BMC Bioinformatics

Methodology article

Target SNP selection in complex disease association studies

Matthias Wjst*

Address: Gruppe Molekulare Epidemiologie, Institut für Epidemiologie, GSF – Forschungszentrum für Umwelt und Gesundheit, Ingolstädter Landstrasse 1, D-85758 Neuherberg/Munich, Germany

Email: Matthias Wjst* - wjst@gsf.de

* Corresponding author

Published: 12 July 2004

BMC Bioinformatics 2004, **5**:92 doi:10.1186/1471-2105-5-92

Received: 19 February 2004

Accepted: 12 July 2004

This article is available from: http://www.biomedcentral.com/1471-2105/5/92

Abstract

Background: The massive amount of SNP data stored at public internet sites provides unprecedented access to human genetic variation. Selecting target SNP for disease-gene association studies is currently done more or less randomly as decision rules for the selection of functional relevant SNPs are not available.

Results: We implemented a computational pipeline that retrieves the genomic sequence of target genes, collects information about sequence variation and selects functional motifs containing SNPs. Motifs being considered are gene promoter, exon-intron structure, AU-rich mRNA elements, transcription factor binding motifs, cryptic and enhancer splice sites together with expression in target tissue.

As a case study, 396 genes on chromosome 6p21 in the extended HLA region were selected that contributed nearly 20,000 SNPs. By computer annotation ~2,500 SNPs in functional motifs could be identified. Most of these SNPs are disrupting transcription factor binding sites but only those introducing new sites had a significant depressing effect on SNP allele frequency. Other decision rules concern position within motifs, the validity of SNP database entries, the unique occurrence in the genome and conserved sequence context in other mammalian genomes.

Conclusion: Only 10% of all gene-based SNPs have sequence-predicted functional relevance making them a primary target for genotyping in association studies.

Background

The massive amount of single nucleotide polymorphism (SNP) data stored at public internet sites provides unprecedented access to human genetic variation. SNPs are thought to be the genetic basis of most human diseases, or at least positional markers for our genetic heritage. In contrast to monogenic diseases, complex diseases require the simultaneous testing of hundreds of genes with thousands of SNPs. Even worse, SNP databases contain redundant, incomplete and even wrong information. Worse still, data are constantly changing during a large genotyping project.

Although public databases offer valuable information about genomic context, submitter, or allele frequency, the annotation is often outdated, incomplete or fragmented among different websites. As a result the US National Human Genome Research Institute is planning to identify all functional elements in the human genome sequence in the ENCODE (Encyclopedia of DNA Elements) project.

Single nucleotide substitutions may influence complex diseases by a variety of mechanisms. Mutations may affect the amino acid sequence of predicted proteins where functions like DNA binding, catalytic activity and receptor

– ligand contact are reduced or abolished. SNPs may interrupt the initiation, the termination codon or introduce errors in the reading frameshift, all with consequences for insufficient or prematurely truncated peptides. SNPs can also have invariable effects on transcription, RNA processing, stability and translation. Mutations in known promoter motifs usually lead to reduced mRNA levels. mRNA splicing mutants are most commonly found at the beginning and end of the donor and acceptor consensus splice sequence and cause either exon skipping or utilization of cryptic splice sites resulting in the absence of normally spliced mRNA. Finally, RNA cleavage-polyadenylation mutants can occur in the AAUAAA sequence upstream of the polyadenylation sites. Other mutations in the untranslated region of the 5' UTR are thought to play a role in controlling mRNA translation while sequence variants in the 3' UTR control RNA cleavage, stability, export and intracellular localization [1].

Many of these biological effects may be predicted by sequence context analysis although identifying repeats, CpG islands, promoter structure, transcription factor binding sites, cryptic and enhancer splice sites, AU-rich elements, different transcripts or inter-species conserved sequences is a time-consuming process. Even web services dedicated to SNP annotation like *SNPper* by Riva [2], *GeneSNPs* implemented by Weiss and Dunn [3], *PicSNP* implemented by Chang and Fujita [4] and *ParseSNP* by Taylor and Green [5] usually implement only a subset of these features. Clearly, it is also important to build a local SNP database, since local data storage is required for documentation.

