D466–D470 *Nucleic Acids Research, 2005, Vol. 33, Database issue*
doi:10.1093/nar/gki045

Improvements to GALA and dbERGE II: databases featuring genomic sequence alignment, annotation and experimental results

Laura Elnitski[1,2,*], Belinda Giardine[1], Prachi Shah[1], Yi Zhang[1], Cathy Riemer[1], Matthew Weirauch[4], Richard Burhans[1], Webb Miller[1,3] and Ross C. Hardison[2]

[1]Department of Computer Science and Engineering, [2]Department of Biochemistry and Molecular Biology and [3]Department of Biology, The Pennsylvania State University, University Park, PA 16802, USA and [4]Department of Computer Science, University of California, Santa Cruz, CA 95064, USA

Received August 14, 2004; Revised and Accepted September 28, 2004

ABSTRACT

We describe improvements to two databases that give access to information on genomic sequence similarities, functional elements in DNA and experimental results that demonstrate those functions. GALA, the database of Genome ALignments and Annotations, is now a set of interlinked relational databases for five vertebrate species, human, chimpanzee, mouse, rat and chicken. For each species, GALA records pairwise and multiple sequence alignments, scores derived from those alignments that reflect the likelihood of being under purifying selection or being a regulatory element, and extensive annotations such as genes, gene expression patterns and transcription factor binding sites. The user interface supports simple and complex queries, including operations such as subtraction and intersections as well as clustering and finding elements in proximity to features. dbERGE II, the database of Experimental Results on Gene Expression, contains experimental data from a variety of functional assays. Both databases are now run on the DB2 database management system. Improved hardware and tuning has reduced response times and increased querying capacity, while simplified query interfaces will help direct new users through the querying process. Links are available at http://www.bx.psu.edu/.

INTRODUCTION

The volume of genomic information from sequenced vertebrate species, including sequence data, extensive annotation and gene expression patterns, requires bioinformatic tools for organization and interpretation. Genome browsers, such as the UCSC Genome Browser (1,2), Ensembl (3) and Map Viewer at NCBI (4), provide views of genes and genomic regions with user-selected annotations.

To provide querying capacity across data types, we developed GALA, a Genome Alignment and Annotation database. The first release recorded whole-genome human–mouse alignments along with extensive annotation of the human genome in a relational database (5). An example of the use of GALA is to find highly conserved regions that do not code for proteins; some of these could have novel functions. A second database, dbERGE, was originally designed as a repository of Experimental Results on Gene Expression (6). The initial implementation used a custom-designed database management system, and the interface for novice users provided limited querying capacity.

This report describes improvements to GALA and dbERGE II, such as new database management systems, improved query interfaces, additional types of data and expansion to include additional species. Technical advances have improved the response time. New connections allow results from queries on dbERGE II to be imported into GALA. Output can be obtained in many formats, including exporting to the UCSC or Ensembl browsers as custom tracks or interactive viewing of alignments with Laj (7). Links to additional data-mining resources and repositories, such as EnsMart (8) and rVista (9), enhance the utility of the database.

GALA DESCRIPTION AND STRUCTURE

Query form and results pages

The GALA database, available at http://www.bx.psu.edu/, has undergone numerous improvements and expanded in size

*To whom correspondence should be addressed. Tel: +1 814 865 4747; Fax: +1 814 863 6699; Email: llb111@psu.edu since its original description in 2003 (see Supplementary Figure 1). Users choose (i) the species of interest (human, mouse, rat or chicken at the time of this writing, chimpanzee will be added soon) and (ii) the assembly for each species (e.g. hg 13-17, mm 3-5). Simple queries on one or a small number of fields are supported on the query page, and complex queries involving operations on one or more query results are supported on the history page. The statistics page shows users the number of entries in each data type and also provides easy access to downloads of the data.

The query page groups categories analogously to those at the UCSC Genome Browser: genes, mRNA, expression, regulation, etc. The current version of this layout allows users to expand the categories to see any or all options simultaneously (using the 'refresh form' button). A new version under development guides a user through the pages with a limited number of choices for ease of use. Long category lists have been moved to alternate pages and can be browsed and selected using buttons and check boxes. This is especially valuable for lists that have hundreds of entries (e.g. transcription factor binding sites), where the web browser's 'find' option can be used to quickly and easily locate terms in the list.

