Documentation for CpG island feature generator

Generating features to distinct between methylated and unmethylated CpG islands

Clemens Wrzodek

February 17, 2012

CpG island feature generator is designed to generate LIBSVM formatted feature files for existing DNA methylation datasets, that allows users to distinguish between methylated and unmethylated CpG islands. This application can be used to read genomic locations with assigned numeric values. For example, read probe locations and intensities from DNA methylation experiments. Furthermore, the application can map the probes to overlapping CpG islands and lift locations between multiple human genome builds. For each of these locations, a LIBSVM compatible feature string is generated, that may include features, representing up to 15 categories. The user may choose the categories to generate features for. Furthermore, the application can generate support vector regression compatible files (by including the actual methylation values) or support vector classification compatible files (by converting given methylation intensities to binary states).

Afterwards, any machine learning approach can be applied to the generated features to create a model on those. The generated model can then be applied to novel data to predict, e.g., unknown methylation states of CpG islands.

Contents

1 Installation

CpG island feature generator comes as a Java JAR file. It can run out-of-the-box on all systems where a Java virtual machine is installed and does not require any further installations.

1.1 Requirements

1.1.1 Software

CpG island feature generator is entirely written in Java™ and runs on any operating system where a suitable Java Virtual Machine (JDK version 1.6 or newer) is installed. See, for example, the Java SE download page[1].

It is recommended to download the histone modification data from Barski *et al.* (2007)[2] and place them in a subdirectory, relative to the applications JAR file, called "res/HistoneMeth/". If features based on "Evolutionary conservation (PhastCons)" should be included, you have to download them from the UCSC download page[3] and place them in the directory "res/Phastcons/", relative to the applications executable (see Section 2 for more information). Further data, e.g., from Ensembl or pre-calculated features, will be downloaded automatically as required by the application.

1.1.2 Hardware

With at least 1 GB main memory, you should be able to perform most tasks without any problem. For large datasets, you should have at least 2 GB of main memory.
An active internet connection is required. It is not possible to run this application without an internet connection.

1.2 Starting the application

Please download the applications JAR file and start the application by typing

```
java -jar -Xms128m -Xmx1024m CGIFtGen.jar --input <Input file> --organism
    <Organism>
```

[1] http://www.oracle.com/technetwork/java/javase/downloads/index.html
[2] http://dir.nhlbi.nih.gov/papers/lmi/epigenomes/hgtcell.aspx - please download all "Tag coordinate bed files"
[3] http://hgdownload.cse.ucsc.edu/

on your command prompt. In this example, a minimum of 128 MB and a maximum of 1024 MB of memory will be available for the program. In most cases, CpG island feature generator needs more than 128 MB memory, so it might be convenient to create a shortcut and start the application with as much memory as available. If you have 2 GB RAM, for example, you might want to start the application with the following command:

```
java -Xms128m -Xmx1400M -jar CGIFtGen.jar --input <Input file> --organism
    <Organism>
```

How much memory you actually need strongly depends on your input dataset and the features you are generating. If you are using Histone modification data, for example, CpG island feature generator will need at least 1 GB memory just to read the input data.

It is strongly recommended to start the program with at least 512 MB memory (`-Xmx512M`).

The application offers the following command-line arguments to process your input data and generate features:

input Path and name of your input file. The file format is automatically detected. In doubt, please use either a BED-formatted file[4] or a tab-separated text file with four columns: chromosome, start, end and methylation value.

organism Specify the organism. One of `human` (*Homo sapiens*), `rat` (*Rattus norvegicus*) or `mouse` (*Mus musculus*).

genomerelease Optional parameter, only available for *Homo sapiens*. Settings this parameter to any other than HG18, will result in automatically lifting all genome coordinates to HG18. One of `HG16`, `HG17`, `HG18`, `HG19`.

nomap Optional parameter that requires no value. If this parameter is specified, the input dataset will not be mapped to CpG islands. The application will treat each location in the input file as CpG island.

flanking Optional parameter. By default, each CpG island is extended by $\pm$500 bp to catch flanking sequence effects. This default behaviour may be changed by specifying this parameter, together with the desired flanking sequence width.

features Optional parameter that allows you to set the categories for generating features. If the parameter is not set, default values will be used. Else, please supply feature numbers as given in Section 3.1. A feature number of 0 will generate features for all categories. For example, `--features 3,5,6` will generate features for genomic attributes, SNPs, and closest CpGs.