To facilitate our own SNP genotyping triage process we built an extension of our previously developed laboratory information system for the MALDI-TOF genotyping platform.

Results and Discussion

We implemented a computational pipeline that scanned the genomic sequence of a target gene, retrieved available information about sequence variation, merged it with local sequences, and selected the most important SNPs. All information is displayed either in list form or in a graphical output (Fig. 1) before being manually edited.

The annotation time depended heavily on the sequence length and was usually not critical up to 50 kB of genomic sequence and up to 250 features found. Many regular expression searches for complicated motifs slowed down the system although finding features in general was not time-consuming. Formatting of the output took about 30 seconds. As longer genes needed up to 60 min, we built a caching mechanism where all annotation is computed and saved to disk for later use. All procedures worked well for more than 95% of the target genes. Problems in the remaining genes arose due to nomenclature discrepancy, wrong transcript boundaries or mismatching sequences between different data sources.

As a case study, we selected 396 genes on chromosome 6p21 contributing 19,495 SNPs. By computer annotation 2,562 functional SNPs could be identified (table 1 and supplementary data cooke.gsf.de/wjst/paper/2004BMCBioinformaticsSuppl that shows all annotated genes). An inspection of the 598 different genomic motifs revealed that 88% were in transcription factor binding sites. Some of these were even redundantly labelled in TRANSFAC, where different transcription factors affect the same target sequences. 1,335 SNPs were found within just one motif, while 630 were found within 2 motifs and 541 in more than 2 motifs. The absolute percentage of conspicuous SNPs in non-redundant genomic motifs therefore was less than 10%.

As a preliminary validation, we compared the SNPs inside of functional motifs with those situated outside (Figure 2). Allele frequency data originated from an earlier study [6] where we used MALDI-TOF for screening of SNP allele frequencies in pooled DNA. From this extended set of SNP allele frequencies, 1,633 SNPs in the Caucasian population (random SNP subset of the 19,495 SNPs annotated) could be assessed by type of functional change. SNP allele frequencies were significantly lower (P = 0.004) if they inserted a new transcription factor compared to SNPs with no functional change. No such effect was seen if a transcription factor binding site was being destroyed.

The reduced allele frequency in these SNPs may be interpreted as a lower adaptive fitness in the population. Spontaneous germline mutations are introduced during DNA replication or recombination and conventional theory says that they are governed by genetic drift and modified by natural selection. Most of these SNPs will be selectively neutral and increase steadily in a fixed population with age [7]. Beneficial variants supporting adaptive changes may even exceed allele frequency while disadvantageous variants may be wiped out or remain at a low steady state. A detailed analysis of single transcription factors that are responsible for this effect, as well as simulation studies of population genetics, might further substantiate this observation.

$$\begin{matrix}A\\C\end{matrix}AG|GT\begin{matrix}A\\G\end{matrix}AGT------YYYYYYYYYYN\begin{matrix}C\\T\end{matrix}AG|G$$

Exon-intron boundaries are being highly conserved. This might be due to the rather short stretch of the acceptor and donor sites,

