ORIGINAL ARTICLE

WEB-BASED SOFTWARE FOR STORAGE, STATISTICAL PROCESSING AND ANALYSIS OF SNP DATA IN STUDIES ON COMPLEX DISORDERS

Betcheva E[1], Betchev C[2], Toncheva DI[1,*]

***Corresponding Author:** Draga Ivanova Toncheva, M.D., Department of Medical Genetics, Medical Faculty, Medical University, 2 Zdrave str., SBALAG "Maichin dom", 6 Fl., 1431 Sofia, Bulgaria; Tel./Fax: +35-92-952-0357; E-mail: dragatoncheva@yahoo.com

ABSTRACT

Single nucleotide polymorphisms (SNPs) have become a very powerful tool for molecular genetics studies. Public databases provide information on over 10 million polymorphisms in the human genome. The candidate gene approach and genome-wide association studies through SNP analysis have opened a new avenue for defining the genetics of complex disorders. However, analysis of large numbers of SNPs is time-consuming, cost-intensive, and requires huge experimental and statistical resources in association studies. We have developed a web-based product that facilitates the processing and statistical analysis of SNP-genotyping data for case-control association studies and provides for custom design, a structured database and practical export layout. Here we describe the software product database and how it helps in high-speed comprehensive SNP analysis.

Key words: Association study; High-throughput genotyping; Information technology (IT); Multifactorial (complex) disorders; Single nucleotide polymorphisms (SNPs); Web-based software

[1] Department of Medical Genetics, Medical University, Sofia, Bulgaria

[2] Department of Menagement and Marketing, Technical University, Varna, Bulgaria

INTRODUCTION

Large bodies of experimental data from family, adoption and twin studies suggest a genetic component of the individual differences in susceptibility to complex disorders. It is clear that multifactorial disorders are, in part, heritable and their etiology results from a complex interaction between environmental and genetic factors [1,2]. In contrast to the single gene (Mendelian) disorders, they have more compound pathogenesis. According to the contemporary models, the potential effect of many genes and genetic variants in several different loci determines genetic susceptibility to such disorders [3]. Emerging data from linkage and association studies support the hypothesis that the triggering effect of certain environmental risk factors, such as a particular lifestyle, might provoke phenotype expression when affecting individuals with certain genetic background [1].

Intense interest has been focused on genome-based studies of complex diseases and accelerated with the completion of the human genome project and the progression of advanced technologies. Comparison of the DNA sequences of people from the major population groups has established a comprehensive map of genetic variants in the human genome, which conveniently serve as genetic markers. Detailed information about genetic diseases, genes, sequences and a great variety of polymorphisms is available in on-line public databases and provides an irreplaceable tool for molecular genetic studies [4,5].

The quest for genetic factors in the susceptibility to complex disorders has focused on single nucleotide polymorphisms (SNPs) which are the most common type of genetic variants in the human genome and occur in approximately every 100 to 300 bp [6,7]. Most SNPs have only two possible alleles that differ between the individuals of the same population group, where the frequency of the minor allele is usually specific. Although SNPs offer a limited number of possible alleles, which is a prerequisite for the selection of markers for DNA analysis, they are very convenient and highly informative for haplotype analysis, because of their abundance (over 10^6 deposited in the dbSNP database of the National Center for Biotechnology Information (NCBI): http://www.ncbi.nlm.nih.gov/About/primer/snps.html) and their genetic stability in the human genome [7-9].

The SNPs occur within coding gene regions, non coding intra- and intergenetic sequences. Most fall in introns, untranslated 3' and 5' regions (UTR3' and UTR5') of the genes and spacer DNA [8]. Although they do not cause gene product modifications, some may play an important role in the control of gene transcription level, by influencing the affinity of promoters or other regulating sequences to trans-regulating factors that modify the gene expression rate, or by affecting pre-mRNA processing [6]. A small portion of SNPs are in coding DNA sequences, however, most are synonymous, *i.e.,* they do not alter the polypeptide structure and only a few are non synonymous SNPs, causing an amino acid exchange [8]. The distribution of SNPs could be explained by a negative effect on survival and fast elimination of expressed variations by natural selection [6].

Single nucleotide polymorphisms are associated with population diversity and individual differences in complex traits [6,8]. Therefore, they are convenient for genetic association studies on identification of susceptibility loci for multifactorial disorders. An association between a disorder and a non synonymous SNP makes the pheno-type-genotype relationship very clear. However, an association with a synonymous SNP or a SNP in a non coding sequence is difficult to explain, and usually another causative marker needs to be identified [7,8,10].