[4] `http://genome.ucsc.edu/FAQ/FAQformat.html`

svr Optional parameter that requires no value. If this option is set, DNA methylation will be quantified and the result feature file will be support vector regression compatible. If this option is not set, CpG islands will be classified as methylated or unmethylated.

debug Optional parameter that requires no value. This will show more messages during the execution of the application.

For example, running the application on a human DNA methylation dataset in BED-format, called `input.BED` would result in the following command:

```
java -Xms128m -Xmx2G -jar CGIFtGen.jar --input input.BED --organism human
```

2 Step 1: Read existing DNA methylation data

Since CpG island feature generator generates features for existing datasets, you first have to specify your DNA methylation dataset. There exist a lot of data formats for this purpose. For example, the BED-format[1] (used, e.g., by the ENCODE project) or the PAIR format. The last one is usually a probe based data format, which has to be mapped to CpG islands with mean probe intensities. Furthermore, the Human Epigenome Project[2] has an own format and many other groups use their self-created files (e.g., excel) to store their experimental results. We implemented parsers for all these data formats. To account for self-created files, we put some effort in creating a custom datafile parser, which autodetects, e.g., the column separator and if the file contains a location column or separate columns for chromosome, start and end.

To address different genome releases and data files, containing general DNA methylation information but not CpG island specific data (e.g., probe intensities), we added several features to the application. We included a liftOver tool that automatically lifts genome coordinates to the required human genome release to use this application. Moreover, we implemented a mapper that maps each given methylation intensity to a CpG island. With these tools in hand, almost any file format can be used with this application.

2.1 Processing input datasets

The application will proceed with processing your data as follows:

1. Reading your input file and mapping to an internal data structure.

2. Lifting to another genome release, if necessary.

3. Downloading and mapping to UCSC CpG islands (can be deactivated by specifying "`--nomap`").

4. Reading DNA sequences for CpG islands from Ensembl.

The last step is the most time-consuming step, but absolutely necessary to proceed with step 2.

[1] `http://genome.ucsc.edu/FAQ/FAQformat.html`
[2] `http://www.epigenome.org/`

2.1.1 Using BED-files

The application required BED-files in the format, described at `http://genome.ucsc.edu/FAQ/FAQformat#format1`. Required columns are 1 to 3 and 5. The first column should contain chromosomal information, always starting with "chr" (e.g., "chr1"). The second column should contain the start location, the third the end location and the fifth a score from 0 to 1000, whereas 0 denotes fully unmethylated and 1000 fully methylated.

3 Step 2: Generate features, representing existing DNA methylation data

This step is for generating various numeric values, characterizing each CpG island. If all features are selected, 948 values per CpG island are generated. You can choose which features should be generated by using the command-line option "`-features`". Some features require additional data files, which have to be downloaded from various sources:

1. The "Histone modification data" feature is tailored for a dataset from Barski *et al.* (2007), which can be obtained from `http://dir.nhlbi.nih.gov/papers/lmi/epigenomes/hgtcell.aspx`. On this homepage, please download all "Tag coordinate bed files" and place them in the directory "res/HistoneMeth/", relative to the applications executable (JAR file). Please note that you should start the application with at least 1 GB memory to use this feature.

2. The "Evolutionary conservation (PhastCons)" feature is tailored for PhastCon files from the UCSC downloads section at `http://hgdownload.cse.ucsc.edu/`. For example, human PhastCon files are obtainable from `http://hgdownload.cse.ucsc.edu/goldenPath/hg18/phastCons44way/vertebrate/`. These files have been generated by Siepel *et al.* (2005) . please download all files, ending with ".wigFix.gz" and place them in the directory "res/Phastcons/", relative to the applications executable (JAR file). Please note that you should start the application with at least 1.5 GB memory to use this feature.