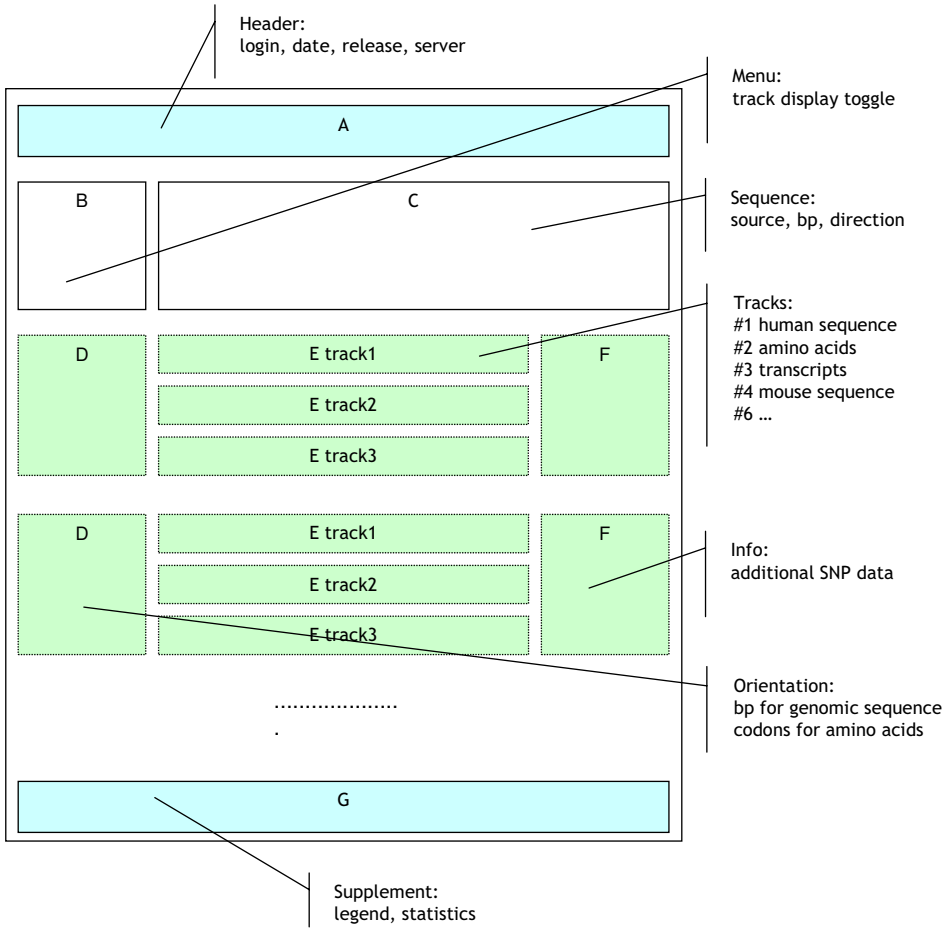

Figure 1
The SNP context view consists of 7 panes with a flexible number of sequence lanes in pane "E". The sequence pane may be exploded by the interlinear display of splice variants or conserved sequences among species.

but also reveals a functionally extremely important region. The Human Gene Mutation Database reported by October 2002 [1] 3,540 deleterious SNPs in exon-intron junctions (10% of all reported SNPs), in comparison to 336 (1%) in regulatory regions, 4,346 (17%) as nonsense and 17,053 (67%) as missense mutations. These figures were obtained from monogenic diseases and it is far from clear that these proportions will also be true for complex diseases.

Amino acid sequence variation was also a possible source for disease-causing SNPs. From table 1 it may be concluded that single base exchanges are seen as often as with amino acid exchanges. Although many SNPs introduce only silent mutations, they affect splice enhancer motifs where an experimental follow of different mRNA copy numbers in cells from these individuals could be important. The coverage of many of these motifs, however, is not exhaustive at the moment. Other locus control regions are also expected to be influenced by SNPs, while only a limited number of these genomic features have been identified so far.

