Published online 21 October 2010 *Nucleic Acids Research, 2011, Vol. 39, Database issue* ***D75–D79***
doi:10.1093/nar/gkq942

NGSmethDB: a database for next-generation sequencing single-cytosine-resolution DNA methylation data

Michael Hackenberg*, Guillermo Barturen and José L. Oliver*

Dpto. de Genética, Facultad de Ciencias, Universidad de Granada, 18071-Granada and Lab. de Bioinformática, Inst. de Biotecnología, Centro de Investigación Biomédica, 18100-Granada, Spain

Received August 14, 2010; Revised September 28, 2010; Accepted September 30, 2010

ABSTRACT

Next-generation sequencing (NGS) together with bisulphite conversion allows the generation of whole genome methylation maps at single-cytosine resolution. This allows studying the absence of methylation in a particular genome region over a range of tissues, the differential tissue methylation or the changes occurring along pathological conditions. However, no database exists fully addressing such requirements. We propose here NGSmethDB (http://bioinfo2.ugr.es/NGSmethDB/gbrowse/) for the storage and retrieval of methylation data derived from NGS. Two cytosine methylation contexts (CpG and CAG/CTG) are considered. Through a browser interface coupled to a MySQL backend and several data mining tools, the user can search for methylation states in a set of tissues, retrieve methylation values for a set of tissues in a given chromosomal region, or display the methylation of promoters among different tissues. NGSmethDB is currently populated with human, mouse and *Arabidopsis* data, but other methylomes will be incorporated through an automatic pipeline as soon as new data become available. Dump downloads for three coverage levels (1, 5 or 10 reads) are available. NGSmethDB will be useful for experimental researchers, as well as for bioinformaticians, who might use the data as input for further research.

INTRODUCTION

DNA methylation is a common epigenetic mark that can be found in eukaryotes exclusively at cytosine residues ($5^{me}C$). This modification has important roles in embryonic development, as shown by early lethality in mice that lack DNA methyltransferases (DNMTs), the inactivation of the X chromosome in female cells or the establishment and maintenance of allele-specific expression of imprinted genes (1–3). Numerous studies over the past decades suggest that cytosine DNA methylation functions to maintain the repressed chromatin state and therefore stably silence promoter activity (4). In animal genomes, the predominantly methylated sequence context is the dinucleotide CpG, while non-CpG methylation exists in plants that is targeted to transposable elements by a mechanism that depends upon small interfering RNAs (5). Recently, methylation at sequence contexts CHH and CHG has been detected in human undifferentiated cells (6).

Many different techniques have been developed for DNA methylation profiling (7,8). The detection methods can be divided into a methylation-dependent pretreatment and an analytical step. The first step is necessary as $5^{me}C$ is not readily distinguished from unmethylated cytosine by hybridization-based methods and PCR amplification erases the DNA methylation information. Basically, three different pretreatments can be distinguished: enzyme digestion, affinity enrichment (immunoprecipiation) and sodium bisulphite conversion. The information on the DNA methylation is finally read out by a gel-based, array-based or sequencing-based analysis. Virtually, all combinations of these two steps exist. Depending on the specific combination used, we can distinguish between 'single cytosine' and 'region wide' profiling of methylation states. The region wide methods detect normally the methylation states of known CpG islands or unmethylated fragments using either enzyme digestion or immunoprecipitation. There are several drawbacks with these methods. Apart from the errors introduced by the methylation-dependent pretreatment, only 'mean values' of the regions can be detected. Although for many experiments it might be sufficient to get information whether a

*To whom correspondence should be addressed. Tel: +34 958243261; Fax: +34 958244073; Email: mlhack@gmail.com
Correspondence may also be addressed to José L. Oliver. Tel: +34 958243261; Fax: +34 958244073; Email: oliver@ugr.es

given region is methylated or unmethylated, for others it will be not. For example, recently it has been shown that many CpG islands show internal fluctuations that can be resolved by means of single-cytosine resolution analysis (9,10). Furthermore, single-cytosine resolution data can be critical to resolve the methylation states, and the possible functionality, of very small islands (islets) or even orphan CpG dinucleotides (10,11).

However, to completely exploit the full potential of single-base resolution whole genome methylation maps, a specifically designed database is needed. Given the lack of single base data in the past, current databases are only focused either on specific regions and/or on pathologic situations (12,13).