3.1 Available feature classes

The following descriptions are an overview of all available feature classes. It is recommended to download the histone modification data as described in number 1. In doubt, the default selection of features is a good choice and should be kept (simply omit the "`-features`" option). Most features are calculated from data that is retrieved online from the Ensembl database via an included MySQL adapter. Thus, if the Ensembl database is offline for maintenance, some features can not be calculated. Figure 3.1 gives hints, which features are correlated to DNA methylation and thus, should be included into the prediction. This figure is taken from a manuscript that is currently prepared for publication. The documentation will be updated with a link to the publication for more information on different feature classes and their correlation to DNA methylation, as soon as the manuscript is published.

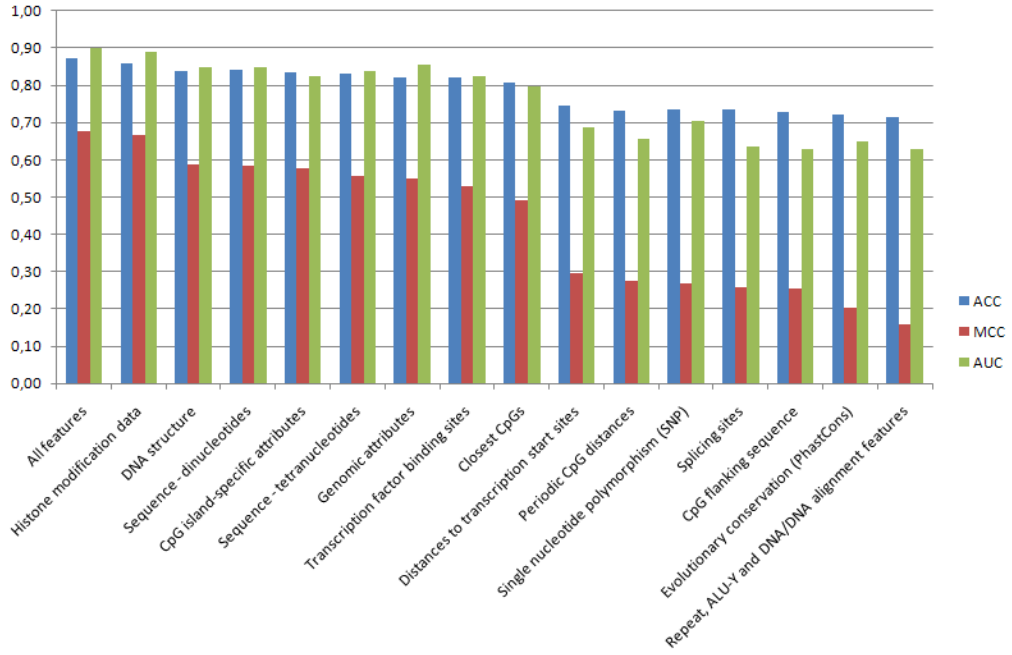

Figure 3.1: Different feature classes and their correlation to DNA methylation.

1. **Distances to transcription start sites (4 features)** This option adds features that represent the distance to the closest gene and closest protein coding gene, based on the Ensembl database. It is known from several studies that the region around a TSS is unmethylated (see Zhang *et al.* (2009); Eckhardt *et al.* (2006)).

2. **CpG island-specific attributes (7 features)** CG content, CG ratio, CG observed/ expected ratio (see Gardiner-Garden and Frommer (1987)), CG/TG ratio (with and without the reverse strand), AT/CG ratio and a boolean flag, if the CpG island is in a coding region.

3. **Genomic attributes (11 features)** Percentage of repetitive base pairs (CpG island length / total length of all self-alignments), number of genes overlapping with the CpG island, total length of all overlapping genes, number of exons overlapping with the CpG island, total and average length of all overlapping exons, number of transcripts for all overlapping genes and number of transcripts divided by number of overlapping genes. For completeness: CpG island length, percentage of CpGs in the whole CpG island and average distance between CpGs are features for this category. All data comes from Ensembl.

