, the observation model for an underlying stochastic process can be more freely chosen to reflect the characteristics of the data gathering procedure. A major obstacle for such models is the intractability of the likelihood, which has to a large extent hampered their practical applicability.

I will discuss recent advances in likelihood-free inference that greatly accelerate the model fitting process by exploiting a combination of machine learning techniques. Applications to several models in infectious disease epidemiology are used to illustrate the potential offered by this approach.
Sparse hierarchical Bayesian models for detecting relevant antigenic sites in virus evolution

Emerging viral diseases pose a substantial threat to public health. Understanding how virus strains offer protection against closely related emerging strains is vital for creating effective vaccines. For many viruses, including the Influenza virus where multiple serotypes often co-circulate, in vitro testing of large numbers of vaccines can be infeasible. Therefore the development of an in silico predictor of cross-protection between strains is important to help optimise vaccine choice. Vaccines will offer cross-protection against closely related strains, but not against those that are antigenically distinct. To be able to predict cross-protection we must understand the antigenic variability within a virus serotype, distinct lineages of a virus, and identify the antigenic residues and evolutionary changes that cause the variability.

In the current work, we use structural and phylogenetic differences between pairs of virus strains to identify important antigenic sites on the surface of the influenza A(H1N1) and A(H3N3) viruses through the prediction of haemagglutination inhibition (HI) assay, a pairwise measure of the antigenic similarity of virus strains. Previously (Davies et al., JMLR W&CP 33:149-158, 2014) we have achieved this using a Sparse hierArchical Bayesian model for detecting Relevant antigenic sites in virus Evolution (SABRE). While the SABRE method can account for the experimental variability in the data and select the variables responsible for the changes in the HI assay via spike and slab priors, it does not fully account for the structure of the dataset. In particular, for any HI assay measurement between the same pair of virus strains the difference in the viral sequence remains the same. Here we propose an extension to the SABRE method (eSABRE) which uses additional variables to represent the underlying HI assay measurement of any given pair of virus strains. This more accurately represents the data in the model while also giving a significant computational improvement. In this work we have applied the eSABRE method to a real H1N1 and H3N2 datasets, identifying some of the key antigenic sites with those serotypes.

Coauthors: Richard Reeve, William T. Harvey and Dirk Husmeier
Bayesian inference for multiple Gaussian graphical models with application to metabolic association networks

We investigate the effect of cadmium (a toxic environmental pollutant) on the correlation structure of a number of urinary metabolites using Gaussian graphical models (GGMs). The inferred metabolic associations can provide important information on the physiological state of a metabolic system and insights on complex metabolic relationships. Using the fitted GGMs, we construct differential networks, which highlight significant changes in metabolite interactions under different experimental conditions. The analysis of such metabolic association networks can reveal differences in the underlying biological reactions caused by cadmium exposure.

We consider Bayesian inference and propose using the multiplicative (or Chung-Lu random graph) model as a prior on the graphical space. In the multiplicative model, each edge is chosen independently with probability equal to the product of the connectivities of the end nodes. This class of prior is parsimonious yet highly flexible; it can be used to encourage sparsity or graphs with a pre-specified degree distribution when such prior knowledge is available. We extend the multiplicative model to multiple GGMs linking the probability of edge inclusion through logistic regression and demonstrate how this leads to joint inference for multiple GGMs. A sequential Monte Carlo (SMC) algorithm is developed for estimating the posterior distribution of the graphs.
Rare and common epilepsies converge on a shared gene regulatory network providing opportunities for novel antiepileptic drug discovery

The relationship between monogenic and polygenic forms of epilepsy is poorly understood, and the extent to which the genetic and acquired epilepsies share common pathways is unclear. Here, we undertook an integrated systems-level analysis of brain gene expression data to identify molecular networks disrupted in epilepsy.

We identified a co-expression network of 320 genes (M30), which is significantly enriched for non-synonymous de novo mutations ascertained from patients with monogenic epilepsy, and for common variants associated with polygenic epilepsy. The genes in M30 network are expressed widely in the human brain under tight developmental control, and encode physically interacting proteins involved in synaptic processes. The most highly connected proteins within M30 network were preferentially disrupted by deleterious de novo mutations for monogenic epilepsy, in line with the centrality-lethality hypothesis. Analysis of M30 expression revealed consistent down-regulation in the epileptic brain in heterogeneous forms of epilepsy including human temporal lobe epilepsy, a mouse model of acquired temporal lobe epilepsy, and a mouse model of monogenic Dravet (SCN1A) disease. These results suggest functional disruption of M30 via gene mutation or altered expression as a convergent mechanism regulating susceptibility to epilepsy broadly. Using the large collection of drug-induced gene expression data from Connectivity Map, several drugs were predicted to preferentially restore the down-regulation of M30 in epilepsy toward health, most notably valproic acid, whose effect on M30 expression was replicated in neurons.

Taken together, our results suggest targeting the expression of M30 as a potential new therapeutic strategy in epilepsy.

Coauthors: Prashant Srivastava, Kirill Shkura, Sarah R. Langley, Liisi Laaniste, Aida Moreno-Moral, Bénédicte Danis, Manuela Mazzuferi, Patrik Foerch, Elena V. Gazina, Kay Richards, Steven Petrou, Rafal M. Kaminski, Enrico Petretto and Michael R. Johnson
Power and Sample Size Determination in Metabolomics

Estimation of statistical power and sample size is a key aspect of experimental design. However, in metabolomics, there is currently no accepted approach for these tasks, in large part due to the unknown nature of the expected effect. In this hypothesis-free approach, we do not know the number or identity of important metabolites, nor do we know the effect size a priori. We introduce an approach (Blaise et al. 2016), based on multivariate simulation, which deals effectively with the highly correlated structure and high-dimensionality of metabolomic data. First, a large data set is simulated based on the characteristics of a pilot study, investigating a given biomedical question. An effect of a given size, corresponding either to a discrete (classification) or continuous (regression) outcome is then added. Different sample sizes are modeled by randomly selecting data sets of various sizes from the simulated data. We investigate different methods for effect detection, including univariate and multivariate techniques. Our framework allows us to investigate the complex relationship between sample size, power and effect size for real metabolomic data sets. For instance, we demonstrate for an example pilot data set, that certain features achieve a power of 0.8 for a sample size of 20 samples, or that a cross-validated predictivity Q2Y of 0.8 is reached with an effect size of 0.2 and 200 samples. We exemplify the approach for both Nuclear Magnetic Resonance (NMR) and Liquid Chromatography – Mass Spectrometry (LC-MS) data from humans and the model organism C. elegans.


Coauthors: see reference.
Gene-based gene-gene interaction with R

Among the statistical methods that have been proposed to identify gene-gene interactions in case-control Genome-Wide Association Studies, gene-based methods have recently grown in popularity as they confer advantages both in statistical power and biological interpretation. Since 2009, several methods have been proposed to model the joint distribution of SNPs within and between the two genes using principal component analysis (PCA - Li et al. 2009), canonical correlation analysis (CCU - Peng et al. 2010), kernel canonical correlation U-based statistic (KCCU - Larson et al. 2014), Partial Least Squares Path Modeling (PLSPM - Zhang et al., 2013), Composite Linkage Disequilibrium (CLD - Rajapakse et al., 2012) as well as information theory (GBIGM - Li et al., 2015). Rather than considering multiple markers in both genes as part of a joint model, an alternative strategy has recently been developed in order to aggregate p-values obtained at the SNP level into a test at the gene level (Aggregator, Emily, 2016).

