ISSN 2079-0597, Russian Journal of Genetics: Applied Research, 2014, Vol. 4, No. 3, pp. 159–167. © Pleiades Publishing, Ltd., 2014.
Original Russian Text © N.L. Podkolodnyy, D.A. Afonnikov, Yu.Yu. Vaskin, L.O. Bryzgalov, V.A. Ivanisenko, P.S. Demenkov, M.P. Ponomarenko, D.A. Rasskazov, K.V. Gunbin, I.V. Protsyuk, I.Yu. Shutov, P.N. Leontyev, M.Yu. Fursov, N.P. Bondar, E.V. Antontseva, T.I. Merkulova, N.A. Kolchanov, 2013, published in Vavilovskii Zhurnal Genetiki i Selektsii, 2013, Vol. 17, No. 4/1, pp. 577–588.

Program Complex SNP–MED for Analysis of Single-Nucleotide Polymorphism (SNP) Effects on the Function of Genes Associated with Socially Significant Diseases

N. L. Podkolodnyy[a], D. A. Afonnikov[a], Yu. Yu. Vaskin[b], L. O. Bryzgalov[a], V. A. Ivanisenko[a], P. S. Demenkov[a], M. P. Ponomarenko[a], D. A. Rasskazov[a], K. V. Gunbin[a], I. V. Protsyuk[b], I. Yu. Shutov[b], P. N. Leontyev[b], M. Yu. Fursov[b], N. P. Bondar[a], E. V. Antontseva[a], T. I. Merkulova[a], and N. A. Kolchanov[a]

[a] *Institute of Cytology and Genetics, Russian Academy of Sciences (RAS), Novosibirsk, 630090 Russia*
e-mail: pnl@bionet.nsc.ru
[b] *Novosibirsk Center for Information Technologies UNIPRO, Novosibirsk, 630090 Russia*
Received 15 August, 2013; in final form, 5 September, 2013

Abstract—We describe development and application of the new SNP–MED modular software system, designed to examine the influence of single-nucleotide polymorphisms (SNPs) on the function of genes associated with the risk of socially significant diseases. The SNP–MED system includes Genomics, Proteomics, and Gene Networks' software components, and the Information Resource Database.

Keywords: bioinformatics; single-nucleotide polymorphism (SNP); personalized medicine
DOI: 10.1134/S2079059714030034

INTRODUCTION

Modern post-genome biology is characterized by a rapid development of high-performance experimental techniques, which allow us to determine parameters within the whole genome, transcriptome, or proteome in a single experiment. Microchip technologies, developed for DNA and transcriptome analysis, allow us to examine the expression patterns of tens of thousands of genes simultaneously. The new generation high-resolution mass spectrometry methods enable us to follow the dynamic changes in the cellular levels of RNA and proteins, and to test the continuous pipeline of small molecules being developed for various biomedical applications and medical treatments. The new genome-wide sequencing technology on the level of the whole organism (the new generation sequencing (NGS) technology) provides an inexpensive and effective way for sequencing individual human genomic DNA, whose cost is constantly decreasing, and could drop to less than $1000 in the near future. These biomedical advances enable a novel approach to medical treatments with a personalized medicine paradigm. Personalized medicine is the modern concept of individual health in the postgenomic era. It postulates that medical treatments must take into account the individual genetic characteristics, such as genetic predisposition to particular diseases and the individual response to various drugs. The foundation of personalized medicine is composed of the individual genetic information and the modern information technologies, designed to assess the individual risk for specific diseases, for instance, by considering the effect of individual gene polymorphisms.

Elucidation of the molecular mechanisms of genetic predisposition to various diseases, such as cardiovascular, cancerous, and neuropsychiatric conditions, are the major goals of modern medical genetics, molecular physiology, and pathology. To address these problems, a number of wide-scale studies have been undertaken around the world aimed at examining the relations between the variations in the genomic DNA sequences and pathologic conditions.

At the present time, a number of bioinformatics methods are being rapidly developed, which are designed to examine the impact of gene polymorphisms at the various levels of molecular genetic systems: the genome, transcriptome, proteome and gene networks.

Mutations of a single nucleotide (single-nucleotide polymorphisms (SNPs)) are the most common and most intensively studied types of DNA sequence polymorphisms. Today, the data on human SNPs is being mined in large quantities, in particular, in the projects aimed at full genome-scale association studies (Genome-Wide Association Studies (GWAS)) (Torkamani et al., 2008). In this context, it becomes possible to conduct an in depth examination of the roles of the gene polymorphisms in the occurrence of

socially significant diseases. These studies are based primarily on two methodologies: the analysis of gene polymorphism association with human diseases, and systems biology.

The first approach is based on statistical identification of relationships between genomic variations and the incidence of the disease (Psychiatric GWAS Consortium..., 2009). For example, genetic association studies have established a link between the number of polymorphic alleles and the risk of skin cancer (Gerstenblith et al., 2010). Results from a large number of genetic studies are freely available from public research databases (Johnson and O'Donnell, 2009). However, these studies require painstaking work on data collection and verification using methodologies of experimental medicine. This approach also has a number of drawbacks. First, statistical analysis of the SNP cannot distinguish between functionally important polymorphisms and those linked with a trait. Therefore, these studies require additional analyses on the same population before the results can be considered clinically significant. Second, it is hard to assess the roles of rare polymorphisms due to the difficulties in recruiting groups large enough to obtain statistically significant data.

