Contribution ID: **179** Type: **Poster**

Genome Wide Association Studies of human complex diseases with EGEE

Describe the scientific/technical community and the scientific/technical activity using (planning to use) the EGEE infrastructure. A high-level description is needed (neither a detailed specialist report nor a list of references).

As part of the research conducted at INSERM U525, the THESIAS software was created in order to analyze statistically, associations between gene polymorphisms and diseases. Given a data set containing the genotypes of case and control individuals, THESIAS measures haplotype frequencies combining several polymorphisms and associations with the disease. This research can lead to the identification of new causes and mechanisms of disease of potential therapeutic and preventive interest.

Report on the experience (or the proposed activity). It would be very important to mention key services which are essential for the success of your activity on the EGEE infrastructure.

For whole genome haplotype analysis we decided to test the EGEE grid. EGEE provides enough computational power to perform these analyses. The EGEE infrastructure was used in our project and proved to be very effective. As a proof of principle, we have analyzed thousands of SNPs for their association with cardiovascular disease in thousands of individuals, and found that the performance improvement relative to a modern PC is about 300. Nevertheless, better reporting and analysis of failures would be very desirable.
A user interface on top of the gLite user interface has been created to simplify batch job submissions, monitoring and automatic resubmission of failed jobs. We have created this interface, which is very easy to use, and allows non computer scientists to use the EGEE grid ressources. We plan to analyse the entire genome as soon as we will have the data.

Describe the added value of the Grid for the scientific/technical activity you (plan to) do on the Grid. This should include the scale of the activity and of the potential user community and the relevance for other scientific or business applications

Until now this kind of analysis was limited to single genes and a few polymorphisms (<25). The recent availability of DNA chips allowing to genotype hundreds of thousands of polymorphisms implies a change in scale in the necessary computations.
The complexity of identifying which DNA sequences variations(SNPs) are associated to a disease on the entire human genome increases exponentially with the number of SNPs. Frequencies of combinations of multiple SNPs must be estimated. Ideally all the possibilities would be analyzed. With at least 10 millions SNPs on the human genome, calculating all the combinations is hardly imaginable. Fortunately, SNPs located close to each other are frequently tightly correlated (linkage disequilibrium); they define haplotype blocks that can be tagged by a limited number of marker-SNPs. The most recent genotyping arrays with 1 million marker- SNPs

are highly informative. The computation may be further reduced by investigating haplotypes in a sliding window

Author: MUNTEANU, Alexandru (INSERM UMR S 525)

Co-authors: GERMAIN-RENAUD, Cecile (LRI); TREGOUET, David (INSERM UMR S 525); CAMBIEN, François (INSERM UMR S 525)

Presenter: MUNTEANU, Alexandru (INSERM UMR S 525)

Track Classification: Demo and Poster session