Currently, SNPs are preferred as genetic markers in case-control and whole genome association studies. They have been used in studies for mapping and discovery of susceptibility genes for many complex disorders: cardiovascular (essential hypertension), neurological (Alzheimer's disease, multiple sclerosis), psychiatric (schizophrenia, bipolar affective disorders), autoimmune (rheumatoid arthritis) disorders, diabetes mellitus type 2, and different types of cancer [11,12]. Linkage studies and genome scans have identified several candidate chromosomal regions for common diseases [7,13]. Selection of SNPs in such loci has become a basic approach in candidate gene(s) association studies [7]. However, the candidate gene approach, is time-consuming, cost-intensive, and insufficient, and has largely failed in prediction of risk for disease susceptibility, since only a limited number of genetic markers in a relatively small region are investigated [7,12]. Results from meta analyses are often inconsistent and demonstrate the need for more efficient and cost-effective high-throughput SNP genotyping technologies, such as DNA-microarray-based technology, for revealing disease causing genes [7,9,12,13].

Application of DNA-microarray technologies in large-scale studies of complex disorders facilitates genotyping of large number of SNPs. A DNA chip consists of an arrayed series of thousands of sequences for detection of tag SNPs from the entire genome [13]. Selection of population-specific tag SNPs has become available since the haplotype block structure of the human genome was established in the International HapMap Project (www.HapMap.org). Tag SNPs are representative markers for a set of variants within a region of high linkage disequilibrium in the genome. Thus, they are useful for economical and efficient genotyping of a relatively small number of markers which provide adequate information on disease-associated genes and loci. Candidate genes identified by such large-scale approaches require further analysis, to elucidate their role in disease etiology (http://www.ornl.gov/sci/techresources/Human_Genome/faq/snps.shtml) [13]. A whole genome association study based on array technology produces large amounts of data and requires a sufficient database, appropriate computational statistical methods, techniques for false-positive error detection and maintenance. Moreover, the use of such technology is allied to high costs and significant time, effort, and resource consumption.

We have performed a whole genome association study (WGAS) of DNA samples from unrelated

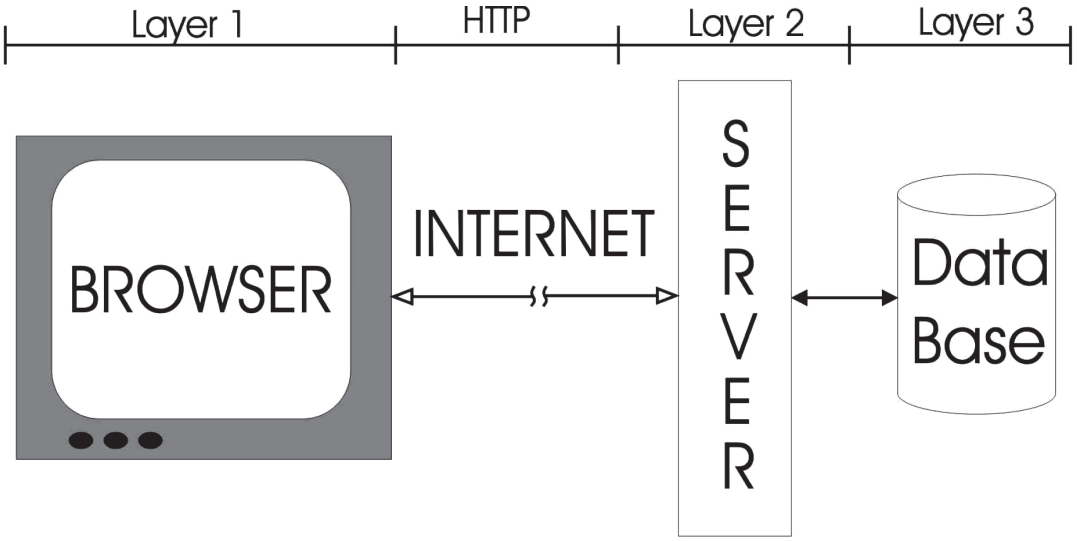
Bulgarian patients with schizophrenia and healthy volunteers (unpublished data). Subsequently to the WGAS, the 100 top SNPs showing lowest p values were validated (genotyped by alternative method in the same samples) and replicated (genotyped in additional DNA samples). The large amount of data produced required comprehensive statistical analysis. For this reason we have created a client-server web-based application for statistical processing and for reliable storage of data from an automated genotyping study in a set of DNA samples as specified below.