The automatic selection of SNPs turned out to be a good starting point for planning a study, but due to the enormous genotyping costs produced by a wrong selection, SNPs need rigorous further manual assessment. As a consequence we developed a set of genotyping triage criteria that are applied before genotyping (table 2). The most critical question was whether a particular SNP really exists in the target population. Hence we always looked for the submitter of the variant (see the online

Table 1: SNP annotation of human chromosome 6p21

total genes examined	396
Total SNP number (mean 49/gene, range 1–463)	19,495
SNPs inside of a motif	2,562
total motifs touched	5,547
SNPs inside of transcription factor site	4,893
SNPs that deleted a binding site	3,053
SNPs that inserted a binding site	1,816
SNPs inside of exon-intron boundary of all known transcripts	44
SNPs leading to amino acid exchange	310
Ala>Thr	15
Glu>Lys	11
Arg>Gln	10
SNPs inside of exon splice enhancer motif	299
3C	80
5C3D	73
5B3A	39

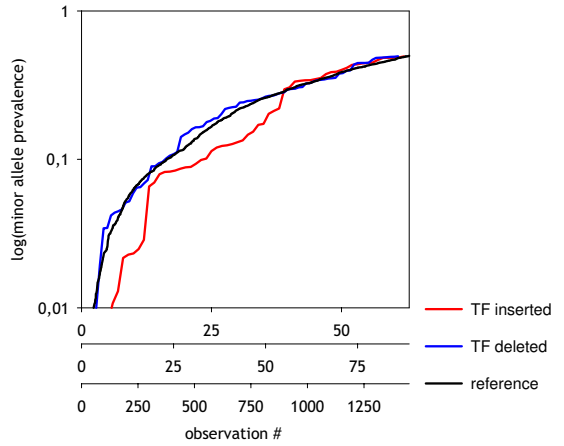

Figure 2
Allele frequencies of 1,633 SNPs in the Caucasian population (which is a random SNP subset of the 19,495 annotated SNPs on chromosome 6p21) by type of functional change. SNP allele frequencies were on average 4% lower in the two bottom tertiles (P = 0.004) of the frequency distribution in those SNPs that insert a new transcription factor binding site (red line) compared to SNPs that destroyed a binding site or were not found in any motif (black line).

supplement of [8] for less reliable submitters) and whether the genomic context suggested sequencing errors during SNP discovery (stuttering, repeats, etc.). Those SNPs with allele frequency available in the target population were preferred as well as those SNPs seen in private databases like the ALLSNPS, REALSNP or in the CELERA Discovery System. SNPs in cross-species conserved regions were also preferred as it could be shown that these are more likely to be of functional importance [9]. Tag information where SNPs are identified on unique haplotypes is only partially available at the moment but may be helpful in some instances.

As a final test we blasted all primer sequences against the whole genome [10] to exclude those with multiple hits in the human genome that would introduce errors by amplification of DNA from distant genomic regions. By applying this criteria, successful genotyping rates of database-obtained SNPs can be increased from about 65% to about 85%.

Conclusion

In summary, we describe an application framework for SNP annotation that allows a rapid decision on whether a SNP should be genotyped in a first instance. The most striking observation so far – except the notion of many database discrepancies – is the fact that only 10% of all SNPs are situated in putative functionally or structurally important regions. Further developments will focus on performance optimization, the implementation of additional annotation features, benchmarking by using published disease-gene associations as well as the experimental follow of SNP-introduced transcription factor binding sites.

Methods

The application was implemented on a 4 CPU, WIN 2000 system, using an MS SQL 2000 database, Cold Fusion 5.0 scripting engine, Apache 2.0 web server and Mozilla Firefox 0.8 as front-end [11]. Scripting was done in standard HTML and CFML (the native Cold Fusion syntax). The system can also be ported to other operating systems where Cold Fusion and Apache binaries are available (Linux, Solaris and others). Statistical procedures have been implemented with R software 1.8.1.

Basically, four scripts were developed: a "grabber" script that connects to external databases or websites, a "compiler" script that compiles all information and preselects SNPs. A "viewer" script displays all data in a coherent view while an "exporter" script transfers all results to the local laboratory information system.