An alternative approach is to employ a methodology for predicting the effects of single nucleotide substitutions in silico, based on the information included in the structural and functional annotations of the human genome and its functional features within the gene and metabolic networks (Weston, 2004). This approach is based on the fact that SNPs may affect human phenotypic traits at different levels of gene regulation: for instance, polymorphisms in the noncoding sequences may damage transcription factor binding sites, affect mRNA splicing, or disrupt other processes important for gene transcription or translation. In turn, polymorphisms in the coding regions can result in amino acid substitutions and lead to changes in the functional or structural properties of the encoded protein. Collectively, such damage on the level of individual genes can affect the functioning of the whole gene network (Computer System Biology..., 2008) and lead to phenotypic defects at the organism level. Contemporary bioinformatics methods enable identification of the damaging mutations both in the noncoding regulatory regions (Savinkova et al., 2009) and at the protein level (Ivanishenko et al., 2011).

Naturally, in silico methodologies cannot provide the same degree of reliability as wide-scale genetic association studies. However, bioinformatics in silico approaches are being intensely developed and used together with GWAS to evaluate the effects of mutations on the occurrence of various pathologies, based on the contemporary knowledge of the affected molecular mechanisms (Na et al., 2013).

To date, a great number of information resources have been developed, such as public databases and computer software, aimed at resolving specific problems on the effects of SNP on particular genomic functions (Mooney et al., 2010). However, the heterogeneity of these resources and the enormous amount and complexity of the information needed to fully assess the risk of pathologies make it difficult for medical and bioinformatics experts to process this information in a manual analysis. An effective solution to this problem requires the development of computer software which can function in an automatic mode to assess the primary risks for socially significant diseases based on individual genetic data. The operation of such a system should be based on the integration of a stream of heterogeneous data produced by several software modules, processing information in a large number of databases. The results of such systemic analysis, obtained based on the known associations between gene polymorphisms and human diseases and bioinformatics predictions, can be used by medical experts for evaluation of an individual patient's characteristics.

In this manuscript, we describe our newly developed modular computer–based information system (MCIS), designed to evaluate the effects of polymorphisms on the incidence of socially significant diseases. The MCIS is structured to link information databases, software algorithms, and mathematical models describing polymorphism effects on the function of numerous genes and networks associated with socially significant diseases, such as neoplasias of various organs and metabolic disorders.

MATHEMATICAL MODELS FOR ANALYSIS OF SNP EFFECTS ON THE INCIDENCE OF SOCIALLY SIGNIFICANT DISEASES

Polymorphisms in the gene coding regions. Models describing SNP effects on the structure and function of proteins, the principal components of the gene networks, are fairly well developed. Amino acid substitutions in proteins can have a significant effect on the function of a particular gene network and can even lead to qualitative changes in network operation. For example, mutations in transcription factors can alter the spectrum of their transcriptional targets (Farnebo et al., 2010).

In general, mutations in the gene coding regions can be subdivided into two major classes based on their effect on the protein product: (1) mutations that disrupt the function of the protein, but maintain its spatial folding; and (2) mutations that disrupt the protein structure (Sanchez-Ruiz et al., 2010).

Impairment in protein activity can be caused by mutations in the sites required for the protein function, such as the catalytically site of an enzyme, a cofactor binding site, or a residue, important for post-translational modifications. Protein activity can be also altered by mutations far from the functional sites if they lead to changes in the spatial structure of the protein active site.

Disruptions in the protein structure can be manifested at the level of protein folding; for instance, they may lead to the formation of proteins fundamentally different in their tertiary structure or entirely lacking a compact structure, or be characterized by reduced thermal stability. In the latter case, mutant proteins may be properly folded, but the lifespan of the protein in this unstable conformation may be shorter then for the native form; such proteins are susceptible to proteolytic degradation at a higher rate.

Large-scale studies of polymorphisms using computer models showed that 90% of single mutations associated with diseases in one way or another reduce the stability of the protein globule. At the same time, based on computer simulations, about 70% of polymorphisms are considered neutral. About 30% of disease-causing mutations are polygenic. Development of novel bioinformatics approaches and data analysis allows us to create information resources for the analysis of the SNP effects on protein activity, and to link the information obtained for the SNPs with annotations of the protein structure and function (Ramensky et al., 2002; Cavallo et al. 2005; Karchin et al., 2005). These bioinformatics technologies empower us to predict the roles of the SNPs in the incidence of human diseases and plan further associated studies for mutations in the human genome (Yue et al., 2006).

Polymorphisms in the noncoding regions of genes. There is considerably less information on the roles of the regulatory SNPs (rSNPs), which can affect the levels of gene expression. The published reports indicate that single nucleotide substitutions in the regulatory regions can disrupt the binding sites of various transcription factors, enable the formation of new sites, or change the transcription factor's affinity to the mutated DNA-binding region. These effects can alter not only the levels of gene expression but can also radically change the gene regulation patterns, including tissue specificity and the response to external signals. With the development of the high performance experimental approaches for identification of the transcription factor binding sites within the whole genome (ChiP-seq). A large volume of data has been accumulated as the result of the works within the international postgenomic project ENCODE (http://genome.ucsc.edu/). These results provide an opportunity to undertake a large scale study to identify regulatory regions in the genomic sequences based on the data on the clustering of the various transcription factors binding sites within particular genomic loci. This opens up the possibility to select SNPs, which localize to these loci, and could, potentially, influence the binding of transcription factors to this region of the DNA. Several studies confirmed the high predictive value of this approach to identify rSNPs. More than 70% of the rSNPs located in these regions are reported to affect DNA–protein binding. Thus, it becomes possible to predict the participation of the regulatory SNPs in the development of various pathologies.