METHOD AND DISCUSSION

The application we have developed is based on a three-layer architecture model (Figure 1). The client system is an arbitrary browser (Mozilla, Internet Explorer, Opera, Safari, *etc.*). The WEB server is Apache on Linux deploying PHP scripting language and MySQL database. Data are transferred from a client's computer to a host mail server as unconverted text file format with particular data structure (Figure 1).

The internet plays an important role in the development of novel systems and algorithmic models for information services application (ISA) [14]. Common internet services are as follows: electronic mail (e-mail), file transfer (ftp), and WEB hypertext transfer protocol for WEB-page browsing (http). Less popular but equally accessible are TELNET, an internet protocol for connecting to a remote server, and USENET, an on-line information interchange service. Connection and communication through the internet is possible regardless of differences in users' platforms (operating systems and hardware) and software. This operational principle enables the establishment of a two-layer model for simple communication between Client and Server. A more complex three-layer model for data interchange is required when abundant and specific information management and its reliable storage are expected. It includes a Server-accessible database for management and storage of large amounts of structured information (Figure 1).

Purposeful use of software and technology and application of computer-based information and communication systems to achieve maximum efficiency of specific task management procedures is defined as modern information technology (IT). This enables highly effective utilization of time and resources, and also reveals new opportunities for task performance. At present, use of the internet is an integral part of contemporary IT and of virtual environment for transfer of large amounts of data. This relates to various specific activities in the field of humanities (http://mysql.com/index.html) [16-18].



Fi

Figure 1. The three-layer structure of the software tool.

In order to reduce the cost and to preserve the power of the analysis, some authors designed the whole-genome approach in a two-step manner. A pilot fraction of samples is selected for high-throughput genotyping by microarray technology. Subsequently, a number of top markers is chosen for genotyping in a second sample set [13,15]. We have adopted such an approach for our study. Subsequently to execution of the validation and replication studies, a highly efficient and reliable statistical processing of genotyping data for 100 genetic markers in 1,000 DNA samples was required. The common procedure for data processing includes interventions such as manual transfer and conversion of text files (containing unnecessary additional information) in electronic spreadsheet (for example in Microsoft EXCEL), animated by macro commands, in order to evaluate certain quantities. Each manual procedure consumes considerable time and resources. It increases the risk of disruption by human error and of completely erroneous interpretations. The quantitative assessment of the obtained data is only an initial step which requires further mathematical processing. A useful approach is the creation of a database for storage of results, followed by data processing, moreover, not all data analysis procedures occur simultaneously. Since the SNP-genotyping machinery is designed for robot control and does not include resources for organization, storage and further processing of data, a three-layer software provides a solution. For these particular tasks an internet connection and an installed WEB browser are quite sufficient.

In our experimental work, the SNP-genotyping detection equipment employed a 384-well plate format. For technical reasons, DNA samples from four 96-well polymerase chain reaction (PCR) plates were used to compose a 384-well plate, where test samples are analyzed along with control samples. Thus, the resulting genotyping text file includes significantly perturbed data from probes of different groups (cases and controls). For the analysis of 100 SNPs in 1,000 DNA samples, we needed to prepare 400 plates of 384-wells, four plates with DNA samples from patients and healthy controls for each polymorphism. Consequently, statistical processing of 400 files with genotyping data was required.

Initially, we developed an appropriate WEB form, enabling the client (researcher) to input the IDs of patients and controls subjects, SNP IDs, the specific position in the PCR-plate of each subject, and the name of the file that contains the assay information (Figures 2A and 2B). From a drop-down menu at the top toolbar, a set of markers (SNPs) can be selected for analysis (custom design). Prior to testing, markers can be assigned into sets of SNPs of interest (*i.e.,* names of SNPs are entered into the database in advance). The SNPs are designated with their unique RefSNP (rs) code according to the NCBI dbSNP, which comprise at least four numbers and a typing mistake can easily occur. By inserting the SNPs of interest in advance the program is enabled to control for typing and other errors.

The plate's number toolbar is custom designed (Figure 2A), and allows the user to insert in a single step the design of all templates (384-well plates with DNA samples) used for the SNP genotyping. In our study, four types of templates (for 1,000 DNA samples) were designed. This step allows each position in a template (*i.e.,* well with a DNA sample) to be recognized as a specific ID number that corresponds to a certain patient or healthy control. Thus, in the database, the genotyping data from individuals with the disease will be separated from the unaffected subjects, and will be arranged according to the list of IDs.