In the next years, however, whole genome methylation data will become available for many new tissues, pathological conditions and species and it will be of critical importance to store and unify this information in an adequate way. We therefore propose here NGSmethDB, a database for single-cytosine resolution methylation data. The database uses a web interface based on GBrowse (14) and coupled to a MySQL backend, which allows to visualize the methylation data in a genomic context together with many other annotations, as well as full data downloads. In addition, a set of powerful data mining tools are also implemented, so the user can filter, analyze and retrieve data in many different ways. For example, the user can search for unmethylated or differentially methylated cytosines in a selected set of tissues, or display and analyze the promoter methylation of RefSeq genes. Finally, the database extends the commonly used focus on CpG dinucleotides to the recently discovered non-CpG targets for DNA methylation in undifferentiated tissues (6).

FEATURES AND SCOPE

The NGSmethDB database can be divided into two parts. First, the content can be visualized, together with many other common annotations, by means of a web interface based on GBrowse (14) coupled to a MySQL backend; and second, several user-friendly data mining tools are provided so the average user can generate its own data sets easily. Currently, the database holds information on three species (human, mouse and *Arabidopsis*) and 52 different tissues (21 unique tissues). Furthermore, two different methylation contexts are considered, CpG and CWG, but other non-CpG contexts, as CAH or CHH, will be soon available. Currently, the database holds methylation data of 696 599 217 cytosines for human (hg18), 69 459 481 cytosines for mouse (mm8) and 16 321 229 cytosines for *Arabidopsis* (TAIR8). A detailed and updated database statistical table is maintained on-line: http://bioinfo2.ugr.es/gbrowse2/StatGraphs/datasourcesrpt.php. A summary of the publications where the data were generated from is also maintained and updated on-line: http://bioinfo2.ugr.es/gbrowse2/DataSource/datasourcesrpt.php?start = 1.

We encourage data submissions of new methylation data in order to populate and maintain updated NGSmethDB.

For most data, the methylation information for the cytosines is directly available for the three mentioned genome assemblies. In these cases, we populate the database with these processed data. For other cases, we used the LiftOver tool (15) to convert the coordinates from other assemblies, or developed scripts to process the raw data (like *fastaq* files) in order to obtain the methylation information for all covered cytosines. All methylation values for both CpG and CWG contexts are calculated taking into account both strands. The assigned methylation value is therefore a weighted mean between the context in the direct and reverse strands. Which means that it is the sum of reads that indicate methylation (cytosine not converted to uracil/thymine) mapped to the specific position in the '+' strand and those mapped to the '−' strand, divided by the total number of reads mapped to the position regardless of the strand.

Genomic browser interface

The GBrowse genome viewer (14) connected to a MySQL backend is used to set up a web browser interface for NGSmethDB. Features of the browser include the ability to scroll and zoom through arbitrary regions of a genome, to enter a region of the genome by searching for a landmark or performing a full text search of features, as well as the ability to enable and disable feature tracks and change their relative order and appearance. The user can also upload private annotations to view them in the context of the existing ones at the NGSmethDB web site.

Apart from the methylation data, the following related annotations are currently available on the NGSmethDB browser: (i) CpGcluster CpG islands (16); (ii) Takai-Jones CpG islands (17); (iii) RefSeq genes (18); (iv) HMR conserved TFBSs (19); (v) CisRED regulatory elements (20); and (vi) the chromosome sequence (hg18, mm8 and TAIR8 genome assemblies) and G + C content.

The methylation information of a given context is represented by the coordinate of the cytosine on the direct strand. To display the methylation values of the cytosines we use a color gradient from white (methylation value = 0, unmethylated in all reads) to red (methylation value = 1, methylated in all reads). To demonstrate the usefulness of the web interface, we analyzed the promoter region of the gene TIAM1 (Figure 1). It can be seen that this promoter is differentially methylated among the different tissues.

Data mining tools

Currently, five different ways are implemented to retrieve raw data from the database. For all five possibilities, two different sequence contexts and three coverage levels exist. We detected not just the methylation values of CpG dinucleotides but also for the cytosines in a CWG (CAG or CTG) context. The methylation value at a given position (cytosine) is calculated as explained before taking both strands into consideration. We stored three different coverage levels in the database: cytosines covered by at least 1, 5 and 10 reads.