Effect of polymorphisms on the function gene networks. Development of systematic and effective approaches to reconstruct the mechanisms responsible for polymorphism effects on the operation of gene networks is becoming increasingly important. For example, Torkamani and coauthors (2008) investigated deviations in the molecular mechanisms in the gene networks associated with a set of diseases: bipolar disorder, coronary artery disease, Crohn's disease, hypertension, rheumatoid arthritis, and diabetes types 1 and 2. The authors analyzed gene functions within the networks and evaluated the impacts of gene polymorphisms based on the data obtained from genome-wide association studies. The results of this analysis revealed that molecular mechanisms leading to pathogenetic changes are similar in many diseases, and are affected by a variety of general and disease-specific risk factors. Namely, the same signaling pathways in the genetic networks may be responsible for the occurrence of various diseases. These important pathways include signaling by various receptors, and activities of adenylate cyclases, protein kinases, and calcium-dependent signaling.

As another example, there are dozens of known mutations in gene networks which control feeding behavior and are linked to the same physiological effect, an abnormal increase in a person's weight. Remarkably, the action mechanisms of these mutations are based on breaches in the regulatory processes in the functions of the gene networks mentioned above.

To date, the emerging bioinformatics approach, aimed at identifying the molecular and genetic mechanisms responsible for multifactorial diseases contains the following steps (Moore et al., 2010): reconstruction of the gene networks whose defects are linked with particular diseases; identification of the genes and proteins involved in these gene networks; functional assessment of the effects of mutations (polymorphisms) at the level of gene expression, and the structure and function of the respective protein products; and evaluation of the character of the functional disruptions at the level of the local and integral gene networks of an organism.

IMPLEMENTATION OF THE SNP–MED SOFTWARE

The software architecture of the MCIS SNP–MED complex includes three functional modules, structured as software components (Fig. 1) and the Informational Resource Database (IRDB), which contains all the necessary information for the execution of these components.

1. The Genomics program component is designed to assess the effects of the SNPs on the functions of the regulatory regions of the genes. At the heart of this module is a wide-scale search for matches between a patient's polymorphisms, with the known data on

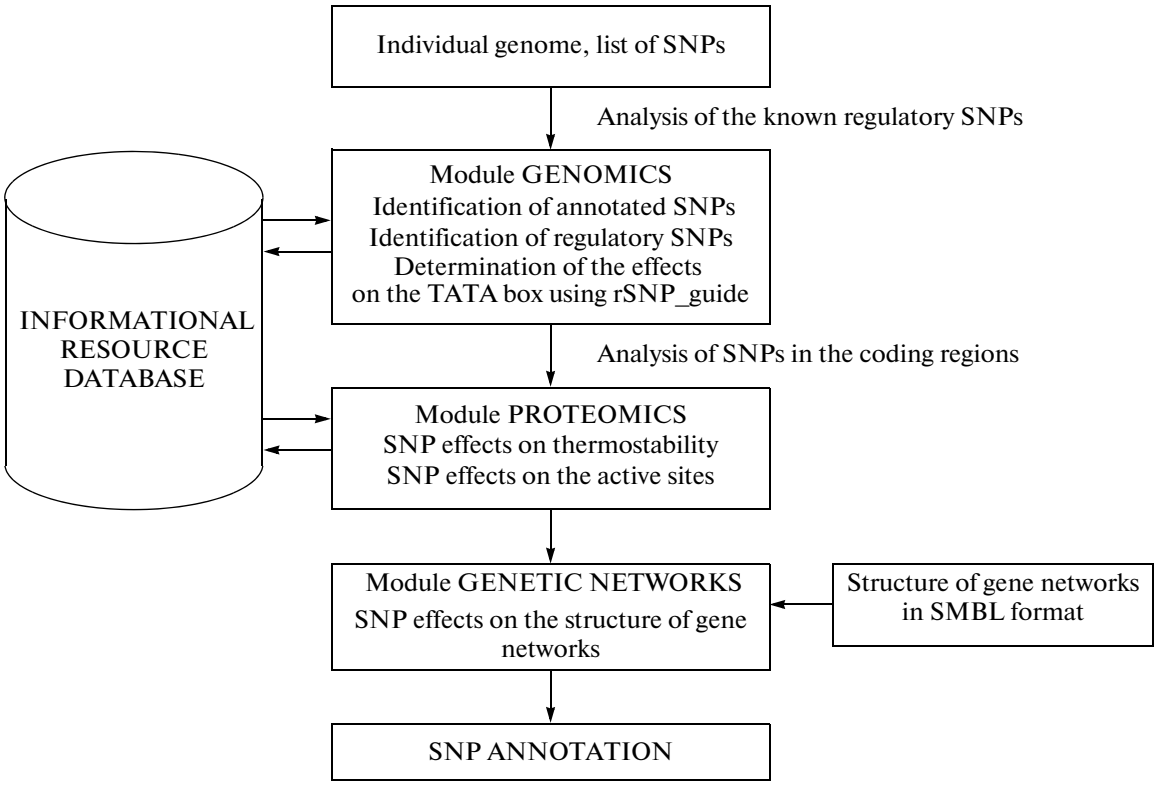

Fig. 1. Basic MCIS SNP–MED modules and their relationship.

polymorphisms and their associations with socially significant diseases, available from public databases; and predictions of the SNP effects in the regulatory regions on gene functions.

2. The Proteomics component is designed to assess the effects of the SNPs in the coding regions of the genes. The basis of this module is a wide-scale analysis of the effects of a patient's polymorphisms on disruptions in the structure and function of proteins encoded by human genes.

