

Trip report- Ushma Galal

Statistical Genetics Short Course (featuring Mendel software)

22-26 August 2011, Barcelona, Spain.

The workshop/short course was jointly hosted by staff members of the University of California, Los Angeles (UCLA) and the Spanish Centro Nacional de Análisis Genómico (CNAG). The presenters were:

UCLA Faculty: Steve Horvath, Ken Lange, Jenny Papp, Janet Sinsheimer, Eric Sobel, and Ram Sripracha (IT)

NCSU Faculty: Hua Zhou

CNAG Faculty: Simon Heath

The purpose of the course was to present some background in genetics & statistical theory, then give an overview of the analysis program Mendel Enterprise, which is being developed by the presenters. It is a user-friendly interface for the program Mendel, and offers more features than Mendel. For example, it can be used for data management, pedigree drawing and editing, as well as statistical analysis. The two biggest contributors to Mendel & Mendel Enterprise are Ken Lange and Eric Sobel, who have developed other programs for the analysis of genetic data. The one we use most for the analysis of our family genetic data is SimWalk, which was created by Eric Sobel. At the moment, SimWalk is the only software package available that can do certain analyses, such as haplotype inference, on the extended families we have in our data. Other packages cannot handle the large number of family members we have.

Overview of Mendel

Mendel is a comprehensive package for the statistical analysis of qualitative (categorical) and quantitative (numerical) genetic traits. For pedigree data, it internally incorporates both the Elston-Stewart and the Lander-Green-Kruglyak algorithms. This means that for some applications it will choose, pedigree by pedigree, whichever algorithm is faster. This is an advantage for us since other packages incorporate only one of the two types of algorithms, implying that they are not always suitable for our data. For example, SimWalk, which incorporates the Elston-Stewart algorithm, can analyse large, complex pedigrees but it can only take a small number of SNPs due to the computation time required. On the other hand, packages such as Merlin, which are based on the Lander-Green-Kruglyak algorithm, are faster and can take more SNPs but have very large computer memory requirements so they can only be used with small, simple families. As a trade-off, Mendel can take medium-sized families of any level of complexity, and 1000's of SNPs. I am not yet sure if it will analyse our large pedigrees; however the developers are confident that it will since we work with small numbers of SNPs (less than 100).

Mendel coordinates with SimWalk, which performs many of the same tasks by stochastic sampling. Mendel also incorporates an enhanced version of the variance component program Fisher for QTL (quantitative trait loci) mapping and classical biometric genetics. There is a Windows version of

Mendel and it can be accessed through the shell program Gregor. Most importantly for us, Mendel gives effect sizes for the regression analyses it performs. This is not the case for some other packages that can be used for family genetic data.

Mendel can be used for data checking, family-based and population association tests, as well as linkage analysis. In addition, there are new features such as pedigree trimming (where uninformative family members are removed from the data prior to analysis) and imputation. For SNP association, Mendel can do SNP-SNP interaction and generate a p-value indicating significant interactions

The Course

Many of the talks focused on genome-wide association studies (GWAS). Thanks to new technologies for gene expression, genotyping, and sequencing, the nature of research is changing and new problems are being defined. There is now a phenomenal amount of data available and fortunately the speed and storage capacity of computers keeps improving or there would be no way to store and manage this data. However, since generating and managing this data is no longer a challenge, the next step is finding ways to analyse the data and interpret the results. This is part of the reason for the development of MendelEnterprise, a tool for data management and statistical analysis. Since Mendel can take a large amount of SNP data, Mendel and MendelEnterprise are useful for genome-wide studies in which there are millions of SNPs.

According to the presenters, GWAS are currently supplanting traditional linkage analysis, to our detriment, as some genetic effects are picked up by linkage but not association. So they say it pays to do both. In addition, more attention is currently being paid to interactions and rare variants, and case-control studies are replacing pedigree studies. However, they do say that they foresee linkage analysis making a comeback. They anticipate a movement back to mapping rare Mendelian disease genes, since we are now able to do this. Thus we should see more of this in the next decade.

The presenters discussed different sources of variation in data, which can lead to false associations. For example, you must be very careful where your sample populations come from, otherwise you might get a case-control association which is actually due to something else, such as geography.

Some time was spent on discussing the quality of data used for analysis. They emphasised that having enough power to find genes for complex traits requires, not only a large amount of data, but also clean data. Mendel can be used to carry out quality control tests on large datasets.

They went through the different types of input files required for Mendel and their structure. They then they demonstrated the use of MendelEnterprise with example datasets which they provided.

They discussed imputation in the case of genotypes which are missing not completely at random (MN-CAR). Imputation describes a set of techniques to infer missing SNPs based on the genotypes of the typed SNPs. They suggested the use of the HapMap data for imputation.

My Experience

One thing I learnt, which was surprising for me, is that South Africa isn't the only country battling to get statisticians involved in analysing genetic data. Countries like Belgium and Malaysia also have a shortage of statisticians with such skills, particularly when studies involve families. This has forced biologists in such countries to learn statistics, a task they find very challenging. Prof. Dr. Hilde Peeters from the University of Leuven is one such biologist. She works with families who have a history of Autism. She was at the point where she was trying to understand the mixed-effects models used to analyse such data. Since I had written up these models in my MSc thesis, we had many conversations about this topic. I also sent her a copy of my thesis after I returned to South Africa, as she really felt a document like that would help her learn what she needed to. Dr Peeters expressed an interest in collaborating with us in future.

The UCLA team came with their own IT expert, Ram Sripracha, which was a huge advantage as any technical problems could be resolved quickly and efficiently. MendelEnterprise was only available to the course participants in our venue, as it was set up for us on a server which Ram brought with him. As a result he could handle any problems immediately. Mendel is an open-source package and is accessible to Windows users through an interface called Gregor. MendelEnterprise does not appear to be available to the public but will be available to the course participants for the next year. Thereafter, if we wish to use it, we need to contact them to sort something out. For larger and longer-term projects, they are willing to help us install MendelEnterprise at our own institutions, as it currently runs off the UCLA server.

It was a wonderful experience being taught and interacting with people like Ken Lange and Eric Sobel who have so much experience in this field. Being at the course meant I could also sit down with them and discuss the problems we are having, in terms of the analyses we are trying to do and the challenges we face with finding software to do those analyses. Eric Sobel spent time with me trying to sort out some of the problems we are having with SimWalk. He gave me some suggestions as to how to solve these issues. Unfortunately, neither he nor Ken could help me with how to do haplotype analysis on X-linked traits. This seems to be something that is not done for family data, and so there is no software available to do it. We hope that, after speaking to me, they will be inspired to implement it into Mendel.

Although the workshop was very intensive, they did take us to see the Barcelona super-computer, MareNostrum, one afternoon. It is housed in a former church and was within walking distance of the course venue. The super-computer is 5 years old. When it was built, it was in the top 5 super-computers in the world. It is now ranked over 100 on the list! However, it is being upgraded so it should be more highly ranked once this is done. It consists of 400 terabytes and has one file system. It is made of standard components, so the 2500 boards need to be changed every two years. It contains over 30 000 processors, enabling it to compute in 1 hour what a standard laptop will take a year to do.

I feel very privileged to have had the opportunity to attend this course and interact with the people I met there. I am very grateful to the Biostatistics Unit and to the Statistics Department at UWC for giving me this opportunity. I have just been given access to MendelEnterprise again and I will be testing it out on some of our data.