Kelly Sheppard
December 12th, 2001
MB&B 452a / 752a2

Large Scale SNP Association Studies: The Importance of Databases

Single nucleotide polymorphisms (SNPs) account for the majority of DNA sequence variation in humans [1]. SNPs are defined as a "stable substitution of a single base with a frequency of more than 1% in at least one population" [2]. Recently completed work on SNP analysis of chromosome 21 has yielded researchers a clue to the topology of SNPs found within the human genome[3,4]. Research is moving now towards doing large-scale genotyping of individuals, in order to associate SNPs with certain disease phenotypes. Bioinformatics tools will be absolutely necessary in order to make these studies meaningful. The most important aspect of which will be designing databases that are capable of handling this load of information in a manner beneficial to researchers.

SNPs are found throughout the human genome, in both coding and non-coding regions. The methods for detecting and discovering SNPs in a population are varied including sequencing, use of microarray technology, PCR assays such as TaqMan and Invader, and mass spectrometry [5]. Large scale SNP discovery has been undertaken by a group comprising of private companies and the public Human Genome Project under the name The SNP Consortium (TSC). Private companies such as Celera have gone out on their own to discover SNPs. All these efforts have created databases to make the information more manageable. In the public databases there is in the range of around two million SNPs [6]. On NCBI, db-SNP (http://www.ncbi.nlm.nih.gov/SNP/) provides researchers with a wealth of information on a given SNP, such as where it is in the human genome, what gene it is in, the method it was detected by, and links to more information on the region of interest.

db-SNP has provided researchers on a small scale a place to begin to try to associate SNPs with diseases. Association studies have been done on APOA genes[7], genes related to narcolepsy[8] and the Tau gene in relation to Parkinson disease[9] to name a few. The goal is to find SNPs that might cause a disease, influence the chance of developing a disease, and/or that could alter a person's response to a treatment, mostly to a certain drug[5]. Through these smaller scale studies only a handful of meaningful SNPs have been found[2].

In order to help speed up the process of associating SNPs with certain phenotypes more genes need to be analyzed at any given time. For complex diseases such as those related to heart disease, association studies are needed in order to distinguish meaningful SNPs, which requires large sample sizes. Projects undertaking large-scale (i.e. examining multiple genes) association studies therefore will generate and collect volumes of data that must be correlated together in an efficient manner. This is where the next generation of SNP relational databases will come into play.

At Stanford University, one such a large scale project, the Reynolds Project[10], is underway to understand atherosclerosis and in particular the role vascular genes play in the disease. The number of such genes is still unknown and a multitude of other factors play a role in the disease such as genes related to lipid metabolism and environmental factors like diet and exercise. Atherosclerosis is therefore a complex disease that requires large numbers of diseased and non-diseased individuals to be studied on a clinical epidemiological and genetic level in order to correlate genetic variation with phenotype changes in a statistically meaningful manner. The key in such a project is properly organizing and inputting the data generated to allow for such correlations.

For the Reynolds Project this is no small task given the amount of data that is to be obtained. The project is subdivided making it easier to develop a database system. First candidate genes must be identified and selected. To do this different cell types are stimulated by a series of conditions that mimic the disease etiology in order to isolate genes differentially expressed which then will be cloned, in the end creating a library of potential genes related to the disease including novel genes. The library enables them to spot the genes on glass slides for microarray analysis of RNA from the vascular walls of diseased and healthy individuals; in the process, select candidate genes for future study. The first part of the project requires a database table of the various cell types, what was expressed under certain conditions, and what genes were detected. It also then requires a table of the array information. The tables being linked by the gene names.

The second part of the project involves taking the candidate genes found and determining if there are any SNPs within the gene in the human population and to select certain SNPs for genotyping. This is done in silico, searching the public SNP databases, and by designing primers to sequence the gene from a small set of individuals. From there genotyping is done on the populations in the association study. This requires a table that indicate how the SNP was obtained, from outside sources or in-house sequencing, what the variation is, and the flanking sequence/position within the human genome and another table indicating the genotypes of each individual in the study for each SNP. The particular SNPs being the key to transverse the two tables. The third part of the project involves collecting the DNA samples from the individuals and collecting their complete medical histories in a tabulated format.

The Reynolds Project has a grant of $24 million over a four year period. Obtaining accurate results is imperative. Inaccurate data entry can totally void correlations drawn from CART algorithms and related k-means clustering adaptations[10]. The properly designed database can minimize these problems. Each subset of the project can have its own interface to enter data into forms that makes it as easy as possible and to create subset specific reports to monitor progress. The links, within and between the subsets, also play a vital role. The candidate gene selection subset is linked to the genotyping through the gene names and the genotyping to the clinical through the individuals. This allows reports to be made across the subsets allowing for the correlations to be made between the genetic and the clinical data. The Reynolds Project is well on its way of doing this through a Sybase database that will be accessible through web browser software, with the aim down the line to open the information up to other research groups to carry out follow up studies and experiments.

Large-scale association studies of SNPs to disease will require databases containing far greater and more accurate information than is currently within SNP databases at the moment, as seen by the Reynolds Project at Stanford, as the genetic information is connected to clinical data. The failure to develop these databases in such a manner that the data within them is properly linked and in an easy to use fashion will have dire consequences. Improper data entry or missed links between the various fields will either cause incorrect correlations to be made or prevent clinically useful ones from being seen and developed further.

[1] Venter, J.C. et al. (2001) The sequence of the human genome. *Science* **291**. 1304-1351

[2] Taylor, J.G., Choi, EH, Foster, C.B., and Chanock, S.J. (2001) Using genetic variation to study human disease. *TRENDS in Molecular Medicine* **7**. 507-512.

[3] Patil, N. et al. (2001) Blocks of Limited Haplotype Diversity Revealed by High-Resolution Scanning of Human Chromosome 21. *Science* **294**. 1719-1723.

[4] Olivier, M. et al. (2001) Complex High-Resolution Linkage Disequilibrium and Haplotype Patterns of Single-Nucleotide Polymorphisms in 2.5 Mb of Sequence on Human Chromosome 21. *Genomics* **78**. 64-72.

[5] Syvanen, A.C. (2001) Acessing Genetic Variation: Genotyping Single Nucleotide Polymorphisms. *Nature Genetics* **2**. 930-942.

[6] Sachidanandam, R. et al. (2001) A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* **409**. 928-933.

[7] Pennacchio, L.A., Olivier, M., Habacek, J.A., Cohen, J.C., Cox, D.R., Fruchart, J.C., Krauss, R.M., and Rubin, E.M. (2001) An apolipoprotein influencing triglycerides in humans and mice revealed by comparative sequencing. *Science* **294**. 169-173.

[8] Olafsdottir, B.R., Rye, D.B., Scammell, T.E., Matheson, J.K., Stefansson, K., and Gulcher, J.R. (2001) Polymorphisms in hypcretin/orexin pathway genes and narcolepsy. *Neurology* **57**. 1896-1899.

[9] Martin, E.R. et al., (2001) Association of single-nucleotide polyphmorphisms of the tau gene with late-onset Parkinson disease. *JAMA* **286**. 2245-2250.

[10] Ralston, L. (Webmaster) Quertermous, T. (2001). Summary of the Reynolds center at Stanford University. http://cvmed.stanford.edu/Reynolds/reysum.htm. Last updated Febraury 26th, 2001. Contains summary of the overall project with links to the subprojects and was the basis for the overview of the project presented in this paper.