3. The Gene Networks' program component will be used to assess the impact of the integral effect of the SNPs on various gene networks. This module will examine the impact of genetic risks factors identified by the Genomics and Proteomics modules at the levels of genes and regulatory interactions, and on the structure and function of gene networks.

The input data, in the form of an individual genome sequence or a list of SNPs, is entered into the system by the end user. These data are first directed for processing by the Genomics module. The module first selects the SNPs for which annotations are already available in the databases and identifies SNPs in the regulatory regions.

Next, the processing of the SNPs follows, localized in the coding regions. Their influence on the protein's thermal stability and integrity of the active centers is determined. At the last step, the identified damaging mutations are projected on the structure of the gene networks, which are uploaded to the system by the end user in one of the standard formats. As result, the Genomics module provides the end user with annotations for the input SNPs; their rating, based on the damage effect; and an assessment of their impact on the function of genes, proteins, and the structure of gene networks.

We constructed the MCIS SNP–MED using the bioinformatics platform UGENE, which is one of the widely used software packages for analysis of various biological data (Okonechnikov, 2012). UGENE is popular due to its wide range of applications, cross-platform properties, and open-source use, as UGENE is distributed under the GNU General Public License, v. 2.0.

Annotations of the single-nucleotide polymorphisms within UGENE are carried out by a digital scheme designer, the Workflow Designer (WD). This module creates multistep conveyors of computational schemes for processing biological information. Each stage in the pipeline represents a separate algorithm. During the operation, these algorithms exchange messages which contain the input and output data for the subsequent circuits.

The WD internal architecture allows us to combine various components of the MCIS within the general interface environment by adding processing elements corresponding to the individual algorithms. This approach will enable the use of these components in any future WD applications.

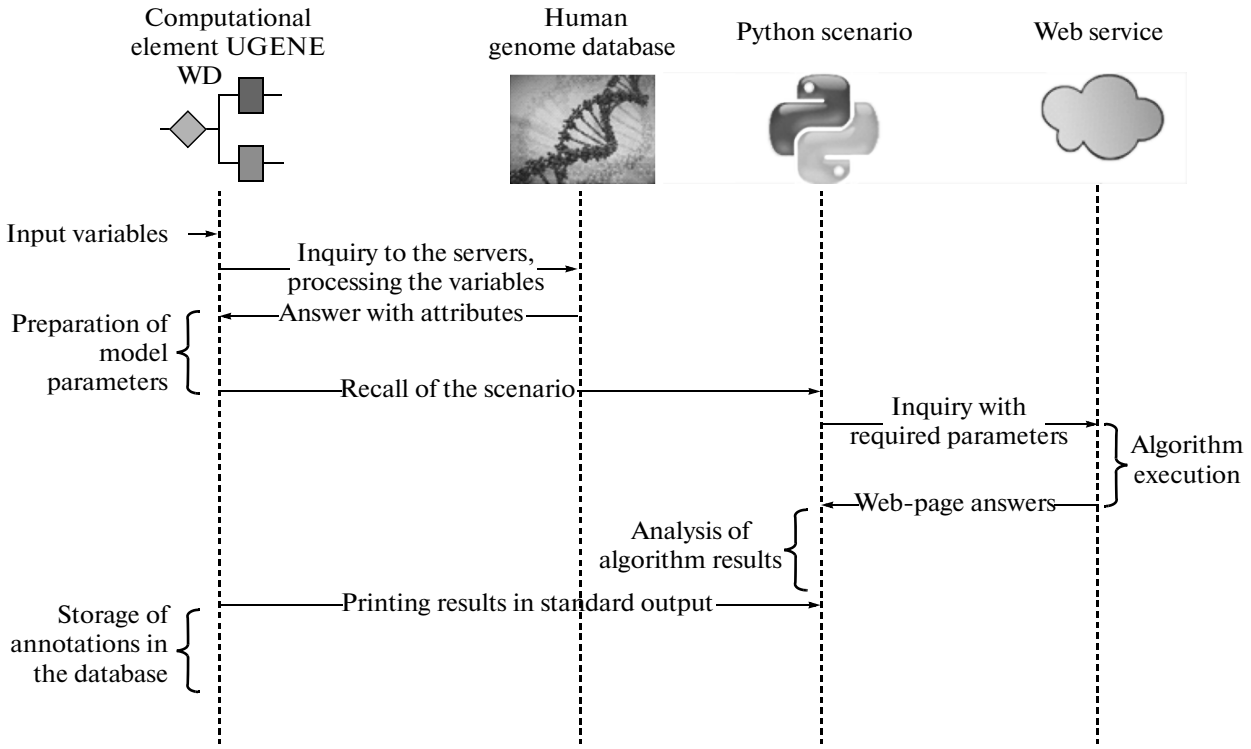

Fig. 2. Algorithm of interactions between the computational elements MCIS SNP_MED and IRBD during the process of data-mining to obtain SNP annotations using the Workflow Designer UNIPRO.

We will further describe the operations of the individual MICIS SNP–MED modules. **The Genomics program component** comprises the following modules.

1. Software module for finding SNPs associated with socially significant diseases represented in public databases. These databases include Diseases, dbSNP (Sherry et al., 2001), Exome variant server, 1000 Genomes, and HapMap. This module also uses the DNA sequence of the human genome and its functional annotations, including the information for localization of genes and regulatory regions, sequence duplications, and evolutionary conserved regions (http://genome.ucsc.Edu/). This module uses the data on the SNP locations to retrieve the available annotations from the databases described above and directs them to the module's output.