All these methods have been implemented in home made softwares that are for most of them available only on request to the authors and at best have a web interface. Thus, searching for gene-gene interaction at the gene level is not straightforward. Furthermore, a comprehensive comparison of such methods, in terms of power and computational performances, remains hardly feasible. To overcome these issues, we propose in this work a novel R package which provides a user-friendly tool for gene-based gene-gene interaction in case-controls association studies.

Our package proposes a set of functions that allow to (1) download data in various standardized format (PED, PLINK, VCF), (2) impute missing genotypes, (3) perform a gene-based gene-gene interaction analysis and (4) visualize the results. To demonstrate the efficiency of our package, we first performed an extensive simulation study with respect to various disease models and different gene correlation structures. Although our results show a large heterogeneity among the different methods, the aggregation of p-values is the most powerful method in many situations. Furthermore, the analysis of the true phenotype in the dataset GSE39428 gives also new insight in the understanding of the etiology of Rheumatoid Arthritis, thus paving the way for further investigation of gene-gene interaction at the gene level.


Coauthor: Magalie Houée-Bigot
Mining big data for environmental epidemiological analyses

Traditionally, environmental epidemiology analysis has aimed to analyse associations between one or multiple pollutants and a single health outcome. Recent advances in computing power and statistical methodology have made multi-level analysis of complex disease aetiologies realistic, but so far, “big data” health analyses have mostly focused on the analysis of -omics data.

We aim to agnostically investigate the associations between air pollution exposure (NO2 and PM10) and all hospital admissions in England (2009-2012) using a range of biostatistical techniques including clustering and classification approaches that easily scale to a large number of participants. The SAHSU health databases contain well over half a billion routinely collected health records with over 500 million in NHS Hospital Episode Statistics (HES) alone. This readily available data source within The Small Area Health Statistics Unit (SAHSU) offers a unique and exciting opportunity to go beyond traditional environmental epidemiological analysis and simultaneously investigate a multitude of exposures and health outcomes.

Applying data mining techniques to analyse health outcomes associated with a common environmental exposure may help validate previously identified associations and identify new combinations of health effects. Potential clustering of certain health outcomes, both within and between individuals, could help to elucidate potential mechanistic pathways involved in the association between health and environmental exposure, highlight areas of future research needs and generate new hypotheses.

Coauthors: Rebeca Ghosh, Fred Piel and Marc Chadeau-Hyam
Bayesian nonparametric approaches to quantifying dependence between random variables

Nonparametric and nonlinear measures of statistical dependence between pairs of random variables are important tools in modern data analysis. In this talk, I will present a Bayesian nonparametric procedure that leads to a tractable, explicit and analytic quantification of the relative evidence for dependence vs. independence.

Our approach uses Pólya tree priors on the space of probability measures which can accommodate known uncertainty in the form of the underlying sampling distribution and provides an explicit posterior probability measure of both dependence and independence. Well known advantages of having an explicit probability measure include: easy comparison of evidence across different studies; encoding prior information; quantifying changes in dependence across different experimental conditions, and the integration of results within formal decision analysis.

I will apply our procedure to the construction of gene expression network from measurements at single-cell resolution as well as for a differential co-expression analysis.

Coauthor: Chris Holmes
Concentrations of different traffic-related air pollutants are often highly correlated. Due to the inadequacy of classical statistical methods to analyse joint effects of correlated covariates, different air pollutants are commonly analysed separately. In this study, we investigated the joint effect of air pollution concentrations, socio-economic factors and the relative risk of cardiovascular mortality, in London and Oxford.

We used a Bayesian approach: Profile Regression mixture model, using Dirichlet Process, to identify clusters, composed of geographical areas, of similar concentrations profiles. Then we used this cluster membership in a multi-level risk model as random effect, with confounders and spatial structured effects. For each area we computed pollutant concentrations maps for PM10, PM2.5, PM.coarse, NO2, NOx based on European Study of Cohorts for Air Pollution effects (ESCAPE) land use regression models, and collected Cardiovascular mortality counts compiled for 2008-2011, together with tobacco expenditure, percentage of Asian population and a measure of deprivation (Index of Multiple Deprivation, 2010).

The study area is comprised by 7,849 Lower Layer Super Output Areas (LSOA). The profile regression mixture model (embedded in PReMiU M R package) can identify the cluster profiles and compute the risk estimate associated with the LSOAs, while adjusting for confounders jointly. However, because of the high-dimensionality in the spatial area the algorithm is crashing when the spatial component is included in the model. Therefore at this stage, with no proper way to estimate such model, we had to break it down in a two step-process. The first make use of the Regression profile for identifying the cluster and the second part make use of INLA capabilities in holding fast spatial computations. The drawback is that uncertainty associated with cluster membership is partially lost. In future, new methods to deal with Bayesian complex model should be develop.

In this study, using profile regression, we were able to identify clusters of geographical areas characterized by joint patterns of risks and profile exposure, areas drifting from global trend and evaluate information of each single variable, unlike classical approaches which combine correlated variables in a unique score.

Couauthors: David Morley, Daniela Fecht, Marta Blangiardo, Anna Hansell and John Molitor
Target-averaged linear shrinkage: high-dimensional covariance matrix estimation in functional genomics

It is well-known that the maximum likelihood estimator (MLE) of the covariance matrix of a multivariate Gaussian distribution is singular/ill-conditioned when the number of variables in a dataset is similar to, or exceeds, the number of samples. Single-Target linear Shrinkage (STS), which weights the MLE with a predefined positive definite target matrix, is now well-established as a computationally efficient solution to this problem. Much of the literature in STS has focused upon optimising the weighting between the MLE and the target in order to maximise mean squared-error reduction [1, 2]. However, the magnitude of this reduction relies heavily upon the selected target matrix.

We present Target-Averaged linear Shrinkage (TAS), an approach to incorporate multiple target matrices in linear shrinkage estimation within a conjugate Bayesian framework. This is attractive both as a way of including statistical uncertainty and as a means of including multiple sources of prior information in the estimation procedure. We show through simulation that TAS outperforms state-of-the-art STS methods. We use protein expression data from 17 different cancer types from The Cancer Proteome Atlas to demonstrate the use of prior information in TAS.

We conclude by discussing the selection of target matrices for situations with or without available prior information. We also discuss the extension of TAS to more general likelihood models and present a future application to the Joint Analysis of Marginal SNP effects [3] in Genome Wide Association Studies.


Coauthors: Gwenaël Leday, Catalina Vallejos and Sylvia Richardson
BeviMed: Bayesian evaluation of variant involvement in Mendelian disease

Rare diseases are often caused by high-penetrance rare variants. Because of genetic heterogeneity and low numbers of cases, statistical procedures testing each variant for marginal association with phenotype are often underpowered. Procedures which aggregate rare variants across regions also sacrifice power because they are likely to include non-pathogenic rare variants in the aggregation. Therefore, models are required which account for a mixture of pathogenic and non-pathogenic rare variants explicitly. Ideally, the composition of this mixture would be informed by additional information such as sequence conservation across species and population allele frequency. It is also important to model the modes of inheritance typical in Mendelian disease, including autosomal dominant, autosomal recessive, and X-linked recessive inheritance.

We present a new method, ‘Bayesian Evaluation of Variant Involvement in Mendelian Disease’ (BeviMed), which assesses the evidence of association between a case-control label and presence of one or more genetic configurations of alleles at rare variant sites. The configurations depend on a latent partitioning of variants into pathogenic and non-pathogenic groups and can be informed by external information. Different modes of inheritance are modelled by conditioning on the number of pathogenic alleles carried by each individual (and gender, in the case of X-linked inheritance). Thus compound heterozygosity and homozygosity can be treated equivalently. Inference is performed by obtaining posterior probabilities of the different models of association and a baseline model of no association.