Candidate genes may be initially selected from linkage regions by disease-specific databases (http://geneticassociationdb.nih.gov, http://archive.uwcm.ac.uk/uwcm/mg/hgmd0.html, http://www.centralmutations.org), from human (OMIM http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=OMIM) and animal studies (OMIA http://www.angis.org.au/Databases/BIRX/omia), by disease gene prediction http://maine.ebi.ac.uk:8000/services/

Table 2: Ten questions that help during the SNP genotyping triage process

[1]	Is the SNP density 1/800 bp or higher?	The gene of interest may need to be resequenced.
[2]	Is the SNP located inside of a DNA repeat?	This SNP could be a sequencing artefact.
[3]	Has a particular SNP been submitted by at least two reliable sources?	The genomic sequence may not be polymorphic.
[4]	Is the SNP seen jointly in different public (dbSNP, HGBASE) and private (ALLSNPS, REALSNP, CELERA) databases?	The SNP may not exist.
[5]	Are allele frequencies available in the target population?	The SNP may not exist in the target population.
[6]	Is the SNP located in a known functional motif?	This will increase the likelihood of a causative mutation.
[7]	Is the SNP situated in a region that is conserved in other species?	Highly conserved regions have an increased likelihood of being functionally important.
[8]	Are there more SNPs in the neighborhood?	These SNPs may interfere with primer design.
[9]	Is the SNP seen also in paralogous genomic regions ("multiple genome hitters">)?	These SNPs could interfere at genotyping by creating artificial alleles.
[10]	Does the SNP tag any specific haplotype?	Tagging SNP can help to capture effects of neighboring SNPs in linkage disequilibrium.

dgp, gene ontologies (http://www.bork.embl-heidelberg.de/g2d/[12]) and from patent databases http://www.uspto.gov. Starting with a loose description of the gene name and searching it against the official HUGO nomenclature database http://www.gene.ucl.ac.uk/nomenclature/ the correct gene name is then used for a query against the Weizman Genecards server (http://genecards.weizmann.ac.il[13]) for further gene descriptions. The SNPper database http://snpper.chip.org and dbSNP http://www.ncbi.nlm.nih.gov/SNP/ provide all SNP information. Other SNP databases queried are ALLSNP http://www.allsnps.com, REALSNP https://www.realsnp.com/default.asp and the CELERA Discovery System https://cds.celera.com/cds/login.cfm. Curated transcript sequences are available from DoTS (Database Of Transcribed Sequences, http://allgenes.org).

FASTA sequences of human, mouse, ape and fugu were retrieved from the UCSC Golden Path http://genome.ucsc.edu. The "grabber" script sends http requests and parses the resulting HTML tables, XML stream or ASCII (FASTA) text. The order in which external databases are being queried is important as some queries rely on the existence of already executed queries. All sequences are treated as binary large objects (BLOBs) and saved to disk. A major problem with most external websites (notably dbSNP and BLAST but not SNPper) is the absence of a computer-readable data exchange format (XML or SOAP) as parsing HTML pages requires a lot of development time.

All sequences were masked with Repeat Masker (http://www.repeatmasker.org[14]). In the following step, SNPs are annotated by transcription factor binding site motifs obtained from TRANSFAC (http://www.gene-regulation.com[15]). This procedure selects all SNPs that result in a modified transcription factor binding site as determined by TRANSFAC3 matrices. A BLAST search follows [10] by scanning for additional motifs not fully included in TRANSFAC (pu.1, c-maf, c-rel, t-bet, xbp1, STAT6) and of modified exon splice enhancer sites (5A3G, 5B3A, 5C3D, 5D, 5E, 3B, 3C, 3E, 3F, 3H) [16]. Promotor sequences are searched with Proscan 1.7 by scoring homologies with putative eukaryotic Pol II promoter sequences (http://bimas.dcrt.nih.gov/molbio/proscan/[17]) and CpG islands (http://www.bioinfo.de/isb/2003/03/0021/main.html[18]) To also cover the 3' UTR we search for 5' adenylate uridylate (AU)-rich element groups (http://rc.kfshrc.edu.sa/ared[19]).