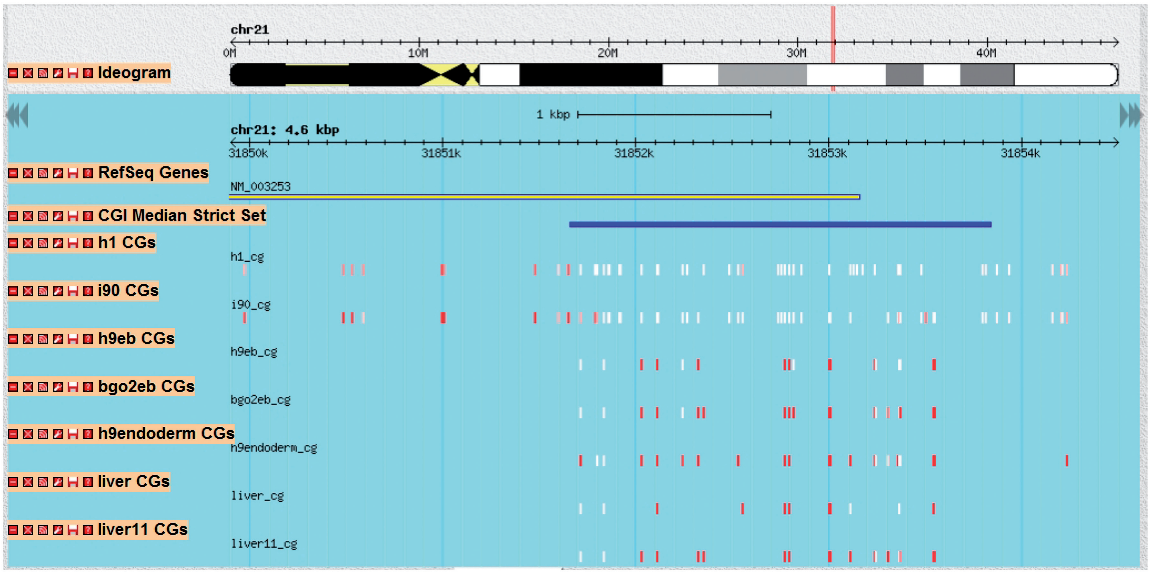


Figure 1. Visualization of methylation states of CpG dinucleotides in different tissues in the NGSmethDB genome browser. The promoter region of the gene TIAM1 (NM_003253) is shown. The different methylation values are displayed by means of a color gradient from white (unmethylated in all reads) toward red (methylated in all reads).

Dump download

This option shows first an overview of current database content, including a short description of the tissue, the genome coverage in %, a link to PubMed, and raw data files for #read ≥ 1, #read ≥ 5 and #read ≥ 10 coverage. The files show the chromosome, chromosome-start and chromosome-end coordinates, the sequence methylation context (either CpG or CWG), the number of reads and the cytosine methylation ratio.

Retrieve unmethylated contexts

This tool can be used to retrieve all unmethylated cytosines in a given set of tissues. The user has to select the sequence context (CG or CWG), the read coverage, the threshold for unmethylation (often a threshold of 0.2 is used, i.e. all cytosines with values ≤0.2 are considered to be unmethylated) and the tissues. The tool will detect all cytosine contexts showing lower methylation ratios than the chosen threshold in all selected tissues. The provided output file holds the chromosome, chromosome start- and end-coordinates and the methylation values in all selected tissues. Note that this tool can be also used to retrieve all CpGs which are present in every single analyzed tissue by setting the threshold to one. In doing so, cytosines with methylation data in all tissues will be reported regardless of its methylation state, i.e. cytosines that are not covered by at least the number of chosen coverage threshold (1, 5 or 10) in any of the analyzed tissues will not be reported in the output.

Retrieve differentially methylated contexts

By means of this tool all differentially methylated cytosine contexts can be determined in a given set of tissues. All parameters of the ‘Retrieve unmethylated contexts’ (see above) are available here, plus one additional parameter: the threshold for the methylation value which defines whether a cytosine is considered to be methylated (often a threshold of 0.8 is used, i.e. all cytosines with higher values than ≥0.8 are considered to be methylated). We define a cytosine as differentially methylated if it is unmethylated in at least one tissue and methylated in at least one other tissue. The tool reports those differentially methylated cytosine contexts that are either methylated or unmethylated in all analyzed tissues, i.e. those contexts that show intermediate methylation in only one tissue will not be reported.