The BROWSE button permits the genotyping data text file to be attached. Each text file obtained from the genotyping machinery, contains data from one 384-well template, where information on the position of up to 384 DNA samples, the alternating alleles of one SNP, the DNA quality, the detection rate quality and other is integrated. After browsing the selected file, the SEND button sends the text files and all descriptive data to the server via the Internet. Server software processess the acquired information and applies the decoding scheme in accordance with the experimental conditions. Interpreted data are converted into a format in compliance with their preliminary properties and functions, and recorded in the database at a predefined position. In other words, the template ID, the position of the DNA sample, the genotyping data and the ID of the patient or the control subject are recognized and matched, and can be preserved structured in a database.

The main page of the interface allows the client to abide for errors in the selection of SNP and template IDs and get information on the progress of

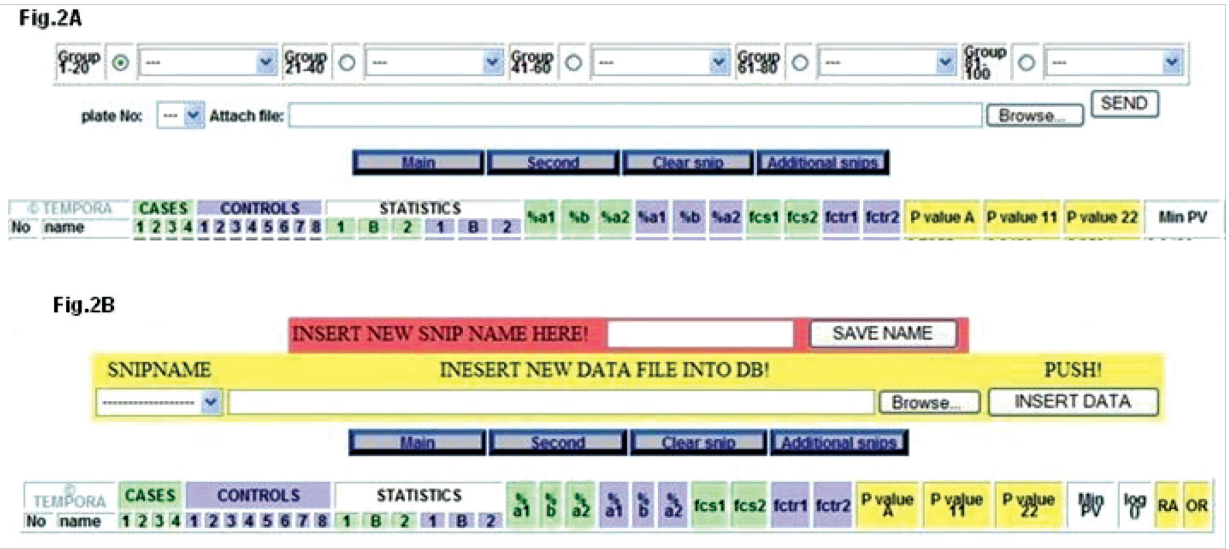


Figure 2. A/B) User interface. Part of the table, after the analysis completion of statistical processing, providing information on: SNPs rs number, identification of the template, number and type of sample (case or control), allele frequencies (% and number), number of homozygotes and heterozygotes, *p* values by Fisher's exact test, OR, risk allele (RA).

statistical processing. In brief, the statistical analysis is completed. The SECOND (part)-button enables quick link to page with precise statistical data in a suitable table format. A MAIN (page)-button enables returning to the main page (Figure 2B).

The Clear SNP-button enables erasing SNP data in DB, if necessary (in case of re-genotyping or detected errors) (Figure 2B). In case of accidental recording of new data over existing one, the client receives a message alert to accept or refuse a new entry. This prevents both duplicates and database disruption. Once stored into database, the data is transformed into a very convenient form for further processing.

Most of the statistical parameters of the investigated markers are obtained by standard operations of the database, whereas some are estimated by specific PHP commands. For each genetic marker the following parameters are presented in a practical format (Figure 2): ***i)*** allele and genotype frequencies in cases and healthy controls as absolute values and as percentage, in order to facilitate the comparison between genotyping data and data from the whole genome association study and the HapMap database; ***ii)*** statistical significance of the association between allele and genotype frequencies and phenotype expression, expressed in *p* values computed by the two-sided Fisher's exact test; ***iii)*** identification of risk allele (the allele associated with increased risk for phenotype expression; the allele that is more common in case samples compared to control samples); ***iv)*** odds ratio (OD) (a statistical measure of the strength of association between having the risk factor if the disease is present compared to if it is absent) in accordance with the risk allele; ***v)*** the 95% confidence interval (95% CI).