The sequence is then loaded in a 1-dimensional array and all annotation written into a second two-dimensional array with the feature name in the first dimension and all other data in the second dimension. The main routine then loops through the feature array and highlights in the genomic sequence all SNPs with overlapping positions. A sequence/codon counter on the left with padded spaces and a comment block on the right completes the graphical output.

Further sequence tracks can be supplemented with additional interlinear sequence tracks, for example with mouse sequence data. It turns out that these sequences need further processing, as UCSC provides only rough coverage of the homologous region but not a base-by-base alignment. We have therefore extended the pipeline with an ungapped alignment of the mouse FASTA sequence by using the BL2SEQ utility from the BLAST package [10]. It is important to use ungapped alignments as otherwise the human sequence would be extended with inserts. As there is also considerable intragenic sequence homology it was important to fill in mouse sequence tracks only with the longest alignment and to not allow short stretches to overwrite the primary alignment.

Only one UCSC assembly was used during the first run as this sequence will serve as a backbone for all other data. Later updates are possible by starting the pipeline from scratch but adding all data into new tables. User generated local SNP annotation will be preserved during migration to a later release as this information is stored in a separate table.

For benchmarking we used a set of validated SNPs from an earlier genotyping project [6] where allele frequencies have been determined in several pools of Caucasians. Briefly, for quantification of individual DNA samples, dsDNA-specific PicoGreen fluorescent dye (Molecular Probes, Eugene, Ore., USA) was used on a Genios fluorescence plate reader (Tecan). Before pooling by mixing equimolar amounts of genomic DNA, all samples were carefully adjusted to the same concentration. The determination of allele frequencies in pooled DNA was based on matrix-assisted laser desorption/ionisation time-of-flight (MALDI-TOF) mass spectrometry of allele-specific primer extension products. Primers were obtained from MWG Biotech AG (Ebersberg, Germany) and Metabion GmbH (Planegg-Martinsried, Germany). The reaction volume of 50 μl contained 17 ng of pooled genomic DNA, 2 pmol of the first sequence-specific primer with a universal sequence at the 5' end, 25 pmol of the second primer, 10 pmol of a biotinylated universal primer, 0.2 mM of each dNTP, 1.5 mM MgCl2, reaction buffer, and one unit of Thermo-Start DNA Polymerase according to the manufacturer's protocol (ABgene, Epsom, U.K.). PCR conditions were an initial denaturation step for 10 min at 95°C followed by 45 cycles of 20 sec at 95°C, 30 sec at 56°C, 30 sec at 72°C, and a final extension step for 10 min at 72°C. Each PCR was replicated three times. The biotinylated universal primer produced DNA strands complementary to the PROBE (primer oligo base extension) primer. Allele-specific primer extensions were performed using the Mass EXTEND Reagents Kit based on biotin-streptavidin binding of the generated PCR products to paramagnetic beads on the MULTIMEK 96 automated 96-channel robot (Beckman Coulter, Fullerton, California, U.S.A.). Primer extension products were loaded onto four positions of a 384-element chip nanoliter pipetting system (SpectroCHIP, SpectroJet, Sequenom) and analyzed using a MassARRAY mass spectrometer (Bruker Daltonik GmbH, Bremen, Germany). The resulting mass spectra were processed and analyzed for peak identification, peak area calculation and allele frequency estimation using the SpectroTYPER RT 2.0 software (Sequenom Inc, San Diego).

Authors contributions

The author developed the ideas presented in this paper, programmed the database and software, did the analysis, drafted and typed the report.

Acknowledgment

I wish to thank M. Emfinger for proof-reading of the manuscript, N. Herbon and Dr. M. Werner for genotyping and many helpful discussions, Dr. T. Faus-Kessler for help with R software, Prof. Wingender for the generous TRANSFAC support and both reviewers for their helpful comments.