Get methylation states of promoter regions

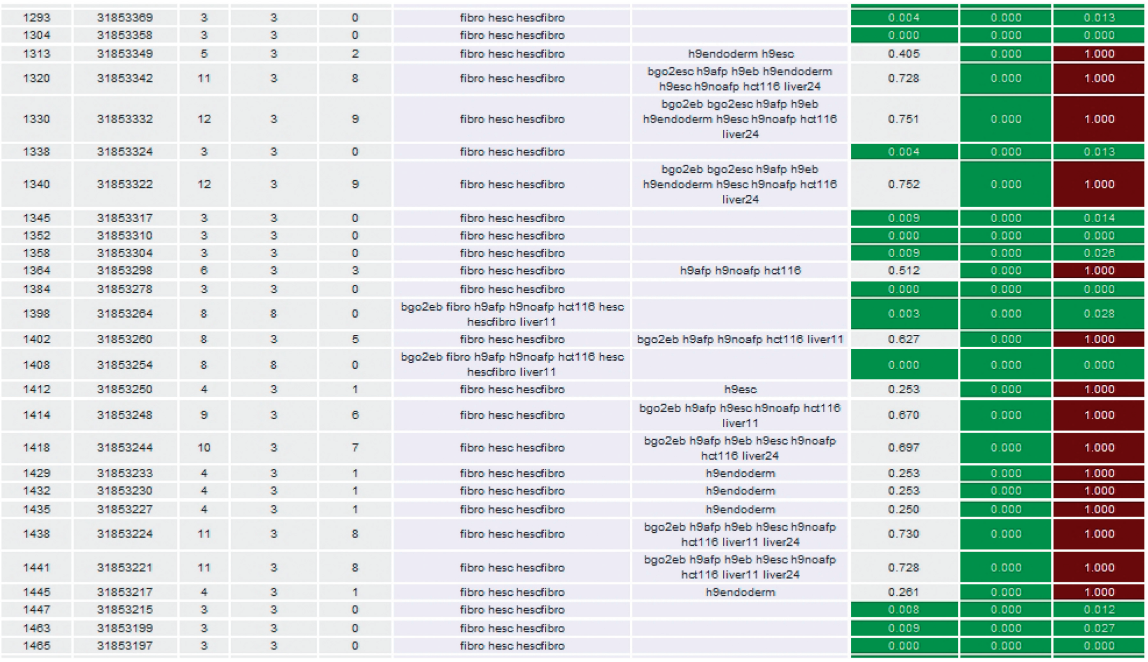
This tool allows depicting the methylation states of all cytosine contexts within the promoter region of RefSeq genes. We define the promoter region as beginning 1.5 kb upstream of the Transcription Start Site (TSS) and ending 500 bp downstream of the TSS. The user needs to provide a valid RefSeq name (NM_*) or a unique TAIR gene id (ATxGxxxxx) and the desired coverage. The output is displayed by default as an overview table that summarizes the fluctuation along the promoter as well as over the different tissues. A detailed table can also be generated (Figure 2).

Retrieve methylation data for chromosome region

All methylation values for a selected set of tissues can be retrieved for a given chromosomal region, once the user provides the start and end chromosome coordinates.

CONCLUSIONS

Over the next years, methylation data for a growing number of tissues, cell types, pathological conditions and diverse species will all be available. In most of the original