4. **Repeat, Alu-Y and DNA/DNA alignment features (19 features)** DNA/DNA self alignments in the CpG island region, various features covering repetitive elements in the CpG island region (e.g., total number of repeats, length of repeats) and multiple features analyzing the

Alu-Y repeat. This special *AGCT*-repeat has been found to occur often in methylated CpG islands (see Fang *et al.* (2006); Kochanek *et al.* (1993)). All features are calculated for three different window sizes: $\pm$900 bp, $\pm$400 bp and exactly covering the CpG island. The Ensembl database is used to calculate the feature values.

5. **Single nucleotide polymorphism (8 features)** Number or known single nucleotide polymorphisms (SNPs) in the CpG island and the distance to the closest SNP from the center of the CpG island are retrieved from the Ensembl database and added as features. In addition, using the same formulas, two special features for the T/C SNPs are added because of its special role in bisulfite sequencing (see Hajkova *et al.* (2002); Bock and Lengauer (2008)).

6. **Periodic CpG distances (15 features)** Based on the hypothesis that CpGs at distances of eight to ten base pairs, relative to other CpGs, are more likely methylated than others (see Jia *et al.* (2007); Zhang *et al.* (2009)), this feature class will add distance scores for multiples of nine from 9 to 45 bps in both directions of a CpG. While studying the flanking sequences of methylated CpGs, we realized a significant difference in CpG occurrence at a distance of 48 bps (CpGs occur almost twice as frequently as on other positions). For this reason, features representing the CpG occurrence at a distance of 48 bp on both strands, are included in addition to the multiples of nine. All values are averaged for all CpGs in the CpG island and three additional features are added, which represent sums of the multiples of nine, of the two 48 bps features and a sum of these two sums.

7. **Closest CpGs (6 features)** These features represent the distances to the three closest CpGs for all CpGs in a CpG island. The three smallest and the average of all values are added as features. This feature category is mainly an extension for the "Periodic CpG distances" category.

8. **Sequence - dinucleotides (16 features)** The occurrences of all possible 16 dinucleotides in the CpG island sequence is counted, divided by the CpG island length and added as features. Correlation between DNA methylation and sequence in general has been proposed by Bock *et al.* (2006).

9. **Sequence - tetranucleotides (257 features)** The occurrences of all possible 256 tetranucleotides in the CpG island sequence is counted, divided by the CpG island length and added as features. This also covers the four bp long Alu repeat (for which an additional feature is included that represents the total (not averaged) count of Alu repeats).

10. **CpG flanking sequence (4 features)** The flanking sequences $\pm$4 bp and $\pm$20 bp for all CpGs in the current dataset are collected and separated into methylated and unmethylated instances. Afterwards, position frequency matrices (PFMs) (see Stormo (2000)) for all these sequences are calculated as follows: Calculate a PFM for all methylated instances and a PFM for all unmethylated instances. Divide the PFM of methylated instances (by dividing the frequency of each nucleotide in each position) by the PFM of unmethylated instances. These PFMs

are then applied to the CpG island sequences and a weight score, covering the quality of the match and the significance, based on the frequency of the actual sequence in the whole genome (see Aerts *et al.* (2003)), is used as a feature. Data and classes from ModuleMaster (see Wrzodek *et al.* (2010)) are used to apply the PFMs to the sequences and calculate the weight scores. This feature category has been added because several authors have found flanking sequence preferences for DNA methyltransferases (for more information on CpG flanking sequences, see Vikas and Albert (2005); Kim *et al.* (2008); Zhang *et al.* (2009)). This category is for human datasets only!

11. Splice sites (5 features) All four PFMs, generated from SpliceDB (Burset *et al.* (2001)) are integrated into CpG island feature generator and are used to identify splice sites. The four PFMs correspond to mammalian frequency matrices of splice sites for GT-AG and GC-AG pairs for donor and acceptor sites respectively. These four features are integrated as weight scores (as described in *CpG flanking sequence*). Additionally, the number of hits from all PFMs is added as fifth feature. This category is for human datasets only!

12. Transcription factor binding sites (457 features) Correlation between DNA methylation and transcription factor binding sites (TFBSs) has already been reported by several groups (see, for example, Fang *et al.* (2006); Bock *et al.* (2006); Das *et al.* (2006); Eckhardt *et al.* (2006)).