2. The software module for estimating the probability of finding the SNPs in the regulatory regions of the genes. This module retrieves the information from the annotations of the genomic binding sites for dozens of transcription factors obtained in the course of the ENCODE project (Rosenbloom et al., 2013). Polymorphisms in these regions can potentially affect the binding of transcription factors to these DNA sites. Based on experimental studies, more than 70% of nucleotide substitutions in the regulatory regions can affect the DNA–protein binding.

3. Software module for assessment of the SNP impact on the function of the TATA-box sites. This analysis is based on the model for the TBP protein (TATA box Binding Protein) binding to a DNA fragment containing the TATA box. The model is described in four successive steps reflecting the critically important stages in the TATA box function (Ponomarenko et al., 2008). This model allows us to accurately estimate the protein binding's affinity to the target sequence, which was confirmed experimentally.

4. Software module for evaluation of the SNPs' influence on the function of the regulatory regions based on the rSNP_Guide methodology. This approach allows us to identify the type of transcription factors, whose binding is impaired most severely by the mutations in the DNA regulatory region.

Software component Proteomics comprises the following modules:

1. Program module for predicting the impact of SNPs on protein thermodynamic stability.

2. Program module to identify SNPs at protein functional sites.

Informational Resource Database (IRDB) includes the information required for all MCIS processes for the analysis of SNP effects on the function of genes associated with the incidence of socially significant diseases. IRDB contains the complete annotation of the human genome and the known SNP annotations which are used by the Genomics module. IRDB continuously accumulates new SNP annotations,

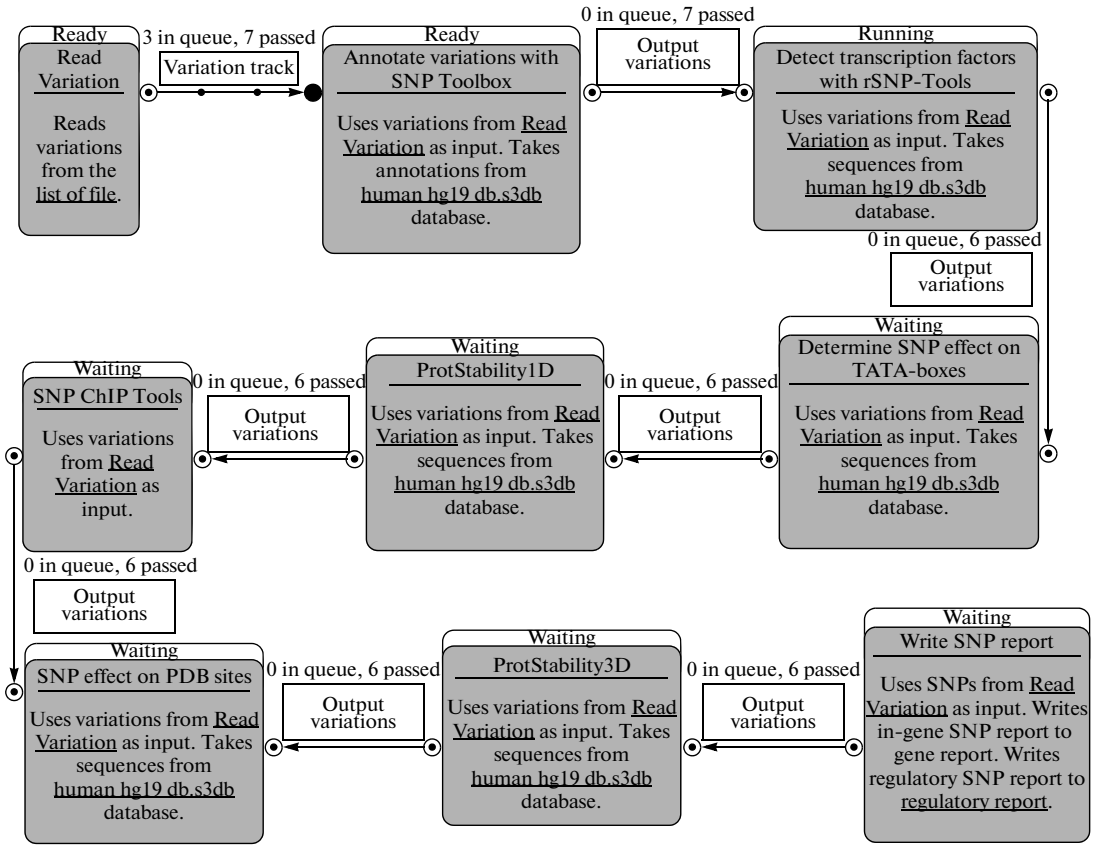

Fig. 3. Detailed algorithm sequence during the operation of the conveyor MCIS SNP_MED, composed within for the UGENE Worflow Designer software environment.

obtained through several freely accessible programs listed below. (1) Results of the analysis of the protein-coding sequences using the algorithm SIFT (Ng and Henikoff, 2003), which evaluates the effects of mutations on protein function. (2) Results of analysis of the protein-coding sequences using the algorithm PolyPen (Adzhubei et al., 2010), which allows us to identify damaging mutations in amino acid sequences. (3) Results of the analysis of the protein-coding sequences using algorithms PhyloP, LRT, and GERP (Pollard et al., 2010), which evaluate the extent of the harmful SNP effects based on the analysis of the evolutionary conserved genomic sequences.

IRDB does not require a separate calculation for each SNP, requested by the end user. The program automatically searches for the available information, which assures fast data processing.