Our approach performs as well or better than existing methods in terms of sensitivity and specificity at the region level (which could correspond to a gene) and at the individual variant level. BeviMed can analyse 10,000 samples genotyped at 100 rare variant sites with minor allele frequencies averaging 2.5% in under a second, enabling fast genome-wide rare variant evaluation. We show the results of analysing a dataset of 5,000 whole-genome sequenced samples from patients with diverse rare diseases.

Coauthors: Sylvia Richardson and Ernest Turro
Are single nucleotides the right scale for Genome-Wide Association Studies?

Genome-Wide Association Studies (GWAS) aim to identify causal genomic variants implied in the expression of rare human diseases. From a statistical point of view, detecting these variants imply to perform hypothesis tests on the population subject to the rare disease (cases) against an healthy population (controls) at several locus on the genome. One individual's genome being characterized by hundreds of thousands of SNPs, the type I error, resulting from a large number of hypothesis tests, can dramatically increase and lead to wrong conclusions about genetic associations with the disease.

Dimension-reduction methods are a way to improve the detection of true genetic associations by reducing the number of hypothesis to test. We thus propose a new dimension-reduction approach which can be applied in the context of GWAS by taking benefit of the structure in haplotype of the human genome.

For this purpose, we first cluster the SNPs with an hierarchical clustering algorithm using the linkage disequilibrium as a measure of dissimilarity. We then propose to:

1. Compress each cluster in the hierarchy into a unique variable which is built to reflect the number of mutation of each SNPs for each individuals compared to an average (most frequent) genotype.
2. Choose an appropriate cut into the tree with a ridge regression procedure.

Replacing the initial genomic matrix with a lower-dimension predictors matrix allow us to perform fewer hypothesis tests and detect associations between the phenotype and clusters of SNPs with a much lower type I error.
Phyloscanner: evolutionary and phylodynamic analysis of NGS viral genome data

Use of pathogen sequence data to answer evolutionary and epidemiological questions has become a well-established area of study in recent years, a field that has been dubbed "phylodynamics". Next-generation sequencing (NGS) technologies have now arrived, providing access to genomic data on a much deeper scale than was possible using older methods. The output is a much richer sample of within-host genetic diversity, but as a set of short reads rather than of full genomes. As a first step, data of this type can be used to generate consensus sequences for input in existing phylogenetic and phylodynamic methodologies, but it is preferable to use all the information contained in the short reads to inform inference. This requires new methodological approaches.

Here, we present phyloscanner, a tool for phylogenetic analysis of NGS viral short read data. The viral genome is divided into short windows, and maximum-likelihood phylogenies generated using the set of short reads that overlap with each window. Within each window, the topological arrangement of the reads from each patient, and well as branch lengths, are then examined to yield insights into, for example, possible transmission pairs, the direction of transmission between patients, the presence of multiple infections, and recombination. Individual per-window observations are then summarised over the entire genome to give final results. We demonstrate the tool using sequence data from HIV seroconverters in the BEEHIVE study.

Coauthors: Chris Wymant, Oliver Ratmann and Christophe Fraser
Novel tSNR as variable selection method for classification: an application to SNP data

Signal-to-Noise Ratio (SNR) is defined as the ratio of the squared amplitude or variance of a signal relative to the variance of the noise. We extend the concept of SNR by calculating the ratio of deviance from a Generalised Linear Model (GLM) as the new ranking measure for Single Nucleotide Polymorphism (SNP) selection. The selected SNPs are then used as the predictors for a classifier. We measure the efficiency of the method by calculating the sensitivity, specificity and area under ROC curve (AUC).
Block testing approach in Genome-Wide Association Studies using a multilevel modeling of the dependence structure

A Genome-Wide Association Study (GWAS) consists in testing the association between Single Nucleotide Polymorphisms (SNPs) and a phenotype, as for example disease status in case/control GWAS. To cover the entire genome, the number of SNPs has to be very large inducing a tremendous number of statistical tests. Thus a big challenge in the interpretation of GWAS is the evaluation of the statistical significance level, for which multiple testing adjustments are commonly performed to either control the family-wise error rate (FWER) or the false discovery rate (FDR).

However SNPs data are known to be dependent because of the block structure of the genome (Conneely and Boehnke, 2007). Dependence between SNPs leads to correlation between statistical tests and seriously affects the consistency of SNP ranking (Friguet et al., 2009), thus highlighting the importance to account for dependence in the correction for multiple testing.

In this work, we therefore propose an original approach to account for the dependency between single-marker tests. Based on the block structure of the genome, our aim is to propose a two-level correction approach defined as a combination of a within-block correction and a between-block correction. Regarding the within-block correction, we focused on a multivariate gaussian modeling of the set of corresponding test statistics. For that purpose, we proposed (1) extensions of existing methods, such as a minP test (Conneely and Boehnke, 2007) or the Versatile Gene-based Association test (Liu et al., 2010), and (2) the application of whitening and decorrelation techniques.


Coauthors: Mathieu Emily and David Causeur
Sequential Dirichlet process mixtures of multivariate skew t-distributions for model-based clustering of flow cytometry data

Flow cytometry is a high-throughput technology used to quantify multiple surface and intracellular markers at the level of a single cell. This allows to identify cell sub-types, and to count the number of cells sampled for each sub-type. Improvements of this technology lead to the ability of describing millions of individual cells from a blood sample using multiple markers. This results in high-dimensional datasets, whose manual analysis is highly time-consuming and poorly reproducible. Several methods have been developed to perform automatic recognition of cell populations. Most of them are suited for the analysis of a single sample from one patient. However in reality, repeated measurements are often available.

We propose to use a Bayesian nonparametric approach with Dirichlet process mixture (DPM) of multivariate skew t-distributions to perform model based clustering of such data. DPM models enable the number of cell populations to be directly estimated from the data avoiding any model selection. The use of skew t-distributions provides robustness to outliers and suits best the usually non elliptical shape of cell populations. In the case of repeated measurements, we propose a sequential strategy relying on a parametric approximation of the posterior, and we show the benefit of such a sequential strategy for detecting rare populations.

We illustrate the good performances of our method on simulated data, on an experimental benchmark dataset, and on new experimental data from the DALIA-1 trial, a clinical trial evaluating a therapeutic vaccine against HIV where flow cytometry data along with genomics data were available longitudinally.
Predicting breast cancer metastasis from blood gene expression using time-to-diagnosis data to derive signature estimates

Screening is nowadays the main tool for breast cancer detection. To determine whether the cancer is metastatic, it is necessary to do a sentinel node biopsy. There is a high false positive rate in the screening program, and the biopsy is invasive. Hence it is valuable to determine if mRNA gene expression levels in blood samples contain some predictive power. The presence of cancer biomarkers in blood mRNA is in general debated.

In this paper we use gene expression levels from blood samples to predict the metastatic spread of breast cancer. We have 88 case–control pairs from the prospective Norwegian Women and Cancer study. These samples are from up to one year before the time of diagnosis. As we have the followup time, we can use time-to-diagnosis as a variable in selecting predictors for our models. We do this by fitting genewise linear models of gene expression on time and metastasis. The coefficient t-statistics provide a ranking of genes.

We use repeated, nested cross-validation to estimate the expected prediction error of several models. We compare models with and without time-to-diagnosis variable selection and find some evidence that this preselection improves predictions. We also find that this preselection, compared to simple genewise t-tests, leads to more stable signatures.