1293	31853369	3	3	0	fibro hesc hescfibro		0.004	0.000	0.013
1304	31853358	3	3	0	fibro hesc hescfibro		0.000	0.000	0.000
1313	31853349	5	3	2	fibro hesc hescfibro	h9endoderm h9esc	0.405	0.000	1.000
1320	31853342	11	3	8	fibro hesc hescfibro	bgo2esc h9afp h9eb h9endoderm h9esc h9noafp hct116 liver24	0.728	0.000	1.000
1330	31853332	12	3	9	fibro hesc hescfibro	bgo2eb bgo2esc h9afp h9eb h9endoderm h9esc h9noafp hct116 liver24	0.751	0.000	1.000
1338	31853324	3	3	0	fibro hesc hescfibro		0.004	0.000	0.013
1340	31853322	12	3	9	fibro hesc hescfibro	bgo2eb bgo2esc h9afp h9eb h9endoderm h9esc h9noafp hct116 liver24	0.752	0.000	1.000
1345	31853317	3	3	0	fibro hesc hescfibro		0.009	0.000	0.014
1352	31853310	3	3	0	fibro hesc hescfibro		0.000	0.000	0.000
1358	31853304	3	3	0	fibro hesc hescfibro		0.009	0.000	0.026
1364	31853298	6	3	3	fibro hesc hescfibro	h9afp h9noafp hct116	0.512	0.000	1.000
1384	31853278	3	3	0	fibro hesc hescfibro		0.000	0.000	0.000
1398	31853264	8	8	0	bgo2eb fibro h9afp h9noafp hct116 hesc hescfibro liver11		0.003	0.000	0.028
1402	31853260	8	3	5	fibro hesc hescfibro	bgo2eb h9afp h9noafp hct116 liver11	0.627	0.000	1.000
1408	31853254	8	8	0	bgo2eb fibro h9afp h9noafp hct116 hesc hescfibro liver11		0.000	0.000	0.000
1412	31853250	4	3	1	fibro hesc hescfibro	h9esc	0.253	0.000	1.000
1414	31853248	9	3	6	fibro hesc hescfibro	bgo2eb h9afp h9esc h9noafp hct116 liver11	0.670	0.000	1.000
1418	31853244	10	3	7	fibro hesc hescfibro	bgo2eb h9afp h9eb h9esc h9noafp hct116 liver24	0.697	0.000	1.000
1429	31853233	4	3	1	fibro hesc hescfibro	h9endoderm	0.253	0.000	1.000
1432	31853230	4	3	1	fibro hesc hescfibro	h9endoderm	0.253	0.000	1.000
1435	31853227	4	3	1	fibro hesc hescfibro	h9endoderm	0.250	0.000	1.000
1438	31853224	11	3	8	fibro hesc hescfibro	bgo2eb h9afp h9eb h9esc h9noafp hct116 liver11 liver24	0.730	0.000	1.000
1441	31853221	11	3	8	fibro hesc hescfibro	bgo2eb h9afp h9eb h9esc h9noafp hct116 liver11 liver24	0.728	0.000	1.000
1445	31853217	4	3	1	fibro hesc hescfibro	h9endoderm	0.261	0.000	1.000
1447	31853215	3	3	0	fibro hesc hescfibro		0.008	0.000	0.012
1463	31853199	3	3	0	fibro hesc hescfibro		0.009	0.000	0.027
1465	31853197	3	3	0	fibro hesc hescfibro		0.000	0.000	0.000

Figure 2. Detailed analysis of the promoter region of the gene TIAM1 (NM_003253) by means of the NGSmethDB data mining tools. The table shows the following columns: relative coordinate towards the point TSS-1.5 kb, the chromosomal coordinate of the cytosine, the number of tissues for which methylation data exists, the number of tissues where the cytosine were found to be unmethylated, the number of tissues where the cytosine were found to be methylated, the tissue names where the cytosine is methylated and unmethylated, respectively, the mean methylation value among all tissues, the minimum and maximum methylation values over all tissues. By means of the color code, green for unmethylation (value ≤ 0.2) and red for methylation (value ≥ 0.8), the situation can be rapidly analyzed. For example, if both minimum and maximum values are green for one cytosine position, this means that this cytosine is unmethylated in all analyzed tissues. On the other hand, if the minimum value is green and the maximum value is red, this indicates differential methylation over the different tissues for the given cytosine.

publications, the authors focus on concrete questions and scarcely the whole potential of the data can be exploited. To get more out of these data, a joint analysis with data from other tissues and/or species is needed. To carry out such analysis, data must be first stored in an appropriate way in a database. We propose here NGSmethDB, a new database with a very broad scope to facilitate the analysis of methylation data from different sources. Heterogeneous methylation data can be either simultaneously visualized through a powerful web interface or selectively downloaded by means of the provided data mining tools that allow the user to design new experiments and retrieve exactly the adequate data for them. Thus, we are confident that the database will be of great usefulness both for experimental and bioinformatics researchers.

FUNDING

The Spanish Government grant (BIO2008-01353 to J.L.O.); 'Juan de la Cierva' (to M.H.); Basque Country 'Programa de formación de investigadores' grant (to G.B.). Funding for open access charge: The Spanish Government grant (BIO2008-01353 to J.L.O.).

Conflict of interest statement. None declared.