During our design of the software for the computing modules, the majority of the algorithms required for SNP annotations were developed in a form of separate web-based applications. Only the analysis of the polymorphism effects on the TATA box function is performed on the local server. Therefore, the assembly of additional computational algorithms in the WD was carried out by incorporating requests for the remote services and direct links to the local processing modules.

From the perspective of the software environment, all algorithms within the system share a common interface: transmission of the attributes of genomic sequences containing nucleotide substitutions and SNP annotations. Figure 2 shows the general scheme of interactions between the system modules and computational elements, assembled in this software package. The query and response steps are not indicated in the diagram. Activation of the appropriate algorithms is performed by the computational elements.

One of the important attributes in each of the computing elements (except the SNP ChIP algorithm) is the local link to IRDB containing the annotations of the human genome. Appeals to IRDB allow us to obtain the parameters for genomic sequences containing the polymorphism required to recall any of the computing algorithms. Using this information, the computing elements execute the appropriate scenario within the Phyton software environment. The Phyton module, in turn, produces an inquiry to a remote server, which implements the requested algorithm. After completion of the remote computing, the scenario receives the processing results scripted as SNP annotations in HTML format, which are used for further analysis. Subsequently, polymorphism annota-

#Chr	Position	Allele	dbSNP	Gene	Clinical_significance	Location	Protein	Codon	Substitution
chr11	94800448	A/G	-	B2R6B8	CDS.	Exon: 94800056..94800900	Q9BRL6	AAC->GAC	N20D
chr13	113634072	G/A	-	AB002360	Factor VII CDS.	Intron: 113623029..113669076	-	-	-
chr13	113634072	G/A	-	CCDS45070	Factor VII CDS.	Intron: 113623029..113669076	-	-	-
chr13	113634072	G/A	-	KIAA0362	Factor VII CDS.	Intron: 113623773..113669076	-	-	-
chr13	113634072	G/A	-	CCDS9527	Factor VII CDS.	Exon: 113633982..113634072. Donor splice-site.	E9PDN9	GAT->AAT	D31N
chr14	22356190	T/A	-	AJ004871	CDS.	Intron: 22192546..22447340	-	-	-
chr14	22356190	T/A	-	FR159098	CDS.	Exon: 22205173..23016490	-	-	-
chr14	22356190	T/A	-	TCRA	Cysticercosis(2.2), Taeniasis(2.2), Leukemia(1.4), Disease(1.2), Toxocariasis(1.2), Echinococcosis(1.1), Cancer(1.0), Lymphoma(0.8), Kyoma(0.8), Pneumothorax(0.7), Arthritis(0.5)	Intron: 22294244..22994626	-	-	-
chr14	22356190	T/A	-	FR004500	CDS.	Exon: 22321073..22981890	-	-	-
chr14	22356190	T/A	-	M97714	CDS.	Intron: 22337544..22733720	-	-	-
chr14	22356190	T/A	-	AV2S1A1	CDS.	Exon: 22356037..22356190. Donor splice-site.	-	TGG->AGG	W16R
chr3	126386191	G/C	-	C3orf46	CDS.	Exon: 126385949..126386191. Donor splice-site.	Q8IVU5	AAG->AAC	K133N
chr4	265207	C/A	-	B4DXR9	CDS.	Exon: 264464..266419	B4DXR9	GGC->GTC	G448V
chr4	265207	C/A	-	B4DXR9	CDS.	Exon: 264464..266419	B4DXR9	GGC->GTC	G480V
chr4	152571673	C/G	-	FAM160A1CDS.	CDS.	Exon: 152570611..152571744	Q05DH4	CCA->CGA	P827R
chr5	33997584	G/C	-	AMACR	Cancer(2.0), Adenocarcinoma(2.0), Carcinoma(2.0), Disease(1.8), Adenoma(1.0), Prostatitis(1.0), Angiomyolipoma(0.8), Adrenoleukodystrophy(0.6)	Intron: 33989608..33998745	-	-	-
chr5	33997584	G/C	-	A5YM47	prostate cancer; effects in AMACR are the cause of alpha-methylacyl-CoAracemase deficiency (AMACRD) effects in AMACR are the cause of congenital bile acidsynthesis defect type 4 (CBAS4) CDS.	Intron: 33989608..33998745	-	-	-

Fig. 4. Sample output report for known SNPs associated with diseases in SNP-report format.
Top row contains the column names of the report. Below, the annotations for polymorphisms obtained during the work MCIS SNP–MED project are given.

tions are directed towards the standard scenario output, read by a computing element.

At the last stage of the algorithm, the newly obtained information is submitted to the shared computational database. Thus, during the sequential movement through the various stages of the conveyor, each SNP annotation is gradually complemented with new database attributes. At the end, the content of the SNP annotation is processed by the final element of the computation scheme Write_SNP_Report (Fig. 3), which creates a two part report on the impact of the gene polymorphism and the regulatory region, respectively (Figs. 4, 5).

These final data are issued in the form of an SNP_report. This is a text file containing polymorphism annotation in one or two rows. The data are arranged in columns separated by tab characters. The number of columns in each report reflects the number of algorithms that contributed to the pipeline process. A full version of the report for each nucleotide substitution (Fig. 3) will contain identifiers for the respective genomic sequence and the following parameters: identity and location of the damaged gene; identity and mutated codon of the damaged protein; list of socially significant diseases linked to the damaged genomic location; assessment of the extent of damage using the SIFT algorithm; evaluation of the polymorphism impact on the protein thermodynamic stability and lifespan; and identification of protein conformers, with variations on the active site. A sample report is shown in Fig. 4.