We have added two new features to GALA's query form; user ranges and an option to retain overlapping regions. User ranges are analogous to custom tracks on the genome browsers, i.e. they are DNA segments of interest to the user. Users can submit ranges of interest on the history page, after which they can be used in compound queries or displayed using any of GALA's output choices. In the second new option, the database will allow overlapping ranges rather than combining them. The default option of collapsing these ranges helps to reduce the memory and time needed to compute compound queries and display the results. However, in some cases users would prefer to see overlapping datasets, such as expressed sequence tags (ESTs) and alternative splice forms of genes.

More types of output are now provided. For instance, the original options for viewing results in the UCSC Genome Browser or seeing alignments in the Laj interactive alignment viewer are now complemented by custom tracks in BED format for the UCSC Browser and textual alignments in machine readable and visual formats. The region and gene summary result pages are useful for browsing results, and are limited to the first 20 pages of output. Buttons at the bottom of these pages can switch the display format to that of other output options. Furthermore, the region summary page provides access to the DNA strand (+ or −) in the FASTA format for all regions (from queries with 5000 or fewer results).

History page

User histories are now stored internally in the database, with only an identifier stored as a cookie on the user's machine; therefore GALA can store a greater number of history items (the previous limit was 15). Furthermore, queries that are frequently selected for use in compound queries are precomputed to save time. The intersection option now has expanded choices to yield more precise results. These include returning coordinates in the first or second datasets, only the overlapping segments, or the union of all regions that overlap (see Supplementary Figure 1 and diagrams at http://gala.cse.psu.edu/help.html).

New data tracks

The number and variety of data types continues to grow; a full list of data fields is at http://gala.cse.psu.edu/outline.html. New genome-wide datasets emphasize gene expression and gene regulation, including (i) three- and five-way multi-species alignments; (ii) regulatory potential scores [which evaluate how proper alignments match patterns that are distinctive for known regulatory regions (10)]; (iii) PhyloHMM Cons scores [which give a likelihood that an alignment results from purifying selection, given local variation in the neutral rate of evolution (11)]; (iv) matches to conserved transcription factor binding sites in any combination of species (based on position weight matrix scores mapped onto all sequences); and (v) microarray expression data from GNF (12). Datasets of novel functional DNA sequences have also been added, such as microRNAs (13), known regulatory regions (10), and functional and predicted promoters (14).

An example of the use of these new data fields is shown in Figure 1. The 450 kb at the end of the short arm of human chromosome 16 contains the cluster of genes *HBZ HBA2*, *HBA1* and *HBQ1*, encoding alpha-like globins, plus several other genes. The distal major regulatory element, or MRE, required for expression of the alpha-like globin genes, is 40 kb telomeric to the *HBZ* gene, and is located in one of the introns of a different gene, called *C16orf35*. Analyses of human–mouse alignments or alignments of multiple vertebrate species were not sufficient to uniquely identify the MRE within the 450 kb region (15); the MRE is conserved between human and mouse but many other non-coding regions were conserved as well. We then used GALA to find all the non-coding regions that were highly likely to be under the purifying selection, based on the phyloHMMcons score (16), and we found 199 segments (Figure 1). We also found 41 matches to GATA-1 binding sites that are conserved among human, mouse and rat. Both these datasets include the MRE, but obviously they include other segments as well. The clustering function on the GALA history page was used to find five conserved matches to the GATA-1 binding sites that have another one within 100 bp. Finally, the intersection operation of GALA was used to find four segments that pass the phyloHMMcons threshold and have a clustered pair of conserved matches to GATA-1 binding sites. Two of these are so close that they show as a single vertical line; these are within the MRE. The other two are close to regions that are erythroid DNase hypersensitive sites (17).

Linkage to other databases

Additional data repositories such as HbVar (18), dbERGE II and rVista are set up to optionally send their results to the GALA database. These external repositories enrich the choice of available data, and a user can take advantage of GALA's compound querying capacity and output formats for use with these automatically imported datasets. Furthermore, GALA is linked to additional output resources such as the alignment servers zPicture (19), Mulan (http://mulan.dcode.org/; I. Ovcherenko, manuscript in preparation) and EnsMart (8). This wider range of output choices allows a user to see customized alignment tracks that are specific to their query results.