Coauthors: Vittorio Perduca, Eiliv Lund, Lars Ailo Bongo and Etienne Birmelé
Multiple hot-deck imputation for network inference from RNA sequencing data

The clinical-research project DiOGenes [5] is based on a dietary intervention led in 8 European centers and investigates the effects of macronutrient composition on weight-loss maintenance and metabolic and cardiovascular risk factors after a phase of calorie restriction in obese individuals. A first phase was an 8-week low-calorie diet with the objective of more than 8% weight loss. The second phase was a maintenance phase. For this purpose, the successful patients were randomized into one of five ad libitum weight maintenance diets during 6 months: four diets combining high and low protein content with high and low glycemic index of carbohydrates and a control diet. Clinical, phenotypic and transcriptomics measures were performed using blood samples and adipose tissues at the baseline and after each phases of the study.

In the present proposal, we will focus on understanding differences between two time steps of the study (CID1, before the dietary intervention, and CID2, after the 8-week low-calorie diet), at a transcriptomic level. A usual method to understand the interaction between genes and their evolutions for transcriptomic data is the inference and analysis of gene co-expression networks [6]. Here, network inference is performed from RNA-seq data [1], that were acquired on about 200-400 individuals (depending on the time step). RNA-seq between CID1 and CID2 are paired for most individuals. However, for experimental reasons, some data were acquired for only one time step for some individuals (only 189 individuals are common to the two time steps among 433 individuals measured at CID1 and 307 measured at CID2). Two approaches are possible: the first one is to infer the gene networks with all available individuals for each time step, in order to obtain a better precision in network inference. A second approach consists in inferring the gene networks with only common individuals between CID1 and CID2. The latter approach is more relevant to compare the two time steps. However, since network inference can be very sensitive to influential individuals [3], the networks resulting from these two approaches can be rather different.

In order to infer more stable and more comparable networks, we propose a method, which is written as a missing data imputation problem: our proposal relies on an auxiliary dataset (e.g., another ‘omic dataset or a less complete transcriptomic dataset obtained with a cheaper technique, such as Rt-qPCR), for which all observations for all individuals are given. The problem is then solved with a multiple hot deck approach [2, 7, 4] which is adapted to unit non-response. Our method is evaluated on simulated datasets and applied to the data of the DiOGenes project to obtain more reliable and more comparable networks.


Coauthors: Caroline Le Gall, Claudia Armenise, Gregory Lefebvre, Jorg Hager, Armand Valsesia, Pierre-Antoine Gourraud, Nathalie Viguerie and Nathalie Villa-Vialaneix
Network-based dimension reduction methods for the integrative analysis of multi-omics data in cancer

Cancer is one of the most complex diseases at multiple molecular levels. No single level of genomic data can fully explain tumor behaviour necessitating analysis at different levels encompassing genomics, epigenomics and proteomics - such as gene expression (using microarray and RNA sequencing), epigenetic variations (by methylation arrays and bisulfite sequencing) and protein variations (assayed in either metabolomic or proteomic studies).

In the last years, thanks to international projects and consortia the access to the genome-wide data at multiple molecular levels has been made available by a variety of high-throughput technologies. For example, the Cancer Genome Atlas (TCGA) program and the European Genome-phenome Archive (EGA) were established to profile large tumor sets at both DNA and RNA levels and to create integrated repositories of all types of sequence and genotype experiments, including the aberrations present in cancer cells. Therefore, data from TCGA and EGA can provide opportunities and challenges to develop sophisticated statistical and computational tools for the analysis, interpretation and validation of cancer data and help the cancer research community to improve the prevention, diagnosis, and treatment of cancer.

However, from a statistical perspective, the most important challenge in integrating multi-omic analyses is the high-dimensionality of the data. In fact, taking more levels into account increases the dimensionality of the problem. In particular, adding more layers of data raises the dimension of unknown parameters, which are often difficult to estimate, thereby making the overall inference weaker. In addition, at every step, there are checkpoints of data compatibility, such as normalisation at the same scale, adequate correction for technical batch effects and use of different platforms. As a consequence, in order to tackle such problems, the combination of dimension reduction methods and network-regularized approaches for the integration of biological information into models need to be developed for the exploratory study of multi-layer datasets in cancer survival analysis.

In our study, we combine dimension reduction techniques and network-penalized methods for the identification of the candidate pathways and genes highly associated with the cancer disease in order to obtain a good prediction and interpretability of survival data. We propose a new multistage computational-statistical strategy for survival analysis based on the following steps. First, we integrate different data types identifying the best approach for combining multiple matrices that include data from different scales and then, we reduce the high-dimensionality of data. Second, we incorporate gene regulatory network information using penalized Cox regression methods. Third, we test the predictive power of the selected gene signatures using independent datasets in order to derive novel disease gene interaction pathways and disease-risk genes.

More precisely, (i) a biological screening is performed if there is enough biological information available; (ii) a statistical screening is applied if there is no biological information available. We achieve this by introducing the concept of sure screening and proposing a network-based cox sure screening method which is based on correlation learning which filters out the features that have weak correlation with the response. Higher the correlation, lower the probability of being removed. (iii) A biological and statistical screening is mixed if there is partial or poor biological information available. It is useful to obtain information not already known about the disease. In addition, we performed a pathway analysis based on KEGG database and on the Human Experimental/Functional Mapper (Huttenhower et al., 2009). In particular, we focused on a gene-gene interaction analysis developing gene-networks that describe the relations between genes in terms of KEGG pathways. Hence, while the dimension reduction methods recruit the features with the best marginal utility to reduce the dimensionality of the omics data, the network incorporates the pathway information used as a prior knowledge network into the survival analysis.

We fuse dimension reduction procedures and network-penalized Cox models for high dimensional survival data. By using this approach, it is possible to obtain a deeper insight of the gene-regulatory networks and investigate the gene signatures related to the survival time. The study focus on multi-omic data downloaded by TCGA and EGA projects. Time-to-event data are used to select prognostic markers and pathways involved in different types of cancer, such as ovarian and breast cancer. Before to reduce the dimensionality of the data, we first analyse individual biological networks, then we integrate multiple biological networks, i.e. networks where each layer stands for a different type of interaction between the same set of nodes.

Using these network, as a-priori biological information, we perform network-based Cox model including Kaplan-Meier curve and log-rank test. In general, to evaluate predictive accuracy of survival risk classifiers, we note that the p-value is very significant when we combine the biological and statistical screening. The choice of the threshold value is data-driven and model based.

Overall this study shows that the new multistage strategy is useful for improving the accuracy of survival prediction and to cope with the often-fatal curse of dimensionality. It is necessary because when we handle high-dimensional data using traditional methods, we experience several computational and mathematical complex problems. Additionally, there is a growing evidence supporting network gene regulatory investigations and its implications in the development of cancer survival analysis. Therefore, this new approach will provide the basis for significant advancements in cancer research for the biomedical science community.

Coauthors: Annalisa Occhipinti, Claudia Angelini, Italia De Feis and Pietro Lio
MDI: Multiple dataset integration using Bayesian correlated clustering

Using genomics datasets to identify meaningful subgroups (whether of patients, genes, DNA motifs, or any other biological units) remains a key task in statistical genomics and stratified medicine. The increasing availability of diverse genomics datatypes presents challenges, as well as opportunities, for subgroup identification.