The final report describing SNP effects in the regulatory region will include the following information: identifier for the affected gene and promoter sequence; assessment of the damage to the regulatory region; list of socially significant diseases associated with the damaged genomic area; list of damaged transcription factor binding sites; and an assessment of the impact of the mutation on the functional activity of the TATA boxes. An example of the annotation report for a regulatory SNP is shown in Fig. 5.

We developed the following bioinformatics scenarios within the MCIS SNP–MED for analysis of the SNP impact on the function of genes associated with the incidence of socially significant diseases.

1. Bioinformatics scenario for data analysis at the genomic level to identify SNPs associated with socially significant diseases.

2. Scenario for assessment of the SNP impact in the regulatory regions of the genes. The end user is

1	#Chr	Position	Allele	dbSNP	Promoter_of_Gene	Clinical_significance	From_transcription_start	rSNPTools_factors	ChIPTools
2	chr17	39993435	T/C	-	C9JKC4	-911	-	-	-
3	chr17	39993435	T/C	-	5NT3L_HUMAN	-911	-	-	-
4	chr17	39993435	T/C	-	KLH10_HUMAN	-608	-	-	-
5	chr17	39993435	T/C	-	AK302141	-642	-	-	-
6	chr17	39993435	T/C	-	AK301797	-642	-	-	-
7	chr17	39993435	T/C	-	5NT3L_HUMAN	-946	-	-	-
8	chr17	39993435	T/C	-	C9JKC4	-911	-	-	-
9	chr17	39993435	T/C	-	5NT3L_HUMAN	-911	-	-	-
10	chr17	39993435	T/C	-	KLH10_HUMAN	-608	-	-	-
11	chr17	39993435	T/C	-	AK302141	-642	-	-	-
12	chr17	39993435	T/C	-	AK301797	-642	-	-	-
13	chr17	39993435	T/C	-	5NT3L_HUMAN	-946	-	-	-
14	chr17	39993435	T/C	-	C9JKC4	-911	-	-	-
15	chr17	39993435	T/C	-	5NT3L_HUMAN	-911	-	-	-
16	chr17	39993435	T/C	-	KLH10_HUMAN	-608	-	-	-
17	chr17	39993435	T/C	-	AK302141	-642	-	-	-
18	chr17	39993435	T/C	-	AK301797	-642	-	-	-
19	chr17	39993435	T/C	-	5NT3L_HUMAN	-946	-	-	-
20	chr17	39993435	T/C	-	C9JKC4	-911	-	-	-
21	chr17	39993435	T/C	-	5NT3L_HUMAN	-911	-	-	-
22	chr6	32635046	A/G	rs76356512	HLA-DQB1	-579	-	-	-
23						Disease(2.0),			
24						Lymphopenia(1.3),			
25						Scleroderma(1.1),			
26						Leukopenia(1.1),			
27						Podoconiosis(1.0),			
28						Sarcoidosis(1.0),			
29						Narcolepsy(0.8),			
30						Dermatitis(0.7),			
31						Schizophrenia(0.7),			
32						Elephantiasis(0.7),			
33						Thyroiditis(0.7),			
34						Glomerulonephritis(0.7),			
35						Hyperthyroidism(0.6),			
36						Arthritis(0.6),			
37						Nephritis(0.6),			
38						Cancer(0.6),			
39						Agammaglobulinemia(0.6)			
40	chr6	32635046	A/G	rs76356512	A2AAZ0	-579	-	-	-
41						Leprosy; diabetes,			
42						type 1; pancreatitis,			
43						autoimmune;			
44						pancreatitis, chronic			
45						calcifying;			
46						periodontitis;			
47						infertility, tubal			

Fig. 5. Sample output report on the influence of polymorphisms on the gene regulatory regions in the SNP-report format. The top row contains the column names of the report. Below, the annotations for polymorphisms obtained during the MCIS SNP–MED project are given.

provided with a list of SNPs located in the regulatory regions and associated with socially significant diseases, and an assessment of their functional significance.

3. Scenario for determination of SNP effects on the functional activity of the TATA boxes in the regulatory gene regions. The client is provided with a list of genes in which the activity of the TATA boxes is found to be particularly vulnerable to the specific SNPs and the description of the SNP effects.

4. Scenario for assessment of the SNP impact on the functional activity of the transcription factor binding the sites in the regulatory regions of the genes. The client is provided with a list of genes which poses binding sites for the particular transcription factors whose functional activity is affected by the specific SNPs, and a description of SNP's effects.

5. Typical scenario for bioinformatics analysis at the proteomic level, including prediction of SNP's effects on protein thermodynamic stability and identification of SNPs at the functionally important protein sites.

6. Typical scenario for bioinformatics analysis of SNP effects on gene networks.

CONCLUSIONS

We developed a modular computer information system, SNP–MED, to analyze the impact of SNPs on the function of genes associated with the incidence of socially significant diseases. This information system comprises the following components:

1. Program component Genomics, which?includes software modules and service interfaces for searches for SNPs associated with socially significant diseases; assessment of the probability of finding SNPs in the regulatory gene regions; and evaluation of SNP's effect on the function of the TATA boxes and regulatory regions.

2. Program component Proteomics, which includes software modules and service interfaces for identification of SNPs in the coding regions of the genes associated with socially significant diseases; prediction of SNP's impact on protein thermodynamic stability; and identification of SNPs at the functional sites of proteins.