References

1. Antonarakis SE, Cooper DN: **Mutations in Human Genetic Diseases.** *Nature Encyclopedia of the Human Genome* 2003:227-253.
2. Rica A, Kohanene S: **SNPper: retrieval and analysis of human SNPs.** *Bioinformatics* 2002, **18:**1681-1685.
3. Marsh S, Kwok P, McLeod HL: **SNP databases and pharmacogenetics: great start, but a long way to go.** *Hum Mutat* 2002, **20:**174-179.
4. Chang H, Fujita T: **PicSNP: A browsable catalog of nonsynonymous single nucleotide polymorphisms in the human genome.** *Biochem Biophys Res Commun* 2002, **287:**288-291.
5. Taylor NE, Greene EA: **ParseSNP: a tool for the analysis of nucleotide polymorphisms.** *Nucl Acid Res* 2003, **31:**3808-3811.
6. Herbon N, Werner M, Braig C, Gohlke H, Dütsch G, Illig T, Altmüller J, Hampe J, Lantermann A, Schreiber S, Bonifacio E, Ziegler A, Schwab S, Wildenauer D, van den Boom D, Braun A, Knapp M, Reitmeir P, Wjst M: **High resolution SNP scan of chromosome 6p21 in pooled samples from patients with complex diseases.** *Genomics* 2003, **81:**510-518.
7. Metzgar D: **Mutation rates: Evolution.** *Nature Encyclopedia of the Human Genome* 2003:215-217.
8. Reich DE, Gabriel SB, Altshuler D: **Quality and completeness of SNP databases.** *Nat Gen* 2003, **33:**457-458.
9. Loots GG, Locksley RM, Blankespoor CM, Wang ZE, Miller W, Rubin EM, Frazer KA: **Identification of a coordinate regulator of interleukins 4, 13, and 5 by cross-species sequence comparisons.** *Science* 2000, **288:**136-140.
10. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool.** *J Mol Biol* 1990, **215:**403-410.
11. Wjst M: **An internet linkage and mutation database for the complex phenotype asthma.** *Bioinformatics* 1998, **14:**827-828.
12. Perez-Iratxeta C, Bork P, Andrade MA: **Association of genes to genetically inherited diseases using data mining.** *Nat Genet* 2002, **31:**316-319.
13. Safran M, Chalifa-Caspi V, Shmueli O, Olender T, Lapidot M, Rosen N, Shmoish M, Peter Y, Glusman G, Feldmesser E, Adato A, Peter I, Khen M, Atarot T, Groner Y, Lancet D: **Human Gene-Centric Databases at the Weizmann Institute of Science: GeneCards, UDB, CroW 21 and HORDE.** *Nucleic Acids Research* 2003, **31:**142-146.
14. Bedell JA, Korf I, Gish W: **MaskerAid: a performance enhancement to RepeatMasker.** *Bioinformatics* 2000, **16:**1040-1041.
15. Matys V, Fricke E, Geffers R, Gossling E, Haubrock M, Hehl R, Hornischer K, Karas D, Kel AE, Kel-Margoulis OV, Kloos DU, Land S, Lewicki-Potapov B, Michael H, Munch R, Reuter I, Rotert S, Saxel H, Scheer M, Thiele S, Wingender E: **TRANSFAC: transcriptional regulation, from patterns to profiles.** *Nucleic Acids Res* 2003, **31:**374-378.
16. Cartegni L, Chew SL, Krainer AR: **Listening to silence and understanding nonsense: exonic mutations that affect splicing.** *Nature Reviews Genetics* 2002, **3:**285-298.
17. Prestridge DS: **Predicting Pol II promoter sequences using transcription factor binding sites.** *J Mol Biol* 1995, **249:**923-932.
18. Takai D, Jones PA: **Comprehensive analysis of CpG islands in human chromosomes 21 and 22.** *Proc Natl Acad Sci U S A* 2002, **99:**3740-3745.
19. Bakheet T, Frevel M, Williams BR, Greer W, Khabar KS: **ARED: Human AU-rich element-containing mRNA database reveals an unexpectedly diverse functional repertoire of encoded proteins.** *Nucleic Acids Res* 2001, **29:**246-254.