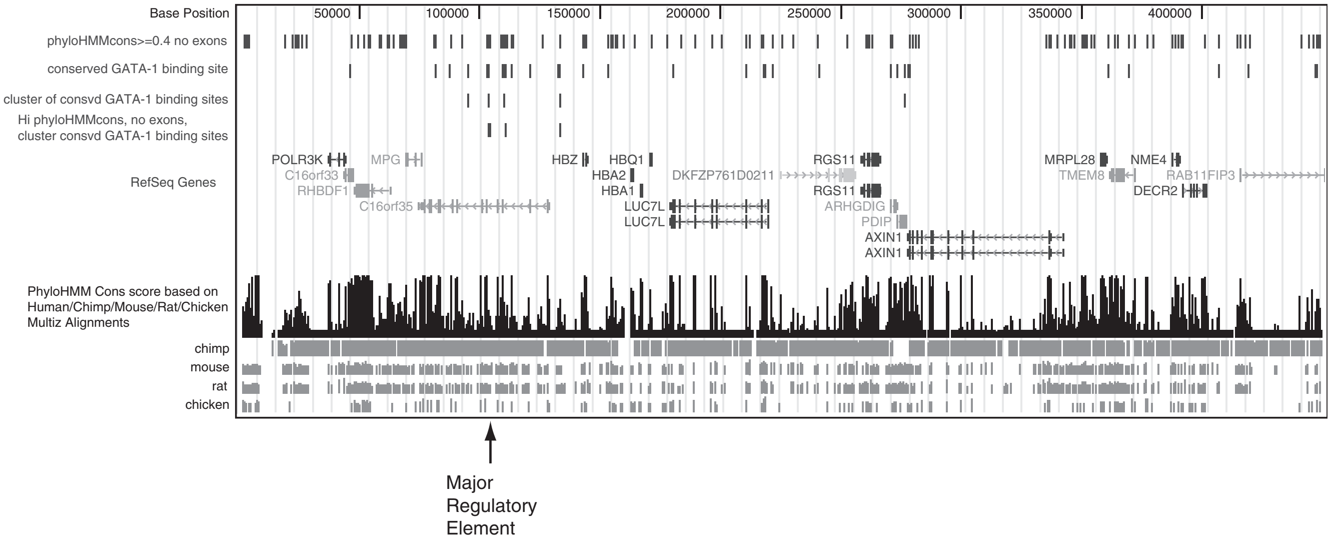

Figure 1. Use of GALA to identify a known regulatory element using features of alignments and annotations. The results of four queries on GALA are displayed as custom tracks in the UCSC Genome Browser (1,2), along with genes and the phyloHMM conservation track (21). The distal major regulatory element for the alpha-like globin genes is marked by an arrow. The track 'phyloHMMcons>=0.4 no exons' shows all the DNA segments with a phyloHMMcons score of at least 0.4, which is a rather stringent threshold, after subtracting all the segments exceeding the threshold that overlapped an exon. The track 'conserved GATA-1 binding site' shows the matches to GATA-1 binding site weight matrices in TRANSFAC (24) that are also conserved in human, mouse and rat. These match the weight matrices in those species as well. The track 'cluster of consvd GATA-1 binding sites' shows the conserved GATA-1 binding sites that have another one within 100 bp. The track 'Hi phyloHMMcons, no exons, cluster consvd GATA-1 binding sites' shows all the segments from the first data track that include one of the clusters on the third data track.

Updates for new assemblies of genome sequences

Apart from adding new data tracks to the current version of GALA, we stay up-to-date with the latest sequence assemblies, maintaining up to three builds for each species. At the time of this writing, we have three human (June and November 2002, and April 2003), three mouse, one rat and one chicken build, chimpanzee will be added soon. Addition to the May 2004 version of the human sequence is underway.

Hardware and database management

GALA is now housed on a SUN V880 computer with eight processors and 16 GB RAM. Queries are notably faster than the previous version, and the total capacity for queriable items (i.e. genomic DNA segments) is currently at 50 million (a 10-fold increase). This increased capability accommodates queries on some of our largest datasets such as transcription factor binding sites or regulatory potential scores (10). We are now using IBM's DB2 as our database management system.

Additional improvements

We have expanded the documentation for GALA including the FAQs, a new outline of the data fields in GALA, and a download page. Other improvements include queuing of background queries, allowing URL links as a method to input data, and creating links between different GALA species/assemblies based on orthologous gene sets and whole genome alignments generated with the program blastz (20). These are the same alignments that are available on the UCSC Genome Browser.