3. Program component Gene Network, which enables evaluation of SNP's effects on the function of gene networks.

4. The Information Resource Database, which accumulates data required for the MCIS SNP–MED function for analysis of SNPs' impact on the function of genes associated with the incidence of socially significant diseases.

ACKNOWLEDGMENTS

This work was supported by the Ministry of Education and Science of the Russian Federation, contract number 14.512.11.0094.

REFERENCES

Adzhubei, I.A., Schmidt, S., Peshkin, L., et al., A method and server for predicting damaging missense mutations, *Nature Meth.*, 2010, vol. 7, no. 4, pp. 248–249.

Cavallo, A. and Martin, A.C., Mapping SNPs to protein sequence and structure data, *Bioinformatics*, 2005, vol. 21, pp. 1443–1450.

Farnebo, M., Bykov, V.J., and Wiman, K.G., The p53 tumor suppressor: a master regulator of diverse cellular processes and therapeutic target in cancer, *Biochem. Biophys Res. Commun.*, 2010, pp. 85–89.

Gerstenblith, M.R., Shi, J., and Landi, M.T., Genome-wide association studies of pigmentation and skin cancer: a review and meta-analysis, *Pigment Cell Melanoma Res.*, 2010, vol. 23, no. 5, pp. 587–606.

Ivanisenko, V.A., Demenkov, P.S., Ivanisenko, T.V., and Kolchanov, N.A., Protein structure discovery: a software package to computer proteomics tasks (review), *Russ. J. Bioorg. Chem.*, 2011, vol. 37, no. 1, pp. 17–29.

Johnson A.D. and Donnell, C.J., An open access database of genome-wide association results, *BMC Med. Genet.,* 2009, vol. 10, no. 1, p. 6.

Karchin, R., Diekhans, M., Kelly, L., et al., LS-SNP: large-scale annotation of coding non-synonymous SNPs based on multiple information sources, *Bioinformatics*, 2005, vol. 21, pp. 2814–2820.

Mooney, S.D., Krishnan, V.G., and Evani, U.S., Bioinformatic tools for identifying disease gene and SNP candidates, *In Genetic Variation*, 2010, pp. 307–319.

Moore, J.H., Asselbergs, F.W., and Williams, S.M., Bioinformatics challenges for genome-wide association studies, *Bioinformatics,* 2010, vol. 26, pp. 445–455.

Na, Y.J., Cho, Y., and Kim, J.H., AnsNGS: an annotation system to sequence variations of next generation sequencing data for disease-related phenotypes, *Healthcare Inform. Res.*, 2013, vol. 19, no. 1, pp. 50–55.

Ng, P.C. and Henikoff, S., Sift: predicting amino acid changes that affect protein function, *Nucleic Acids Res.*, 2003, vol. 31, pp. 3812–3814.

Okonechnikov, K., Golosova, O., Fursov, M., et al., Unipro UGENE: a unified bioinformatics toolkit, *Bioinformatics*, 2012, vol. 28, pp. 1166–1167.

Pollard, K.S., Hubisz, M.J., Rosenbloom, K.R., and Siepel, A., Detection of nonneutral substitution rates on mammalian phylogenies, *Genome Res.,* 2010, vol. 20, no. 1, pp. 110–121.

Ponomarenko, P.M., Savinkova, L.K., Drachkova, I.A., et al., A step-by-step model of TBP/TATA box binding allows predicting human hereditary diseases by single nucleotide polymorphism, *Dokl. Biochem. Biophys.*, 2008, vol. 419, pp. 88–92.

Psychiatric GWAS Consortium Steering Committee. A Frame work for interpreting genome-wide association studies of psychiatric disorders, *Mol. Psychiatry,* 2009, vol. 14, no. 1, p. 10.

Ramensky, V., Bork, P., and Sunyaev, S., Human non-synonymous SNPs: server and survey, *Nucleic Acids Res.*, 2002, vol. 30, pp. 3894–3900.

Rosenbloom, K.R., Sloan, C.A., Malladi, V.S., et al., Encode data in the UCSC genome browser: year 5 update, *Nucleic Acids Res.*, 2013, pp. D56–D63.

Sanchez-Ruiz, J.M., Protein kinetic stability, *Biophys. Chem.*, 2010, vol. 148, pp. 1–15.

Savinkova, L.K., Ponomarenko, M.P., Ponomarenko, P.M., et al., TATA box polymorphisms in human gene promoters and associated hereditary pathologies, *Biochemistry* (Moscow), 2009, vol. 4, no. 2, pp. 117–129.

Sherry, S.T., Ward, M.H., Kholodov, M., et al., dbSNP: the NCBI database of genetic variation, *Nucleic Acids Res.*, 2001, no. 29, pp. 308–311.

Sistemnaya komp'yuternaya biologiya (Computer System Biology), Kolchanov, N.A., Goncharov, S.S., Likhoshv, V.A., and Ivanisenko, V.A., Eds., Novosibirsk: SO RAN, 2008.

Torkamani, A., Topol, E.J., Schorkn, J., Pathway analysis of seven common diseases assessed by genome-wide association, *Genomics*, 2008, no. 92, pp. 265–272.

Weston, A.D., L h. systems biology, proteomics, and the future of healthcare: toward predictive, preventative, and personalized medicine, *J. Proteome Res.*, 2004, vol. 3, no. 2, pp. 179–196.

Yue, P., Melamud, E., and Moult, J., SNPs3D: candidate gene and SNP selection for association studies, *BMC Bioinformatics*, 2006, no. 7, p. 166.

Translated by I. Grishina