We present a Bayesian correlated clustering algorithm, which we refer to as MDI (Multiple Dataset Integration). MDI is able to integrate the information from a wide range of different datasets and data types simultaneously (continuous/categorical, single/multiple time points, ...). Each dataset is modelled using a Dirichlet-multinomial allocation (DMA) mixture model, with dependencies between these models captured via parameters that describe the levels of agreement among the datasets. We establish the effectiveness of MDI using a number of examples, including a case study in which we integrate gene expression, transcription factor binding, and protein-protein interaction data, to identify a set of protein complexes for which genes are co-regulated during the cell cycle. We also demonstrate an accelerated implementation of MDI that uses general purpose GPU computing in order to parallelise computation, thereby allowing MDI to be applied to big datasets.
Causal inferences in cross-sectional data - opportunities and drawbacks in the joint analysis of genomic, transcriptomic and metabolomic data

Post-GWAS analyses are frequently characterized by the integration of information from additional omic-levels, e.g. transcriptome data or metabolome data. Thereby, questions related to causality are of central interest, e.g. "Is the association caused by genetically altered gene expression?", or "Is there evidence for an unconfounded link between gene-expression levels and the analysed endpoint?". Commonly adopted methods to provide answers for those questions include Mendelian randomisation and mediation analysis. However, the use of these methods requires acceptance of rather strong assumptions, assumptions that are frequently not met in the available data.

At the example of our GWAS on blood metabolites that integrated analysis of gene expression data we show insight and limitations achieved by these methods. We compare our analysis with related published data and provide improvements for Mendelian randomisation to increase robustness in case that assumptions are not met. These results are complemented by simulations to gain a deeper insight and to better apply methods of causal inference in the identification of rationales for subsequent functional studies.

Coauthors: Frank Beutner, Joachim Thiery, Uta Ceglarek and Markus Scholz
Data-driven hypothesis weighting increases detection power in genome-scale multiple testing

Hypothesis weighting is a powerful approach for improving the power of data analyses that employ multiple testing. However, in general it is not evident how to choose the weights.

I will describe IHW (independent hypothesis weighting), a method for data-driven hypothesis weighting that makes use of informative covariates that are independent of the test statistic under the null, but informative of each test's power or prior probability of the null hypothesis. Covariates can be continuous or categorical and need not fulfill any particular assumptions. Examples include the sum of read counts across all samples in an RNA-Seq differential expression study or the distance between the genetic variant and the locus of expression in eQTL studies.

The method increases statistical power in applications while controlling the false discovery rate (FDR) and produces additional insight by revealing the covariate-weight relationship. Independent hypothesis weighting is a practical approach to discovering associations in genomics, high-throughput biology and other large data sets. IHW is available from Bioconductor (http://www.bioconductor.org/packages/IHW).

Coauthors: Nicolaos Ignatiadis, Judith Zaugg and Wolfgang Huber
Prediction of childhood asthma risk by high-dimensional genetic and environmental data

Genome-wide association studies typically contain hundreds of thousands to millions of SNPs, which are typically uni-variately tested for association to a certain phenotype. More rarely multivariate statistical learning techniques have been used on these kinds of large data sets in order to predict a certain phenotype on an independent data set.

The GABRIELA study’s phenotype of interest was childhood asthma. The study contained 1708 individuals, 2.5 mio SNPs and further features on environmental exposures. Since one was especially interested in the impact of specific farm-related exposure an efficient complex survey design was conducted. Our aim was to build a prediction model for the risk of suffering from childhood asthma using the genome-wide information as well as the environmental features.

Quality control and linkage-disequilibrium pruning were performed in order to reduce the dimension of the data in a first step. Then feature selection and machine learning techniques were applied and adjusted via weighting for correcting sample selection bias caused by the complex survey design. Prediction models were trained and compared via cross validation, and the best-performing model was applied on an independent held-out validation data set. Several SNPs and several farm-related features were identified as predictive but were not exceeding predictive power of the family history of an individual. Best performance could be achieved when both SNPs and environment information were used by incorporating their group structure in a weighted LASSO approach assigning different penalty factors to different data groups.

Coauthors: Christiane Fuchs, Fabian Theis and Markus Ege
Personalized health: challenges in data science

The promise of personalized health is driven by the wide availability of data, but we don’t need to talk so much about where we want to be, rather how we should get there. What are the challenges that need to be bridged technologically to unlock the potential in the much greater availability of data we now have?

In this talk we’ll consider three challenges of data science in the context of personalized health, the three challenges each need to be bridged to bring the era of true precision, or personalized, medicine within the reach of an affordable health care service.
Gene network reconstruction using global-local shrinkage priors

High-throughput biotechnologies such as microarrays provide the opportunity to study the interplay between molecular entities, which is central to the understanding of disease biology. The statistical description and analysis of this interplay is naturally carried out with graphical models in which nodes represent molecular variables and edges between them represent conditional dependencies.

Inferring the structure of the graph (i.e. the edge set) from high-throughput molecular data is an important but challenging task, as the number of parameters to estimate easily is much larger than the sample size. A conventional remedy is to regularize or penalize the model likelihood. In network models, this is often done locally in the neighbourhood of each node or gene. However, estimation of the many regularization parameters is often difficult and can result in large statistical uncertainties.

In this talk we propose to combine local regularization with global shrinkage of the regularization parameters to borrow strength between genes and improve inference. We employ a simple Bayesian model with non-sparse, conjugate priors to facilitate the use of fast variational approximations to posteriors. We discuss empirical Bayes estimation of hyper-parameters of the priors, and propose a novel approach to rank-based posterior thresholding. Using model- and data-based simulations, we demonstrate that the proposed inference strategy outperforms popular (sparse) methods, yields more stable edges, and is more reproducible.

The proposed approach is also shown to lend itself well to the incorporation of prior biological knowledge and we show how the network reconstruction may benefit substantially from the inclusion of prior knowledge via structured shrinkage and empirical Bayes.


Coauthors: Mark A. van de Wiel, Gino B. Kpogbezan, Aad W. van der Vaart, Wessel N. van Wieringen and Mathisca C.M. de Gunst
Because evolutionary processes unfold within a phylogenetic context, comparative phylogenetic methods take up a prominent role in ecological and evolutionary analyses. Widely used to investigate the evolutionary variation of phenotypic characters and life-history traits, these analyses typically consider a phylogenetic reconstruction or sequence evolution process that is divorced from the trait evolutionary process.

Here, I will discuss a comprehensive and coherent statistical framework for integrated analyses of sequences and traits equipped with dedicated models and inference techniques to tackle complex trait data. These include generalised Brownian processes for multivariate trait diffusion, multidimensional scaling implementations, and latent liability extensions that can capture the joint variation in multiple trait observations. For discrete traits, I will present generalised linear modelling extensions that allow integrating and testing covariates. Applications will be illustrated in the field of infectious diseases, in particular for phenotypic and spatial evolution of rapidly evolving viruses such as Influenza and Ebola.
Bayesian inference on high-dimensional Seemingly Unrelated Regressions, applied to metabolomics data

Increasingly, epidemiologists are collecting multiple high-dimensional molecular data sets on large cohorts of people. The interest is in finding associations between these data sets and with genetic variants. In order to do this effectively these multi-variate data sets should be modelled jointly, taking into account correlations in the data. Sparse solutions are usually required, and performing variable selection in this setting is critical.

We present a Bayesian Seemingly Unrelated Regressions (SUR) model for associating metabolomics outcomes with genetic variants, allowing for both sparse variable selection and correlation between the outcomes. This model can be fit using a Gibbs sampler, but this quickly becomes computationally unfeasible as the dimensions of the problem grow. Previously people have made use of either the assumption of independence between the outcomes (Bottolo et al. 2011, Lewin et al. 2015) or selected predictors jointly for all the outcomes (Bhadra and Mallik 2013, Bottolo et al. 2013).

In order to overcome some of the computational difficulty with the general SUR model, Zellner and Ando (2010) proposed a reparametrisation of the model in which the likelihood factorises completely into a product of conditional distributions, and used a Direct Monte Carlo (DMC) approach to estimate the posterior. This improves computational time, however their method requires re-sampling of the regression coefficients in order to obtain the correct posterior distribution.