Data and methods from ModuleMaster are applied to calculate weight scores for transcription factors as described in *CpG flanking sequence*. The transcription factors have been selected among a large PFM database, consisting of transcription factor binding information from TRANSFAC professional, NUBIScan and predicted TFBSs. This dataset is described in more detail by Wrzodek *et al.* (see Wrzodek *et al.* (2010)). We took all CpG island sequences from the NAME21 dataset (see Zhang *et al.* (2009)) and performed a matrix scan with all PFMs on those. All PFMs which had a weight score below one were removed, because of lack of significance (good matches should have weight scores of at least five. Smaller scores indicate that either the putative TFBS is not well recognized by the PFM or that the putative TFBS occurs very often by chance throughout the human genome). This resulted in a total of 456 PFMs. All those PFMs are encrypted and included in this application. In addition to these 456 features, the logarithm of the sum of all TFBSs is included as feature.

13. DNA structure (43 features) Data from Gardiner *et al.*, who measured and published DNA energies of octamers, is used to calculate sequence dependent DNA structure energies (7^{th} order hidden Markov models – see Gardiner *et al.* (2003)). A strong correlation of these energies and DNA methylation has been reported by Bock *et al.* (2006). This category is for human datasets only!

14. Evolutionary conservation (PhastCons) (4 features) Data from the UCSC Genome Browser (see Rhead *et al.* (2010)) is used to calculate several features, representing the evolutionary

conservation of the CpG island. This feature class requires at least 1.5 GB memory and pre-downloaded PhastCon files as described in number 2 on page 6.

15. **Histone modification data (92 features)** The correlation between histone modification and DNA methylation has already been reported by several authors (for example Ting *et al.* (2006); Bird (2002); Cedar and Bergman (2009); Ooi *et al.* (2007); Fan *et al.* (2008)). Barski *et al.* (2007) generated 23 genome-wide datasets, covering 20 different histone modification variants (H3K4me1, H3K4me2, H3K4me3, H3K9me1, H3K9me2, H3K9me3, H3K27me1, H3K27me2, H3K27me3, H3K36me1, H3K36me3, H3K79me1, H3K79me2, H3K79me3, H3R2me1, H3R2me2, H4K20me1, H4K20me3, H4R3me2, H2BK5me1) and the distribution of H2A.Z, RNA polymerase II, and the insulator binding protein CTCF. This data is mapped onto each CpG island and four numerical features for each histone modification dataset are generated per CpG island. It has been shown that, e.g., H3K4me prevents DNA methyltransferase enzymes from *de novo* methylating CpG islands (see Cedar and Bergman (2009)). Thus, some histone modifications dictate DNA methylation in the embryo and taking one dataset for any cell types is feasible. This feature class requires at least 1 GB memory and pre-downloaded files as described in number 1 on page 6. This category is for human datasets only!

In addition to the feature classes, the user may choose to generate binary data or quantify DNA methylation. If DNA methylation is quantified, a support vector regression may be performed. Therefore, the predicted results also have a quantified DNA methylation value but may fail to separate methylated from unmethylated CpG islands. Else, the application will treat each CpG island with a methylation value higher than 60 % as methylated and CpG islands with methylation values lower or equal than 40 % as unmethylated. Differentially methylated CpG islands (methylation values between those thresholds) will be skipped.

The generated features will be saved to a file called "lastFeatures.txt". The coordinates of all CpG islands extracted in step 1 will be saved to "lastCoords.txt". Each line in the coordinates file corresponds to the same line number in the feature file.

4 FAQ / Troubleshooting

Warning messages, saying that something has been deleted in liftOver target, appear.
This message tells you that something has been deleted or changed in the new genome build. It occurs, if you read an input datasets and specify a genome release, other than hg18. It means, that the old coordinates could not be mapped to the new genome release. Usually you can simply ignore this error.

