Contributed session 2

A comparison of methods for inferring causal relationships in genetic data
Holly F. Ainsworth, So-Youn Shin and Heather J. Cordell

Many novel associations between genetic variants and human disease have been successfully identified using genome-wide association studies (GWAS). However, a typical GWAS gives little insight into the biological function through which these associated genetic variants are implicated in disease. Indeed, rather than finding variants which directly influence disease risk, the variants implicated by GWAS are typically in linkage disequilibrium with the true causal variants. Understanding the causal role the genetic variants play in disease and moving towards therapeutic interventions is not simple. Integration of additional data such as gene expression, proteomic and metabolomic data, measured in relevant tissue in the same individuals for whom we have GWAS data, could potentially provide further insight into disease pathways.
We review currently available statistical methods for inferring causality between variables that include a genetic variant, which can be used to anchor the direction of the causality. We consider Mendelian Randomisation, Structural Equation modelling, a Causal Inference Test and several Bayesian methods. We present a simulation study assessing the performance of the methods under different conditions, assuming throughout that we have genotype data along with two observed phenotypes. In particular, we consider how the causal inference is affected by the presence of common environmental factors influencing the observed traits.

Extracting Genetic Determinants from De Bruijn Graphs in Bacterial GWAS
Magali Jaillard, Maud Tournoud, Leandro Ishi, Vincent Lacroix, Jean-Baptiste Veyrieras and Laurent Jacob

Antimicrobial resistance has become a major worldwide public health concern, calling for better definition of existing and novel resistance mechanisms. GWAS methods applied to bacterial genomes showed encouraging results for new genetic marker discovery. Most existing approaches either look at SNPs obtained by sequence alignment or consider sets of k-mers, whose presence in the genome is associated with the phenotype of interest. We propose an alignment-free GWAS method, targeting any region of the genome and selecting genotypes of variable length associated to the resistance phenotype, using De Bruijn graphs. While this graph implicitly contains all k-mers of all sizes, its structure allows to drastically reduce the number of feature to explore, without loss of information, thus increasing the statistical power of the tests. In particular considering De Bruijn graph nodes instead of k-mers reduces up to 30 times the parameter space, even for species with a high genome plasticity such as Pseudomonas aeruginosa.

Longitudinal genetic modelling: revisiting associations of SNPs associated with blood fasting glucose in normoglycemic individuals
Mickaël Canouil, Ghislain Rocheleau, Loïc Yengo, Philippe Froguel

New statistical methods need to be proposed as an alternative to the current cross-sectional design predominantly used in genome-wide association studies (GWAS). When longitudinal (repeated) measures of a trait are available, an ecient modelling of the temporal trajectories is expected to increase statistical power to detect genetic loci associated with that trait.
Using genotypes assayed with the Metabochip DNA arrays (Illumina) from 4,500 subjects recruited in the French cohort D.E.S.I.R. (Données Épidémiologiques sur le Syndrome d'Insulino-Résistance), we re-examine published GWAS ndings for some conrmed loci associated with fasting plasma glucose (FPG).

We compared several approaches to test the SNP main eect, on the one hand, and to test the interaction SNP-by-time eect, on the other hand. For the former, we compared ve methods: linear regression models using only baseline measures or using the average of measures across all time-points, Two-Step approach with random intercept, Generalised Estimating Equations (GEE) and Linear Mixed Model (LMM); while for the latter we compared Two-Step approach with random slope, Conditional Two-Step, GEE and LMM with interaction term.
Type I error and power were computed using permutations and resampling procedures on the full dataset for the SNP eect, and using numerical simulations for the interaction eect.
Across all models tested, the type I error was not inated. In contrast, power analysis sometimes showed an increased statistical power for the baseline approach compared to methods dealing with repeated measures of FPG. We provide mathematical conditions showing why this counterintuitive situation might happen.
In the context of large GWAS with millions of imputed SNPs and when repeated measures are available for exploration, implementing methods which approximate a full longitudinal model seems at present the most ecient and fastest way to identify genetic associations without major loss in power. More importantly, these approximate methods run much faster than the full modelling approaches like GEE or LMM and could help picking the most associated SNPs for further testing in full models.