Proceedings of the Joint Meeting AGI-SIBV-SIGA
Assisi, Italy – 19/22 September, 2011
ISBN **978-88-904570-2-9**

Oral Communication Abstract – 3B.02

IMPROVING THE ACCURACY OF DEMOGRAPHIC MODELS USING WHOLE-GENOME SEQUENCE DATA

COLONNA V.*,**, XUE Y.**, TYLER-SMITH C.**

*) Institute of Genetics and Biophysics "A. Buzzati-Traverso", National Research Council, Naples (Italy)
**) The Wellcome Trust Sanger Institute, Hinxton (UK)

Whole- genome sequence, human demography, 1000 Genomes project

Accurate knowledge of human genome sequence variation is important for the study of human disease inheritance and understanding of the history of human populations. Demographic models inferred from genetic data can explain patterns of genome variation and have been successfully used to decipher the history of human migrations within a framework of archaeological evidence. Demographic models are also essential to distinguish signals of natural selection from patterns of neutral variation resulting from demographic processes. This has important implications for mapping genes underlying complex human diseases.

Current models represent simple approximations to the true complexities of human history. Still, in some respects, they have also offered good fits to empirical data and often the quality of the empirical data has been the limiting factor. Indeed, it has been shown that the use of unbiased empirical data sets can considerably improve the accuracy of model predictions. A major bias in empirical data comes from the fact that most consist of Single Nucleotide Polymorphisms (SNPs) that were discovered in a small set of samples then genotyped in a larger set, so that rare alleles tend to be missed or over-represented and a description of the true distribution of allele frequencies is not achieved. Data sets for analysis of evolutionary history without this bias do exist but have been small and not representative of the whole human species. Thus, significant contributions and new exciting insights are expected from the use of unbiased empirical data.

The 1000 Genomes Project now provides a nearly unbiased resource describing human genetic variation, including SNPs and structural variants. Within this project, genome sequences of 2,500 anonymous people from 27 populations distributed around the world are being produced. At the moment, the 1000 Genomes Pilot Project has generated low-coverage sequence data over the complete genomes of 179 individuals from four HapMap samples, discovering 14.5 million SNPs, 8 million of which are novel, with their genotype in each individual, and many structural variants.

We are exploring ways of calibrating the existing models of human evolution using the newly-available pilot sequence data from the 1000 Genomes Project, with a view to then applying these methods to additional populations.