Future prospects

Planned improvements to the GALA database include extracting data from additional databases such as EnsMart and the UCSC Table Browser. This approach could be extended to additional databases as well. Additional updates are aimed at minimizing query times, and automating the addition of new data (both new releases and new tracks within releases). We also plan to provide a wider variety of user interfaces for more convenient access to GALA's capabilities.

dbERGE II DESCRIPTION AND STRUCTURE

Query form and results pages

The general aim of the dbERGE II database, available at http://www.bx.psu.edu/, has not changed with this database upgrade, i.e. to store detailed data from various types of experiments such as DNA transfer (transfections and transgenic mice), binding assays (gel shift, *in vitro* and *in vivo* footprints and methylation interference), DNase hypersensitivity and chromatin immunoprecipitation microarray (ChIP-chip) experiments. However, almost every other aspect of this database has changed since it was last reported in 1998. For example, the nested tables of the original database are now in a relational database that can be queried with a graphical query interface similar to that of the GALA database, rather than having to write arcane programmed queries to access the full capability of the database. The new guided, step-wise query interface enables both simple and complex querying capacities for users with varying levels of experience without the need to learn a programming language.

The query interface has been redesigned for user convenience. Each query is a step-by-step process that guides a user in a simple and efficient way through the large number of query options. The interface provides complex querying capacity while maintaining an intuitive and non-intimidating approach by splitting the query process into stages. Queries can be constructed to narrow-down and refine results from any data-type recorded in the database, including type of assay, tissue type, protein name, expression levels, author, etc. Experimental regions are recorded with respect to DNA ranges or chromosome locations on the genome. Thus, data are easily output to linked databases like GALA and the UCSC Genome Browser, and upgraded between genome assemblies.

Once the query choices are selected, a user is prompted to select the output format. Output choices can be text-based, graphical or the results can be uploaded into other databases. For instance, DNA transfer experiments are displayed graphically in the UCSC Genome Browser as DNA segments corresponding to the genomic regions used in each transfer experiment, flanked by the expression level and type of tissue. Multiple constructs are stacked in a custom browser track for visual comparison with genomic annotations such as the position of known genes. Viewing several tracks simultaneously allows a user to see multiple experimental features such as hypersensitive sites, expressed genes and functional promoters, in order to build evidence for the location of functional elements. These views can be scaled to show varying levels of detail through the variety of query display options, such as 'dense, packed and full' offered in the Genome Browser (http://genome.cse.ucsc.edu/). Textual displays allow a choice of the amount of detail to be returned in the output file, describing information about the references, DNA transfer experiments and binding assays. Users wishing to parse large data files can soon request XML output as a form of computer readable, structured results. A schematic illustration of the types of experimental data, the data sources and output options is illustrated in Supplementary Figure 2.

In addition to detailed experimental results, the database accepts and displays a second type of data known as a summary. This simplified data type is optimal for entering high-throughput data along with general conclusions, such as 'hypersensitive site' 'functional promoter', etc., without all the experimental details. Additional bits of information can be included, such as tissue type. Summaries can be queried as a group or by limiting the query to those that fall within a specific genomic locus.

Data entry

Data is entered into the dbERGE II database through a graphical data entry interface with permission, or through a more automated manner using a formatted spreadsheet. Individuals working on expression assay studies are encouraged to load their data into the dbERGE II database.

History page

The dbERGE II database has a user history page analogous to the one in GALA. It stores results of simple queries and allows a user to combine them as unions, intersections or exclude them with a 'NOT' statement. Formats for output files created by compound queries are the same as for simple queries. Each query listed on the history page is associated with an automatically generated query description and a count of the number of experiments returned from the query.

Data

The database holds several types of data including DNase hypersensitive sites (21), ChIP–chip binding sites data (22), functional promoters (23), locus-specific experimental data from the β-globin and CFTR loci, with plans to house non-microarray data generated from the Encyclopedia of DNA Elements (ENCODE) project.

Future prospects

In addition to the features described, the database is scalable to include more experiment types and species. Current efforts are aimed at reducing response times, expanding experimental data types and interaction with other databases, such as Array Express (http://www.ebi.ac.uk/arrayexpress/) and GEO (http://www.ncbi.nlm.nih.gov/geo/), which hold microarray data. Additionally, data from the ENCODE project (http://www.genome.gov/ENCODE/) can be handled and displayed by the dbERGE II database. Most data types (http://www.genome.gov/11009066) are already implemented, such as DNase hypersensitivity, ChIP-chip analyses, enhancer and promoter assays, etc. Furthermore, having links to GALA and UCSC for the comparative display of data will help to analyze the biological implications of the results.