We extend their work by allowing for a more general prior distribution, and we show that it is possible to build a Gibbs-DMC sampler without the need for re-sampling. Zellner and Ando (2010) demonstrated their DMC method on examples with up to 3 responses. We are aiming higher, with real molecular biology data involving 100's or 1000's of responses. The proposed method is applied to both simulated data, to illustrate the computational gains, and real metabolomics analysis where the dimension of the data precludes the use of the traditional sampler.

Coauthor: Marco Banterle
Modelling collinear and spatially correlated data

We present a statistical approach to distinguish and interpret the complex relationship between several predictors and a response variable at the small area level, in the presence of i) high correlation between the predictors and ii) spatial correlation for the response.

Covariates which are highly correlated create collinearity problems when used in a standard multiple regression model. Many methods have been proposed in the literature to address this issue. A very common approach is to create an index which aggregates all the highly correlated variables of interest. For example, it is well known that there is a relationship between social deprivation measured through the Multiple Deprivation Index (IMD) and air pollution; this index is then used as a confounder in assessing the effect of air pollution on health outcomes (e.g. respiratory hospital admissions or mortality). However it would be more informative to look specifically at each domain of the IMD and at its relationship with air pollution to better understand its role as a confounder in the epidemiological analyses.

In this talk we illustrate how the complex relationships between the domains of IMD and air pollution can be deconstructed and analysed using profile regression, a Bayesian non-parametric model for clustering responses and covariates simultaneously. Moreover, we include an intrinsic spatial conditional autoregressive (ICAR) term to account for the spatial correlation of the response variable.

Coauthors: Aurore Lavigne and Marta Blangiardo
Adaptive group-regularized logistic elastic net regression

In high dimensional data settings, additional information on the variables is often available. Examples of such external information in genomics research are: (a) p-values on the same variables, obtained from a previous study, (b) information from a publicly available database such as pathway membership, (c) genomic annotation (e.g. the location on the genome), or (d) response independent summary statistics like sample standard deviation of the variables. The inclusion of this information into the analysis may benefit classification performance and variable selection, but is not straightforward in the standard regression setting.

As a solution to this problem, we propose a group-regularized (logistic) elastic net regression method, where each penalty parameter corresponds to a group of variables. The grouping of the variables is based on a partitioning according to the external information. The method makes use of the Bayesian formulation of logistic elastic net regression to estimate both the model and penalty parameters in an approximate empirical Bayes - variational Bayes framework.

Simulation results show that in settings where the partitioning of the variables is informative, the group-specific penalization of the variables does indeed benefit classification performance. In addition to the simulation results, we present the application of the proposed method to a real-world cancer genomics data set.

Coauthors: Mark van de Wiel and Wessel van Wieringen
Integrating biological information through differential shrinkage in a Bayesian joint model: improving the analysis of genetic association studies

Genome-Wide Association Studies are usually analysed by estimating SNP effects individually (standard analysis) and adjusting for multiple testing. However, SNPs detected by this method explain only a small fraction of the predicted heritability for most traits. Here we aim to improve SNP detection by integrating external biological information about the SNPs in a Bayesian hierarchical shrinkage model that jointly estimates SNP effects.

We assume that the SNP effects follow a normal distribution centered at zero and use prior biological information retrieved from online databases to modulate the variance of the SNP effects through differential shrinkage. SNPs with more biological support are less shrunk towards zero, thus being more likely detected. Additionally, we propose a novel measure to determine the best set of shrinkage parameters to use in this setting of differential shrinkage.

The performance of the method was tested in a simulation study on 1000 datasets, each with 500 subjects and ~200SNPs divided in 10 linkage disequilibrium (LD) blocks, using both a continuous and a binary outcome. The method was further tested in an empirical example on body mass index and on normal vs. overweight individuals (continuous and binary outcomes) in a dataset with 1,829 subjects and 2,614 SNPs in 30 blocks. Biological knowledge was retrieved using the bioinformatic tool Dintor and the Pfam, MGI and Reactome databases. The Bayesian model with inclusion of prior information outperformed the standard analysis: in the simulation study, the mean ranking of the true LD block was 2.8 vs. 3.6 in the standard analysis (continuous outcome) and 2.9 vs. 3.7 (binary outcome); in the empirical example, the mean ranking of the six true blocks was 8.3 vs. 11.7 in the standard analysis (continuous outcome) and 7.3 vs. 8.2 (binary outcome). These results suggest that our method is more powerful than the standard approach, and we expect its performance to improve even further as more biological information about SNPs becomes available.

Coauthors: John Thompson, Christian Weichenberger, Duncan Thomas and Cosetta Minelli
A latent model to infer tumoral heterogeneity from DNA copy number profiles

Single-polymorphism nucleotide (SNP) arrays enables to measure the DNA copy number across thousands of locations in a genome but also the Loss of Heterozygosity (LOH). In cancerology, the main goal of the analysis of these data is to identify the regions where the DNA copy number has been altered. However, it is common that primary tumors are not composed of only one type of cells but several. This phenomenon is called intra-tumor heterogeneity. Although there are many methods for analyzing heterogeneity in a single-sample, this raises the issue of identifiability and a way to solve it could be to infer heterogeneity in a multi-sample analysis.

We propose an extension to previous models that provides a statistical framework for modeling heterogeneity for a multi-sample analysis by including new constraints and the LOH information. The method involves modeling parental copy number as a weighted sum of a fixed number of latent profiles after a joint segmentation step. The simulation analysis shows that our model outperforms the existing ones. In addition, an analysis of a public data set on kidney cancer has been conducted. We studied samples from the same patient at various locations. This analysis shows that the proposed model provides interesting results in terms of the composition of the primary tumor that are coherent with the ones of the initial paper.

Coauthors: Julien Chiquet and Pierre Neuvial
Evaluation of different approaches to stratify somatic mutation profiles

One of the fundamental question of cancer informatics is tumor stratification, whereby a heterogeneous population of tumors can be classified into clinically and biologically meaningful subtypes. Tumor stratification based on somatic mutations is challenging because cancers are usually highly heterogeneous and because driver events are in general rare and not necessarily shared among patients.

Because cancer is a network disease, i.e., driven by combinations of genes acting in networks corresponding to hallmark processes such as cell proliferation and apoptosis or affecting the whole network region, Hofree et. al. (2013) proposed to integrate to the somatic mutation profiles the knowledge of the molecular network architecture of human cells to reveal the shared networks affected by these mutations.

Here, we propose a new standardized evaluation for stratification algorithms of somatic mutation profiles, which integrates different indexes into synthetic scores of stability and internal measures, and is able to select a good stratification algorithm. Our method is based on the following steps: i/ Obtention of stability and intrinsic quality indexes measured on clustering outputs. ii/ Construction of synthetic scores representative of all quality criteria containing the most possible of quality indexes knowledge. iii/ Ranking of stratification algorithms by the values predicted by the best model (in BIC’s sense) fitted for each of the retained syntetic score.

We build several strategies of stratification, each integrating a protein-protein interaction database (PPI), a kernel propagation method linking the PPI to the somatic mutation profiles and a clustering algorithm. These strategies were applied on several TCGA somatic mutation datasets.

In this work, we therefore present the results of our protocol of evaluation to the combination of one dataset and one stratification strategy.


Coauthors: Isabel Brito and Philippe Hupe
Graphical model inference with unobserved variables via latent tree aggregation

In many biological case studies, the empirical covariance matrix of the variables of interest displays large blocks of uniform correlation. This suggests the existence of one or several unobserved (missing) variables having a simultaneous influence on a series of observed ones, and that we observe a sample drawn from a distribution where the unobserved variables have been marginalized out. The inference of underlying networks is compromised in this context because marginalizing variables yields locally dense structures that challenge the generally accepted assumption that biological networks are sparse.