"Could not connect to the Ensembl database."
This is a critical error. The application needs to connect to the Ensembl MySQL Server for retrieving DNA sequences in step 1 and for generating various features in step 2. These steps can not be performed without a connection to the Ensembl database. Please check your internet connection to resolve this issue. If you have an active internet connection, the Ensembl database is probably offline for maintenance – just try again later.

Warning messages, saying that methylation state of a CpG island could not be determined.
This warning is issued if a dataset, consisting of several amplicons from several cell types, has been loaded and applied to only one cell type. In this case, it may happen that an amplicon for a specific CpG island is missing in a cell type and this warning is issued. This happens, e.g., if using this application to predict CpG island methylation for the NAME21 data (see Zhang *et al.* (2009)).

"CpG island is differentially methylated [...]".
This warning is issued for CpG islands that are between the defined "is methylated" and "is not methylated" threshold (see Section 3.1).

I'm getting a "java.lang.OutOfMemoryError: Java heap space"
Some operations need a lot of memory. If you simply start CpG island feature generator, without any JVM parameters, only 64 MB of memory are available. Please append the argument `-Xmx1024M` to start the application with 1 GB of main memory. See Section 1.2 for a more detailed description of how to start the application with additional memory. If possible, you should give the application 2 GB of memory. A minimum of 1 GB memory should be available to the application.

Is an internet connection required to run CpG island feature generator?
An internet connection is required for step1 and for certain feature classes in step 2.

Which organisms are supported? Currently mouse, human and rat are supported.

Bibliography

Aerts, S., Loo, P. V., Thijs, G., Moreau, Y., and Moor, B. D. (2003). Computational detection of cis -regulatory modules. *Bioinformatics*, **19 Suppl 2**, ii5–i14.

Barski, A., Cuddapah, S., Cui, K., Roh, T.-Y., Schones, D. E., Wang, Z., Wei, G., Chepelev, I., and Zhao, K. (2007). High-resolution profiling of histone methylations in the human genome. *Cell*, **129**(4), 823–837.

Bird, A. (2002). Dna methylation patterns and epigenetic memory. *Genes Dev*, **16**(1), 6–21. Methylation of CpG islands silences the corresponding gene.

Bock, C. and Lengauer, T. (2008). Computational epigenetics. *Bioinformatics*, **24**(1), 1–10.

Bock, C., Paulsen, M., Tierling, S., Mikeska, T., Lengauer, T., and Walter, J. (2006). Cpg island methylation in human lymphocytes is highly correlated with dna sequence, repeats, and predicted dna structure. *PLoS Genet*, **2**(3), e26+.

Burset, M., Seledtsov, I. A., and Solovyev, V. V. (2001). Splicedb: database of canonical and non-canonical mammalian splice sites. *Nucleic Acids Res*, **29**(1), 255–259.

Cedar, H. and Bergman, Y. (2009). Linking dna methylation and histone modification: patterns and paradigms. *Nat Rev Genet*, **10**(5), 295–304.

Das, R., Dimitrova, N., Xuan, Z., Rollins, R. A., Haghighi, F., Edwards, J. R., Ju, J., Bestor, T. H., and Zhang, M. Q. (2006). Computational prediction of methylation status in human genomic sequences. *Proc Natl Acad Sci U S A*, **103**(28), 10713–10716.

Eckhardt, F., Lewin, J., Cortese, R., Rakyan, V. K., Attwood, J., Burger, M., Burton, J., Cox, T. V., Davies, R., Down, T. A., Haefliger, C., Horton, R., Howe, K., Jackson, D. K., Kunde, J., Koenig, C., Liddle, J., Niblett, D., Otto, T., Pettett, R., Seemann, S., Thompson, C., West, T., Rogers, J., Olek, A., Berlin, K., and Beck, S. (2006). DNA methylation profiling of human chromosomes 6, 20 and 22. *Nat Genet*, **38**(12), 1378–1385. 10.1038/ng1909.

Fan, S., Zhang, M. Q., and Zhang, X. (2008). Histone methylation marks play important roles in predicting the methylation status of cpg islands. *Biochemical and Biophysical Research Communications*, **374**(3), 559 – 564.

Fang, F., Fan, S., Zhang, X., and Zhang, M. Q. (2006). Predicting methylation status of CpG islands in the human brain. *Bioinformatics*, **22**(18), 2204–2209.