DISCUSSION

We have described improvements to two separate, yet connected, complementary databases, GALA and dbERGE II. EnsMart is another example of a queriable database of genomic data, with a different approach to querying the data. For example, GALA queries on DNA coordinates whereas EnsMart queries on genomic entities (genes or SNPs), asking a user to further refine a query from there. Thus genomic sequences can only be queried with EnsMart if they are within or adjacent to a gene or SNP. By providing a history page, GALA is flexible in allowing a wide variety of combinations of queries. All operations are available from the history page, regardless of the order in which the component queries were conducted.

dbERGE II is unlike any other database currently available to the public. It houses both detailed experimental data and summary of experimental results and provides a convenient query interface and display options for quick retrieval of interesting data. Its applications include archiving data in a queriable manner, finding all experimental results in a targeted genomic locus and assessing biological meaning from high-throughput data. The ENCODE project (http://www.genome.gov/10005107) is submitting 1% of the human genome to exhaustive experimental testing for transcription, transcription factor occupancy, DNase hypersensitive sites, replication timing and other features. Pilot experiments have shown that dbERGE II supports the recording and querying of data similar to those in the ENCODE project, and query results can be readily displayed in a genome browser or in the graphical viewer resident in dbERGE II. Thus, we anticipate that dbERGE II will be one of the tools used to record and analyze the ENCODE data.

SUPPLEMENTARY MATERIAL

Supplementary Material is available at NAR Online.

