Motivation

The identification of genes and SNPs involved in human pathologies remains a challenge. Although public resources collect biological data for increasing the global biological knowledge of specific diseases, the need for tools oriented to SNPs prioritization is emerging due to the lightning diffusion of new high-throughput genotyping technologies. Even if statistical methods are widely used in epidemiology and many positive results have been achieved in genome-wide association studies, they show clear limits. First, they only allow the identification of regions involved in a disease, due to the necessity of selecting a restricted set of TAG SNPs as markers for larger lists; second, they are prone to bias in the selection of the study population; lastly, they are often very computationally intensive, especially when dealing with large amounts of data produced by high-throughput techniques. FastSNP [Yuan et al., 2006] and F-SNP [Lee et al., 2006] are examples of available tools that prioritize a list of SNPs but none of them is oriented to investigate relations between SNPs and pathologies. The presented work concerns a system aimed at supporting both the analysis of SNPs associated to particular pathologies/pathways and for designing custom chips for specific disease-oriented studies. It relies on the identification of a set of crucial features characterizing SNPs related to an input dataset of genes or genome-wide association studies, while the output is a ranked list of SNPs. Although some databases and resources are available and represent the primary source for the development of the models employed in this work, a user-friendly, integrated, powerful and reliable resource for SNP ranking is still an unmet need for the scientific community interested in genetics and epidemiology.

Methods

Starting from a list of SNPs, genes or a biological process involved in specific diseases, the resource allows the identification (when needed) and the ranking of an annotations enriched SNPs’ set. When using genes, the initial input list can be expanded through an ontology-based engines based on biological processes and biomolecular pathways. Moreover, the system takes into consideration all the SNPs occurring within coding and regulatory regions. For each SNP associated to the input dataset the tool scores different biomolecular features, which represent the a-priori knowledge. Among crucial features appear: snp localization on gene (dbSNP [Smigielski et al., 2000]), epigenetics experimental data concerning several cell lines (UCSC [Fujita et al., 2010]), knowledge about alternative splicing and transcription regulation factors (UCSC), protein domain...
characterization (SwissProt [Boeckmann et Al., 2003]), genetic disease evidence (OMIM [McKusick, 1998]) even based on genome wide analysis studies (GWASStudies [Hindorff et Al., 2009]). Also the linkage disequilibrium information is maintained in the database and appears among the annotations of ranked SNPs. The final score is computed by weighting each feature through an automated approach, which can be modulated by the user according to the scientific content of the experiment. The system provides a ranked list of SNPs, together with their relevant annotations.

Results
The SnpRanker tool has been validated using OMIM data, considering a set of pathologies influenced by recognized SNPs. For each disease the list of associated genes has been given as input for the system and the list of retrieved ranked SNPs has been compared to the set of SNPs provided by OMIM for the same disease. In the considered cases, such as cystic fibrosis, Sickle cell anemia, hemophilia, the top ranked SNPs show a statistically significant enrichment (P < 0.05, hypergeometric test) concerning SNPs known to be associated to the tested pathologies. In Huntington’s disease the first three SNPs appearing in the ranked list are exactly those reported in OMIM for this pathology.

Availability
http://www.itb.cnr.it/snpranker/

References

Contact email
federica.viti@itb.cnr.it