We present a procedure for inferring Gaussian graphical models from an independent sample in the presence of unobserved variables. Both the graph structure and the unobserved nodes are considered as latent variables.

We propose to exploit the simple properties of trees in terms of identifiability to build an inference strategy based on the EM algorithm and spanning trees. Considering only trees is very restrictive. We overcome this by treating trees as random variables and relying on an algebra result called the Matrix-Tree theorem, that enables us to sum posterior probabilities over the space of trees containing a given edge, therefore computing quantities that can be interpreted as posterior probabilities of edge appearance.

We compare our method to existing graph inference techniques on synthetic and flow cytometry data.

Coauthors: Stéphane Robin and Christophe Ambroise
Metabolic effects of environmental pollutants in pregnant women - an exposome approach

Environmental pollutants are suspected of adversely affecting child development through, among other mechanisms, perturbation of endogenous metabolism. Few studies have investigated the metabolic effects of exposure to environmental pollutants among human populations, with pregnant women representing a particularly important subgroup. We aimed to assess the association between the exposome, measured by multiple pollutant biomarkers, and the metabolome, characterised by 1H NMR urinary metabolic phenotypes.

Women were enrolled into the Spanish INMA birth cohort at the 1st trimester routine antenatal visit in Sabadell and Gipuzkoa during 2004-2007. In both centres, biomarkers of persistent pollutants were measured in blood (organochlorine pesticides, PCBs, PFAS, mercury). In Sabadell only, non-persistent pollutants were measured twice in urine (metals, phthalates, BPA) collected at 12 and 32 weeks of pregnancy. Repeated 1H NMR metabolic profiles were acquired for both centres in repeat urine samples (N = 806 at 12 weeks and 886 at 32 weeks). Exposome-metabolome associations were assessed by partial correlations, separately for persistent and non-persistent pollutants, with correction for false discovery rate (FDR).

50 exposure-metabolite correlations were statistically significant (FDR 5%) in the persistent pollutant analysis. In the non-persistent pollutant analysis, 149 associations were detected in the first trimester and 89 in the third trimester, with 39 of these associations replicated in both trimesters. The strongest replicated association was between arsenic and the mammalian -microbial co-metabolite trimethylamine oxide ($p = 1.3 \times 10^{-13}$), suggesting a role for the gut microflora in arsenic metabolism. These results indicate that low level pollution may influence metabolism during pregnancy, which may have potentially far-reaching implications for the developing fetus.

Coauthors: Lea Maitre, Mireille Toledano, Jesús Ibarluzea, Loreto Santa Marina, Jordi Sunyer, Cristina Villanueva, Martine Vrijheid and Muireann Coen
The OrMachine

We introduce the OrMachine, a model for dimensionality reduction of binary data with interpretable latent representations. Common latent variable models rely on continuous parameters and/or continuous latent representations which severely limits their interpretability and is inappropriate if they intend to capture the presence or absence if biological processes.

The OrMachine generates observations from joint assignment of multiple binary prototypes, each of which captures dominant patterns of correlation in the data. Inference of the posterior distribution is fast and scales to large datasets using a Metropolised Gibbs sampler. Prior knowledge can readily be built in by clamping model variables or parameters. Furthermore, OrMachines can be stacked to build a multi-layer hierarchy of representations.

We present the OrMachine on scRNAseq data, as well as on mutational landscape in different cancer types. For both examples we uncover prototypes that correspond to the underlying biology. This demonstrates the capability of the OrMachine to unravel complexity in high-dimensional biological data, producing readily interpretable features for the visualisation and discovery of biological processes.
Transcriptional analysis of two large-scale cancer cell line panels identifies novel insights for pharmacogenomics studies

In order to test hundreds of drugs on numerous tumor models, the scientific community commonly uses tumor derived cell lines. The goal of these studies is to find genomic biomarkers that can help predict whether a given patient will respond to a drug and ultimately to select the best drug for each patient. A recent comparative analysis of large-scale pharmacogenomic high-throughput screening (HTS) data set, has found inconsistency between the measured cell line drug sensitivities across several studies. We explored the possibility to improve our understanding of the results from cell lines drug sensitivity and their consistency.

We present a comprehensive reanalysis of these pharmacogenomic data from the Cancer Genome Project (CGP) and the Cancer Cell Line Encyclopedia project (CCLE) studies. We focused on common cell lines (n=471) and molecular information available as well as the IC50 of 15 drugs that have been screened by CGP and CCLE. We propose a novel cell lines classification that groups cell lines by their transcriptomic profiles based on a biological network-driven gene selection process. This robust molecular classification of cancer cell lines defined eleven clusters. Only four clusters of cell lines were characterized by a majority of tumors with same tissue of origin. The cell-of-origin signal then appears to not being the key driver characteristic of cell lines transcriptomic profiles. Furthermore, this novel cell lines clustering show higher homogeneity in term of drug sensitivity, compared to the cell line tissue of origin organization. Finally, we were able to find associations of cell line clusters and drug sensitivity or resistance for six drugs: erlotinib, lapatinib, palbociclib, vemurafenib, PD0325901 or selumetinib. These associations were robustly found in both large-scale pharmacogenomics data-sets.

This study defines a robust molecular classification of cancer cell lines, which provides new biological insights for drug sensitivity data reproducibility. This classification has a great potential for use in the identification of new biomarkers of response.

LEAN discovery of hot spots in networks

“Everything should be made as simple as possible but not simpler,” said Einstein. But what does this mean for new computational models that link complex disease ‘omics data with relevant phenotypes? To be useful in practice, new models need to be simple enough to be computationally tractable, and yield biologically interpretable outcomes. Yet, models also need to be complex enough to allow the discovery of new, non-classical relationships between molecular and clinical measurements and disease phenotypes.

In my talk, I will discuss a simple subnetwork model for identifying ‘hot spots’ in interaction networks. Methods based on the classical subnetwork model tend to have long running times, provide single or partial, often heuristic, solutions, contain user-tuneable parameters, or lead to solutions that are difficult to interpret. An alternate approach (termed Local enrichment analysis, or LEAN) substitutes the general model by simpler model. The simpler model is more constrained, but, in return, allows exact, parameter-free, efficient, and exhaustive identification of local subnetworks that are statistically dysregulated, and directly implicates single genes for follow-up experiments.

A first empirical evaluation on simulated and biological data suggests that LEAN detects dysregulated subnetworks, and reflects biological similarity between experiments better than standard approaches. A strong signal for the local subnetwork around Von Willebrand Factor (VWF), a gene which showed no change on the mRNA level, was identified by LEAN in transcriptome data in the context of a genetic disorder, Cerebral Cavernous Malformations (CCM). Targeted follow-up experiments revealed an unexpected strong cellular phenomenon around VWF. The LEAN method can be used to pinpoint statistically significant local subnetworks in any genome-scale data set.
Detecting interactions in GWAS with the gene-gene eigen epistasis approach

In past years, numerous methods have been proposed for studying epistatic interactions in genome-wide association studies (GWASs). They vary in terms of data analysis (genome-wide or filtering) and statistical methodology (Bayesian, frequentist, machine learning or data mining). Most of them focus on single-locus interactions, but considering interactions at gene level may offer many advantages. In the past few years several gene-gene methods have been proposed, they rely on a summarizing step to obtain information at the gene level and a modeling phase to represent interactions. Directly modeling all gene-gene interactions would be inefficient due to computational challenge and lack of power. For the most recent methods, filters or penalized models are used to make the method applicable to a large number of genes.

