Hindawi Publishing Corporation
BioMed Research International
Volume 2013, Article ID 709042, 6 pages
http://dx.doi.org/10.1155/2013/709042

Research Article

CpGislandEVO: A Database and Genome Browser for Comparative Evolutionary Genomics of CpG Islands

Guillermo Barturen,[1,2] Stefanie Geisen,[1,2] Francisco Dios,[1,2] E. J. Maarten Hamberg,[1,2] Michael Hackenberg,[1,2] and José L. Oliver[1,2]

[1] *Departamento de Genética, Facultad de Ciencias, Universidad de Granada, 18071 Granada, Spain*
[2] *Laboratório de Bioinformática, Instituto de Biotecnología, Centro de Investigación Biomédica, 18100 Granada, Spain*

Correspondence should be addressed to José L. Oliver; oliver@ugr.es

Received 20 April 2013; Revised 12 July 2013; Accepted 19 August 2013

Academic Editor: Stephan Koblmüller

Hypomethylated, CpG-rich DNA segments (CpG islands, CGIs) are epigenome markers involved in key biological processes. Aberrant methylation is implicated in the appearance of several disorders as cancer, immunodeficiency, or centromere instability. Furthermore, methylation differences at promoter regions between human and chimpanzee strongly associate with genes involved in neurological/psychological disorders and cancers. Therefore, the evolutionary comparative analyses of CGIs can provide insights on the functional role of these epigenome markers in both health and disease. Given the lack of specific tools, we developed *CpGislandEVO*. Briefly, we first compile a database of statistically significant CGIs for the best assembled mammalian genome sequences available to date. Second, by means of a coupled browser front-end, we focus on the CGIs overlapping orthologous genes extracted from *OrthoDB*, thus ensuring the comparison between CGIs located on truly homologous genome segments. This allows comparing the main compositional features between homologous CGIs. Finally, to facilitate nucleotide comparisons, we lifted genome coordinates between assemblies from different species, which enables the analysis of sequence divergence by direct count of nucleotide substitutions and indels occurring between homologous CGIs. The resulting *CpGislandEVO* database, linking together CGIs and single-cytosine DNA methylation data from several mammalian species, is freely available at our website.

1. Introduction

Short stretches of CpG dinucleotides (CpG islands or CGIs) predominantly hypomethylated in healthy tissues [1, 2] are key epigenomic markers in mammalian genomes [3]. Almost all housekeeping genes and a half of the tissue-specific genes are associated to CGIs [4]. DNA methylation plays an important role in the origin as well as in the function of CGIs. Aberrant methylation (mostly hypermethylation) of CGIs can lead to several syndromes, such as cancer [5–10]. Moreover, although it has been shown that certain human diseases may have evolutionary epigenetic origins [11, 12], it remains largely unknown how patterns of DNA methylation differ between closely related species and whether such differences contribute to species-specific phenotypes [11]. Some methylation databases [13–15] and CGI databases [16] have been developed, but, to our knowledge, no existing genome browser addresses specifically the evolutionary relationships between the CGIs from different species. To help describing and understanding the function as well as the mechanisms generating and maintaining CGIs within an evolutionary context, we develop here *CpGislandEVO* (http://bioinfo2.ugr.es/CpGislandEVO/index.php). The database, coupled to a powerful genome browser, links together experimental and predicted CGIs, as well as single-cytosine-resolution DNA methylation data from different mammalian species.

Early analyses of CGI evolution were based on compositional comparisons between islands from different species but located on homologous gene contexts [17, 18]. Recently, the rapidly increasing number of sequenced genomes enabled evolutionary studies relying on multiple-sequence alignments [19]. Here, we combine both approaches to envisage accurate sequence comparisons between CGIs located on homologous gene contexts.

2. Material and Methods

2.1. Genome Assemblies. Updated chromosome sequences for the best assembled mammalian genomes (*Homo sapiens* (hg19), *Pan troglodytes* (panTro3), *Gorilla gorilla* (gorGor3), *Pongo abelii* (ponAbe2), *Macaca mulatta* (rheMac2), *Mus musculus* (mm10), and *Rattus norvegicus* (rn5)) were downloaded from the UCSC genome browser [20].

2.2. CGI Predictions. CGIs were predicted by means of an improved version [21] of the *CpGcluster* algorithm [22]. We used the genome intersection as distance threshold to define the clusters of CpG dinucleotides and a *P* value threshold of 1E-5 for the statistical significance. For comparison, the database also includes the CGI predictions for hg19 made by a window-based program [23], as well as the UCSC island track for the different species [20].

2.3. Experimental CGIs. Experimental CGI datasets include the 13,277 nonoverlapping promoter regions which are unmethylated in at least one of the two tissues (fibroblast and sperm) analyzed by Weber et al. [24] and the 17,383 CpG-islands experimentally detected in blood cells [25].

2.4. Orthologous Gene-Contexts. To ensure that we are comparing truly homologous genome segments, we focus on the CGIs located around orthologous genes extracted from *OrthoDB* [26]. The *OrthoDB* implementation employs a best-reciprocal-hit clustering algorithm based on all-against-all Smith-Waterman [27] protein sequence comparisons. In particular, we take into account all the islands within the gene-body of each of the *OrthoDB* genes. We defined the gene-body as the region extending from 500 bp upstream from the transcription start site (txStart) to 500 bp downstream the transcription stop site (txEnd).

2.5. Sequence Comparisons. Base level comparisons of homologous CpG-island sequences from different species were carried out by lifting genome coordinates between assemblies by means of the Galaxy implementation [28, 29] of the *LiftOver* utility, based on the *Chain* and *Net* tracks from the UCSC Genome Browser (http://genome.ucsc.edu/cgi-bin/hgLiftOver/). Default parameters were used, except for the "minimum matching region size for the query" which was set to 8 bp, which corresponds to the smallest CGI length.

