Data Management and Genome Wide Association Screening

ID - 222
Orro Alessandro[1], Guffanti Guia[2], Salvi Erika[3], Macciardi Fabio[2], Milanesi Luciano[3]
[1]Consorzio Interuniversitario Lombardo per l'Elaborazione Automatica, Segrate
[2]Istituto di Tecnologie Biomediche, Segrate
[3]Dipartimento di Scienze e Tecnologie Biomediche, Università di Milano

Motivation

Recent progresses in genotyping technologies allow to generate high-density genetic maps using hundreds of thousands of genetic markers for each DNA sample. The availability of this large amount of genotypic data makes a whole genome search for genes underlying genetic diseases considerably easy. In particular, one of the most common forms of human genetic variation, Single Nucleotide Polymorphisms (SNPs), can be used to discover the sequence variants affecting common diseases by examining them for statistically significant association with measurable phenotypes.

To efficiently manage the data flow produced by whole genome genotyping and make it available for further analyses, we need a suitable information management system.

Methods

We have developed an information system mainly devoted to the storage and management of SNP genotype data produced by the Illumina platform from the raw outputs of genotyping into a relational database. Data reports can be extracted and used as input for the genetic analyses. Particular attention has been given to whole genome screening analyses that allow selecting the set of markers with high degree of statistical significance for the disease under consideration.

The main features of the system are: (1) automatic import of genotype data, (2) definition and assignment of phenotypes to the subjects, including both qualitative and quantitative traits, (3) control of the quality of the data in order to select markers with high genotyping score, (4) analysis of the genetic population structure to identify stratification and/or admixture (sub-populations), thus avoiding false association between markers and phenotypes, (5) statistical descriptive analysis (Hardy-Weinberg equilibrium test, Minimal Allele Frequency - MAF), (6) single point analysis of association between genotype and quantitative or qualitative traits, and (7) multi locus analysis to combine genotypes of adjacent markers and find associations between haplotypes and phenotypes.

Results of the analyses can be summarized in reports that allow visualizing and selecting significant SNPs and genes. Results can also be exported in CSV format, so that it is easy to use them in other analyses. Data are organized in sessions of acquisition and analysis so that multiple studies can be managed in logical projects and shared between users.

Results

The proposed infrastructure allows managing a relatively large amount of genotypes for each sample and an arbitrary number of samples and phenotypes. Moreover, it enables the users to control the quality of the data and to perform the most common screening analyses and identify genes that become 'candidate' for the disease under consideration.

Email: alessandro.orro@itb.cnr.it