Here we propose a Group LASSO based method [Yuan and Lin, 2006] that takes into account the group structure of each gene in order to detect epistasis at the gene level. We introduce the Gene-Gene Eigen Epistasis (G-GEE) as a new approach to compute the gene-gene interaction part of the model. The method first compute interaction variables for each gene pair by finding its Eigen-epistasis Component defined as the linear combination of Gene SNPs having the highest correlation with the phenotype. The selection of the significant effects results from a group LASSO penalized regression method combined to an adaptive ridge approach [Becu et al., 2015] allowing to control the False Discovery Rate.

We conduct a simulation study to compare G-GEE with recent alternative proposals and demonstrate the power of our approach by detecting new gene-gene interactions on genome-wide association studies.


Coauthors: Christophe Ambroise and Cyril Dalmasso
Using co-data to improve the predictive accuracy of a Random Forest

Prediction with high-dimensional data is inherently a difficult problem due to the typical needle in the haystack issue. We demonstrate how we can improve the performance of a Random forest (RF) on high dimensional (genomics) data by guiding it with "co-data". Co-data here are any type of data that contains information about the primary set of covariates used for training.

Guiding a prediction model by co-data may lead to improved predictive performance and variable selection in high dimensional settings. A number of methods have been developed that allow for incorporation of co-data, all based on penalized regression, and they either allow for one type of continues co-data (weighted lasso), or do allow to model multiple types of co-data, but requires co-data to be divided into groups (group ridge). We suggest a method that allows the joint usage of multiple types of co-data, continues co-data, and (flexible) modelling of the co-data. The described methodology can in principle be used in combination with any bagged classifier, we only focus on RF. Additionally advantages of a RF are that is invariant to data transformations, which makes it especially useful for RNASeq (count) data, and it is able to seamlessly handle nonlinearities.

The method is exemplified with the prediction of a lymph node metastasis for patients with a diagnosed oral squamous cell carcinoma (OSCC). The primary data used for prediction in this example is TCGA RNASeq data. Co-data used to guide the RF consists of a DNA copy number data, also downloaded from TCGA, p-values of an external independent set of microarray data, and a validated gene profile for lymph node metastasis. The co-data supported RF outperformed the base RF on an independent validation data set and raised the AUC from 0.63 to 0.70.

Coauthor: Mark van de Wiel
Empirical Bayes learning from co-data in high-dimensional prediction settings

Empirical Bayes is an approach to ‘learn from a lot’ in two ways: first, from a large number of variables and second, from a potentially large amount of prior information on the variables, termed ‘co-data’, for example available in public repositories. We review empirical Bayes methods in the context of regression-based prediction models. We discuss ‘formal’ empirical Bayes methods which maximize the marginal likelihood, but also more informal approaches based on other data summaries. We contrast empirical Bayes to cross-validation and full Bayes. Hybrid approaches are proposed.

Empirical Bayes is particularly useful to estimate multiple hyper-parameters that model the information in the co-data. Some examples of co-data are: p-values from an external study, additional molecular measurements or genomic annotation. The systematic use of co-data can considerably improve predictions and variable selection, which we demonstrate on two applications: blood-based molecular cancer diagnostics and molecular cervical cancer screening. Finally, some extensions to other prediction methods, such as the random forest, and to other problems, such as network estimation, are shortly discussed.
A tale of two networks

The two-sample problem is addressed from the perspective of Gaussian graphical models (GGMs). It concentrates on the particular situation in which partial correlations (i.e. edge strength measures between node pairs in a GGM) are systematically smaller/larger (in an absolute sense) in one of the groups. Biologically, systematically weaker/stronger partial correlations represent an inactive/active pathway.

Data in both groups are assumed to follow a GGM but their partial correlations are proportional, differing by a multiplier (common to all partial correlations). The multiplier reflects the overall strength of the conditional dependencies. Model parameters are estimated by means of penalized maximum likelihood, using a ridge-like penalty. A permutation scheme to test for the multiplier differing from zero is proposed.

A re-analysis of publicly available data (from six studies) on the Hedgehog pathway in normal and cancer prostate tissue shows its activation in the disease group. The analysis is accompanied by extensive diagnostics to assess the value of this conclusion.
Within-host genetic diversity and large transmission bottlenecks confound phylodynamic inference of epidemiological dynamics. Conventional phylodynamic approaches assume that nodes in a time-scaled pathogen phylogeny correspond closely to the time of transmission between hosts that are ancestral to the sample. However, when hosts harbour diverse pathogen populations, node times can substantially pre-date infection times. Imperfect bottlenecks can cause lineages sampled in different individuals to coalesce in unexpected patterns.

To address realistic violations of standard phylodynamic assumptions we developed a new inference approach based on a multi-scale coalescent model, accounting for nonlinear epidemiological dynamics, heterogeneous sampling through time, non-negligible genetic diversity of pathogens within hosts, and imperfect transmission bottlenecks.

We apply this method to HIV-1 and Ebola virus outbreak sequence data, illustrating how and when conventional phylodynamic inference may give misleading results. Within-host diversity of HIV-1 causes substantial upwards bias in the number of infected hosts using conventional coalescent models, but estimates using the multi-scale model have greater consistency with reported number of diagnoses through time. In contrast, we find that within-host diversity of Ebola virus has little influence on estimated numbers of infected hosts or reproduction numbers, and estimates are highly consistent with the reported number of diagnoses through time. The multi-scale coalescent also enables estimation of within-host effective population size using single sequences from a random sample of patients. We find within-host population genetic diversity of HIV-1 p17 to be $2 N \mu = 0.012$ (95% CI: 0.00066, 0.023), which is lower than estimates based on HIV envelope serial sequencing of individual patients.

Coauthors: **Ethan Romero Severson** and **Thomas Leitner**
Gene regulatory networks from single cell data

It is very tempting to use snapshot gene expression data from single cells to obtain a glimpse of the underlying dynamic cell processes. Several methods have been proposed to turn these cross-sectional into time-series data by pseudotemporal ordering. The assumption is that cells are at different stages of a common dynamic process and can be ordered accordingly. Gaussian processes are well suited to capture the uncertainty due to uncertainty in the ordering and measurement error. I will explore how far this uncertainty can be incorporated in autoregressive and state space models of dynamic processes.
Contributed talk

Verena Zuber  
European Bioinformatics Institute (EBI), Hinxton  
verena@ebi.ac.uk

Two-dimensional fine-mapping of global and molecular traits to identify gene-disease linkages

A major challenge in the post genome-wide association studies (GWAS) era is to link genetic variants associated with a trait of interest to mediating expression of genes. Here we propose a statistical framework to integrate summary statistics from large (meta-)analysis of GWAS and expression quantitative trait loci (eQTL) studies in order to identify regions with shared genetic variants between two traits, i.e. the trait of interest and gene expression.

Extending existing Bayesian strategies for fine-mapping of one trait, our model estimates the posterior probability for shared causal variants between GWAS and eQTL using a closed form Bayes factor. Our model allows for both, one or multiple causal variants, and accounts for the correlation structure between genetic variants. To explore efficiently the combinatorial space of causal configurations, we implement a shotgun stochastic search strategy.

We first validate our model through simulations, showing that our model does identify loci with shared genetic architecture more accurately than alternative methods, in particular if multiple causal variants are considered. Finally, we present results from an application on summary statistics from GWAS on inflammatory bowel disease, for which we identify target genes based on eQTL data from the GTEx study.

Coauthors: Will Jones, Daniel Zerbino and Oliver Stegle