Gardiner, E. J., Hunter, C. A., Packer, M. J., Palmer, D. S., and Willett, P. (2003). Sequence-dependent dna structure: a database of octamer structural parameters. *J Mol Biol*, **332**(5), 1025–1035.

Gardiner-Garden, M. and Frommer, M. (1987). CpG islands in vertebrate genomes. *Journal of Molecular Biology*, **196**(2), 261–282. Traditionelle CpG island definition. Formeln wie obs/exp.

Hajkova, P., el Maarri, O., Engemann, S., Oswald, J., Olek, A., and Walter, J. (2002). Dna-methylation analysis by the bisulfite-assisted genomic sequencing method. *Methods Mol Biol*, **200**, 143–154.

Jia, D., Jurkowska, R. Z., Zhang, X., Jeltsch, A., and Cheng, X. (2007). Structure of dnmt3a bound to Dnmt3L suggests a model for de novo DNA methylation. *Nature*, **449**(7159), 248–251. Correlation between periodic distribution of CpG and DNA Methylation.

Kim, S., Li, M., Paik, H., Nephew, K., Shi, H., Kramer, R., Xu, D., and Huang, T. (2008). Predicting dna methylation susceptibility using cpg flanking sequences. *Pacific Symposium on Biocomputing*, pages 315–326.

Kochanek, S., Renz, D., and Doerfler, W. (1993). Dna methylation in the alu sequences of diploid and haploid primary human cells. *EMBO J*, **12**(3), 1141–1151. ALU sequences.

Ooi, S. K. T., Qiu, C., Bernstein, E., Li, K., Jia, D., Yang, Z., Erdjument-Bromage, H., Tempst, P., Lin, S.-P., Allis, C. D., Cheng, X., and Bestor, T. H. (2007). Dnmt3l connects unmethylated lysine 4 of histone h3 to de novo methylation of dna. *Nature*, **448**(7154), 714–717.

Rhead, B., Karolchik, D., Kuhn, R. M., Hinrichs, A. S., Zweig, A. S., Fujita, P. A., Diekhans, M., Smith, K. E., Rosenbloom, K. R., Raney, B. J., Pohl, A., Pheasant, M., Meyer, L. R., Learned, K., Hsu, F., Hillman-Jackson, J., Harte, R. A., Giardine, B., Dreszer, T. R., Clawson, H., Barber, G. P., Haussler, D., and Kent, W. J. (2010). The ucsc genome browser database: update 2010. *Nucleic Acids Res*, **38**(Database issue), D613–D619.

Siepel, A., Bejerano, G., Pedersen, J. S., Hinrichs, A. S., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L. W., Richards, S., Weinstock, G. M., Wilson, R. K., Gibbs, R. A., Kent, W. J., Miller, W., and Haussler, D. (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res*, **15**(8), 1034–1050.

Stormo, G. D. (2000). Dna binding sites: representation and discovery. *Bioinformatics*, **16**(1), 16–23.

Ting, A. H., McGarvey, K. M., and Baylin, S. B. (2006). The cancer epigenome–components and functional correlates. *Genes Dev*, **20**(23), 3215–3231.

Vikas, H. and Albert, J. (2005). Profound flanking sequence preference of dnmt3a and dnmt3b mammalian dna methyltransferases shape the human epigenome. *Journal of Molecular Biology*, **348**(5), 1103–1112.

Wrzodek, C., Schröder, A., Dräger, A., Wanke, D., Berendzen, K. W., Kronfeld, M., Harter, K., and Zell, A. (2010). ModuleMaster: A new tool to decipher transcriptional regulatory networks. *Biosystems*, **99**(1), 79–81.

Zhang, Y., Rohde, C., Tierling, S., Jurkowski, T. P., Bock, C., Santacruz, D., Ragozin, S., Reinhardt, R., Groth, M., Walter, J., and Jeltsch, A. (2009). Dna methylation analysis of chromosome 21 gene promoters at single base pair and single allele resolution. *PLoS Genet*, **5**(3), e1000438.