REFERENCES

1. Kent,W.J., Sugnet,C.W., Furey,T.S., Roskin,K.M., Pringle,T.H., Zahler,A.M. and Haussler,D. (2002) The human genome browser at UCSC. *Genome Res.*, **12**, 996–1006.
2. Karolchik,D., Baertsch,R., Diekhans,M., Furey,T.S., Hinrichs,A., Lu,Y.T., Roskin,K.M., Schwartz,M., Sugnet,C.W., Thomas,D.J. *et al.* (2003) The UCSC Genome Browser Database. *Nucleic Acids Res.*, **31**, 51–54.
3. Birney,E., Andrews,D., Bevan,P., Caccamo,M., Cameron,G., Chen,Y., Clarke,L., Coates,G., Cox,T., Cuff,J., Curwen,V., Cutts,T., Down,T., Durbin,R., Eyras,E., Fernandez-Suarez,X.M., Gane,P., Gibbins,B., Gilbert,J., Hammond,M., Hotz,H., Iyer,V., Kahari,A., Jekosch,K., Kasprzyk,A., Keefe,D., Keenan,S., Lehvaslaiho,H., McVicker,G., Melsopp,C., Meidl,P., Mongin,E., Pettett,R., Potter,S., Proctor,G., Rae,M., Searle,S., Slater,G., Smedley,D., Smith,J., Spooner,W., Stabenau,A., Stalker,J., Storey,R., Ureta-Vidal,A., Woodwark,C., Clamp,M. and Hubbard,T. (2004) Ensembl 2004. *Nucleic Acids Res.*, **32** (Database issue), D468–D470.
4. Wheeler,D.L., Church,D.M., Edgar,R., Federhen,S., Helmberg,W., Madden,T.L., Pontius,J.U., Schuler,G.D., Schriml,L.M., Sequeira,E., Suzek,T.O., Tatusova,T.A. and Wagner,L. (2004) Database resources of the National Center for Biotechnology Information: update. *Nucleic Acids Res.*, **32** (Database issue), D35–D40.
5. Giardine,B.M., Elnitski,L., Riemer,C., Makalowska,I., Schwartz,S., Miller,W. and Hardison,R.C. (2003) GALA, a database for genomic sequence alignments and annotations. *Genome Res.*, **13**, 732–741.
6. Riemer,C., ElSherbini,A., Stojanovic,N., Schwartz,S., Kwitkin,P.B., Miller,W. and Hardison,R. (1998) A database of experimental results on globin gene expression. *Genomics*, **53**, 324–337.
7. Wilson,M.D., Riemer,C., Martindale,D.W., Schnupf,P., Boright,A.P., Cheung,T.L., Hardy,D.M., Schwartz,S., Scherer,S.W., Tsui,L.C., Miller,W. and Koop,B.F. (2001) Comparative analysis of the gene-dense ACHE/TFR2 region on human chromosome 7q22 with the orthologous region on mouse chromosome 5. *Nucleic Acids Res.*, **29**, 1352–1365.
8. Kasprzyk,A., Keefe,D., Smedley,D., London,D., Spooner,W., Melsopp,C., Hammond,M., Rocca-Serra,P., Cox,T. and Birney,E. (2004) EnsMart: a generic system for fast and flexible access to biological data. *Genome Res.*, **14**, 160–169.
9. Loots,G.G. and Ovcharenko,I. (2004) rVISTA 2.0: evolutionary analysis of transcription factor binding sites. *Nucleic Acids Res.*, **32**, 217–221.
10. Elnitski,L., Hardison,R.C., Li,J., Yang,S., Kolbe,D., Eswara,P., O'Connor,M.J., Schwartz,S., Miller,W. and Chiaromonte,F. (2003) Distinguishing regulatory DNA from neutral sites. *Genome Res.*, **13**, 64–72.
11. International Mouse Genome Sequencing Consortium (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature*, **420**, 520–562.
12. Su,A., Cooke,M., Ching,K., Hakak,Y., Walker,J., Wiltshire,T., Orth,A., Vega,R., Sapinoso,L., Moqrich,A. *et al.* (2002) Large-scale analysis of the human and mouse transcriptomes. *Proc. Natl Acad. Sci. USA*, **99**, 4465–4470.
13. Griffiths-Jones,S. (2004) The microRNA Registry. *Nucleic Acids Res.*, **32**, 109–111.
14. Trinklein,N.D., Aldred,S.J., Saldanha,A.J. and Myers,R.M. (2003) Identification and functional analysis of human transcriptional promoters. *Genome Res.*, **13**, 308–312.
15. Flint,J., Tufarelli,C., Peden,J., Clark,K., Daniels,R.J., Hardison,R., Miller,W., Philipsen,S., Tan-Un,K.C., McMorrow,T. *et al.* (2001) Comparative genome analysis delimits a chromosomal domain and identifies key regulatory elements in the alpha globin cluster. *Hum. Mol. Genet.*, **10**, 371–382.
16. Siepel,A. and Haussler,D. (2003) Combining phylogenetic and hidden Markov models in biosequence analysis. In *Proceedings of the Seventh Annual International Conference on Computational Molecular Biology (RECOMB 2003)*, Berlin, Germany, April 10–13. ACM press, NY, pp. 277–286.
17. Anguita,E., Johnson,C.A., Wood,W.G., Turner,B.M. and Higgs,D.R. (2001) Identification of a conserved erythroid specific domain of histone acetylation across the alpha-globin gene cluster. *Proc. Natl Acad. Sci. USA*, **98**, 12114–12119.
18. Patrinos,G.P., Giardine,B., Riemer,C., Miller,W., Chui,D.H.K., Anagnou,N.P., Wajcman,H. and Hardison,R.C. (2004) Improvements in the HbVar database of human hemoglobin variants and thalassemia mutations for population and sequence variation studies. *Nucleic Acids Res.*, **32**, 537–541.
19. Ovcharenko,I., Loots,G.G., Hardison,R.C., Miller,W. and Stubbs,L. (2004) zPicture: dynamic alignment and visualization tool for analyzing conservation profiles. *Genome Res.*, **14**, 4772–4777.
20. Schwartz,S., Kent,W.J., Smit,A., Zhang,Z., Baertsch,R., Hardison,R.C., Haussler,D. and Miller,W. (2003) Human–mouse alignments with BLASTZ. *Genome Res.*, **131**, 103–107.
21. Crawford,G.E., Holt,I.E., Mullikin,J.C., Tai,D., Blakesley,R., Bouffard,G., Young,A., Masiello,C., Green,E.D., Wolfsberg,T.G. and Collins,F.S., National Institutes Of Health Intramural Sequencing Center (2004) Identifying gene regulatory elements by genome-wide recovery of DNase hypersensitive sites. *Proc. Natl Acad. Sci. USA*, **101**, 992–997.
22. Cawley,S., Bekiranov,S., Ng,H.H., Kapranov,P., Sekinger,E.A., Kampa,D., Piccolboni,A., Sementchenko,V., Cheng,J., Williams,A.J., Wheeler,R., Wong,B., Drenkow,J., Yamanaka,M., Patel,S., Brubaker,S., Tammana,H., Helt,G., Struhl,K. and Gingeras,T.R. (2004) Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding RNAs. *Cell*, **116**, 499–509.
23. Trinklein,N.D., Aldred,S.J., Saldanha,A.J. and Myers,R.M. (2003) Identification and functional analysis of human transcriptional promoters. *Genome Res.*, **13**, 308–312.
24. Wingender,E., Chen,X., Fricke,E., Geffers,R., Hehl,R., Liebich,I., Krull,M., Matys,V., Michael,H., Ohnhauser,R. *et al.* (2001) The TRANSFAC system on gene expression regulation. *Nucleic Acids Res.*, **29**, 281–283.