REFERENCES

1. Chen,T. and Li,E. (2004) Structure and function of eukaryotic DNA methyltransferases. *Curr. Top. Dev. Biol.*, **60**, 55–89.
2. Karpf,A.R. and Matsui,S. (2005) Genetic disruption of cytosine DNA methyltransferase enzymes induces chromosomal instability in human cancer cells. *Cancer Res.*, **65**, 8635–8639.
3. Dodge,J.E., Okano,M., Dick,F., Tsujimoto,N., Chen,T., Wang,S., Ueda,Y., Dyson,N. and Li,E. (2005) Inactivation of Dnmt3b in mouse embryonic fibroblasts results in DNA hypomethylation, chromosomal instability, and spontaneous immortalization. *J. Biol. Chem.*, **280**, 17986–17991.
4. Bird,A.P. and Wolffe,A.P. (1999) Methylation-induced repression: belts, braces, and chromatin. *Cell*, **99**, 451–454.
5. Chan,S.W., Henderson,I.R. and Jacobsen,S.E. (2005) Gardening the genome: DNA methylation in Arabidopsis thaliana. *Nat. Rev. Genet.*, **6**, 351–360.
6. Lister,R., Pelizzola,M., Dowen,R.H., Hawkins,R.D., Hon,G., Tonti-Filippini,J., Nery,J.R., Lee,L., Ye,Z., Ngo,Q.M. *et al.* (2009) Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature*, **462**, 315–322.
7. Laird,P.W. (2010) Principles and challenges of genome-wide DNA methylation analysis. *Nat. Rev. Genet.*, **11**, 191–203.
8. Beck,S. and Rakyan,V.K. (2008) The methylome: approaches for global DNA methylation profiling. *Trends Genet.*, **24**, 231–237.
9. Hodges,E., Smith,A.D., Kendall,J., Xuan,Z., Ravi,K., Rooks,M., Zhang,M.Q., Ye,K., Bhattacharjee,A., Brizuela,L. *et al.* (2009) High definition profiling of mammalian DNA methylation by array capture and single molecule bisulfite sequencing. *Genome Res.*, **19**, 1593–1605.
10. Hackenberg,M., Barturen,G., Carpena,P., Luque-Escamilla,P.L., Previti,C. and Oliver,J.L. (2010) Prediction of CpG-island

function: CpG clustering vs. sliding-window methods. *BMC Genomics*, **11**, 327.

11. Wong,N.C., Wong,L.H., Quach,J.M., Canham,P., Craig,J.M., Song,J.Z., Clark,S.J. and Choo,K.H. (2006) Permissive transcriptional activity at the centromere through pockets of DNA hypomethylation. *PLoS Genet*., **2**, e17.
12. Ongenaert,M., Van Neste,L., De Meyer,T., Menschaert,G., Bekaert,S. and Van Criekinge,W. (2008) PubMeth: a cancer methylation database combining text-mining and expert annotation. *Nucleic Acids Res*., **36**, D842–D846.
13. Amoreira,C., Hindermann,W. and Grunau,C. (2003) An improved version of the DNA Methylation database (MethDB). *Nucleic Acids Res*., **31**, 75–77.
14. Stein,L.D., Mungall,C., Shu,S., Caudy,M., Mangone,M., Day,A., Nickerson,E., Stajich,J.E., Harris,T.W., Arva,A. *et al.* (2002) The generic genome browser: a building block for a model organism system database. *Genome Res*., **12**, 1599–1610.
15. LiftOver: http://genome.ucsc.edu/cgi-bin/hgLiftOver (May 2010, date last accessed).
16. Hackenberg,M., Previti,C., Luque-Escamilla,P.L., Carpena,P., Martínez-Aroza,J. and Oliver,J.L. (2006) CpGcluster: a distance-based algorithm for CpG-island detection. *BMC Bioinformatics*, **7**, 446.
17. Takai,D. and Jones,P.A. (2002) Comprehensive analysis of CpG islands in human chromosomes 21 and 22. *Proc. Natl Acad. Sci. USA*, **99**, 3740–3745.
18. Pruitt,K.D., Tatusova,T. and Maglott,D.R. (2007) NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res*., **35**, D61–D65.
19. Weirauch,M. and Raney,B. (2007) HMR Conserved transcription factor binding sites. *UCSC Genome Browser*. http://genome.ucsc.edu/cgi-bin/hgGateway (May 2010, date last accessed).
20. Robertson,G., Bilenky,M., Lin,K., He,A., Yuen,W., Dagpinar,M., Varhol,R., Teague,K., Griffith,O.L., Zhang,X. *et al.* (2006) cisRED: a database system for genome-scale computational discovery of regulatory elements. *Nucleic Acids Res*., **34**, D68–D73.