The main advantage of the described product compared to the common electronic spreadsheet approach is the opportunity for establishing a structured database, which may be further processed further if necessary. For example, haplotype analysis, evaluation of correlations within different subgroups of subjects according to their age, gender, drug therapy applied.

ACKNOWLEDGMENTS

We acknowledge the support of the Bulgarian Consortium for Structural Genomics and in silico drug design (Contr.No DRI-5/07.02.2006).

REFERENCES

1. Lang UE, Puls I, Muller DJ, Strutz-Seebohm N, Gallinat J. Molecular mechanisms of schizophrenia. Cell Physiol Biochem 2007; 20(6): 687-702.

2. McGuffin P. Gene polymorphisms and behaviour. Pediatr Blood Cancer 2007; 48(7): 736-737.

3. Kurland L, Liljedahl U, Lind L. Hypertension and SNP genotyping in antihypertensive treatment. Cardiovasc Toxicol 2005; 5(2): 133-142.

4. Ayme S. Bridging the gap between molecular genetics and metabolic medicine: access to genetic information. Eur J Pediatr 2000; 159(Suppl 3): S183-185.

5. Shi MM. Enabling large-scale pharmacogenetic studies by high-throughput mutation detection and genotyping technologies. Clin Chem 2001; 47(2): 164-172.

6. Wang X, Tomso DJ, Chorley BN, Cho HY, Cheung VG, Kleeberger SR, Bell DA. Identification of polymorphic antioxidant response elements in the human genome. Hum Mol Genet 2007; 16(10): 1188-1200.

7. Lee JE. High-throughput genotyping. Forum Nutr 2007; 60: 97-101.

8. Shastry BS. SNP alleles in human disease and evolution. J Hum Genet 2002; 47(11): 561-566.

9. Tamiya G, Shinya M, Imanishi T, Ikuta T, Makino S, Okamoto K, Furugaki K, Matsumoto T, Mano S, Ando S, Nozaki Y, Yukawa W, Nakashige R, Yamaguchi D, Ishibashi H, Yonekura M, Nakami Y, Takayama S, Endo T, Saruwatari T, Yagura M, Yoshikawa Y, Fujimoto K, Oka A, Chiku S, Linsen SE, Giphart MJ, Kulski JK, Fukazawa T, Hashimoto H, Kimura M, Hoshina Y, Suzuki Y, Hotta T, Mochida J, Minezaki T, Komai K, Shiozawa S, Taniguchi A, Yamanaka H, Kamatani N, Gojobori T, Bahram S, Inoko H.. Whole genome association study of rheumatoid arthritis using 27 039 microsatellites. Hum Mol Genet 2005; 14(16): 2305-2321.

10. Thorisson GA, Stein LD. The SNP Consortium website: past, present and future. Nucleic Acids Res 2003; 31(1): 124-127.

11. Gray IC, Campbell DA, Spurr NK. Single nucleotide polymorphisms as tools in human genetics. Hum Mol Genet 2000; 9(16): 2403-2408.

12. Grant SF, Hakonarson H. Recent development in pharmacogenomics: from candidate genes to genome-wide association studies. Expert Rev Mol Diagn 2007; 7(4): 371-393.

13. Steemers FJ, Gunderson KL. Whole genome genotyping technologies on the BeadArray platform. Biotechnol J 2007; 2(1): 41-49

14. Ivanova Z, Stoilova K, Stoilov T. Wallet optimization - information service in Internet. Marin Drinov: Academic Press. 2005; 276.

15. Hao K, Schadt EE, Storey JD. Calibrating the performance of SNP arrays for whole-genome association studies. PLoS Genet 2008; 4(6): e1000109.

16. Boar B. Implementing Client Server Computing: a strategic perspective. New York: McGraw-Hill 1992; 1947.

17. Cash J, McFarlau F, McKenney J. Corporate Informations Systems Management: The Issues Facing Senior Executives. Third Edition. Homewood IRWIN 1992; p 329.

18. Schussel G. Client/Server: past, present and future. (http://www.dciexpo.com/geos/).