2.6. Methylation Data. Since the lack of CGI methylation is a very good indicator of function [30], we linked *CpGislandEVO* to a relevant subset of *NGSmethDB* [31, 32], where a wide variety of single-cytosine-resolution methylation methylome maps from different tissues and species are available. Methylomes were obtained with *NGSmethPipe* [33] (http://bioinfo2.ugr.es/NGSmethPipe/) and *MethylExtract* [34] (http:// bioinfo2.ugr.es/MethylExtract/), two tools implementing stringent quality controls to minimize important error sources, as for example sequencing errors, bisulfite failures, clonal reads, or single nucleotide variants. Likewise important, the use of a single bioinformatics protocol homogenizes the database content making the different methylomes comparable among each other even if they are from different studies.

2.7. Database and Genome-Browser Implementation. The orthologous genes were taken from *OrthoDB* [26], which does not provide information about gene names or coordinates. Therefore, this information is obtained from *ensGene* (using *ensemblIDs* as identifiers) and *refGene* (using gene names) from UCSC [20]. The curated online repository of HGNC-approved gene nomenclature [35] is then used to link names between *ensGene* and *refGene* databases.

We implemented an autocomplete function to help the user locate human genes in *OrthoDB* [26] with at least one orthologous gene in any of the other species (via its gene name or its *UniProtID*). Once chosen a gene name, the *refGene* database (214,898 entries) and, if no results are found, the *ensGene* database (647,600 entries) are searched for this gene. As an output, the chromosome and the start and stop positions of this gene are obtained; a final check for at least one CGI within this gene-body is performed. Then, *OrthoDB* is queried with the *ensemblID* of the human gene, and a table with the *ensemblIDs* of the orthologous genes found in any of the six animal species is generated (Table 1). This table also contains known gene names obtained by converting back *ensemblIDs* via *biomart* databases [36].

When the user introduces a chromosome and an approximate coordinate, the script searches the *ensGene* table for the closest upstream and/or downstream gene with at least one orthologous gene in *orthoDB*, then returning the exact chromosome position and gene name. The corresponding *ensemblID* is then used as above to generate Table 1.

The most recent version (currently 1.9.7) of the cross-platform genome browser *JBrowse* [37, 38] is used to display genes, CpGislands, LiftOver-mapped tracks, and methylation tracks for the hg19 assembly and to compare it to the other six mammalian species. A pair-wise comparison is performed by means of two frames within a window: the top one is always used to display hg19 tracks and the bottom one for one of the six animal species. Note that, by the moment, both frames are not synchronized. This feature will be implemented as soon as Jbrowse_syn is available (http://gmod.org/w/images/a/aa/ISyIPGMODforComparativeGenomics.pdf, slide 15).

Currently, *CpGislandEVO* includes the mammalian genomes with comparable genome-wide methylation data (human, chimpanzee, rhesus monkey, and mouse). In this way, the platform allows the user to compare CGIs from these mammalian species. The number of species and methylation datasets will be increased according to the advent of new comparable genome-wide methylation datasets.

2.8. Data Download and Script Availability. The datasets in *CpGislandEVO* can be freely downloaded by the user from NGSmethDB (http://bioinfo2.ugr.es/NGSmethDB/database.php) in a wide variety of standard data formats: BED, GFF3, bedGraph, Wiggle, and so forth. The Perl script for the most recent version of *CpGcluster* is also freely available to download from the group webserver (http://bioinfo2.ugr.es/CpGcluster/CpGcluster.zip).

Table 1: List of orthologous genes to the human query gene KDM1A for the lysine-specific histone demethylase 1A, offering links to Ensembl and UCSC genome browsers, as well as to a specific island viewer based on *JBrowse*.

(a) Query gene

Species	Gene name	Link Ensembl	Link UCSC	Link JBrowseViewer
Homo sapiens	KDM1A	hg_KDM1A_ensembl	hg_KDM1A_ucsc	hg_KDM1A_evoDB

(b) Orthologous genes

Species	Gene name (*EnsemblID*)	Link Ensembl	Link UCSC	Link JBrowseViewer
Gorilla gorilla	KDM1A (ENSGGOG 00000003664)	gorgor_KDM1A_ensembl	gorgor_KDM1A_ucsc	gorgor_KDM1A_evoDB
Macaca mulatta	KDM1A (ENSMMUG 00000009773)	rhemac_KDM1A_ensembl	rhemac_KDM1A_ucsc	rhemac_KDM1A_evoDB
Mus musculus	Kdm1a (ENSMUSG 00000036940)	mm_kdm1a_ensembl	mm_kdm1a_ucsc	mm_kdm1a_evoDB
Pan troglodytes	KDM1A (ENSPTRG 00000000321)	pantro_KDM1A_ensembl	pantro_KDM1A_ucsc	pantro_KDM1A_evoDB
Pongo abelii	KDM1A (ENSPPYG 00000001747)	ponabe_KDM1A_ensembl	ponabe_KDM1A_ucsc	ponabe_KDM1A_evoDB
Rattus norvegicus	Kdm1 (ENSRNOG 00000022372)	rn_kdm1a_ensembl	rn_kdm1a_ucsc	rn_kdm1a_evoDB

3. Results and Discussion

We first compiled a CGI database (http://bioinfo2.ugr.es/CpGislandEVO/launch.php) for the best assembled mammalian genomes using the *CpGcluster* algorithm [22] with the genome intersection as distance threshold [21, 22, 39]. This setup is especially appropriate for interspecies comparative genomic studies as (i) the distance threshold is directly obtained from the genome sequence and (ii) a *P* value is assigned to each CGI. Therefore, exactly the same criterions are used in all species to detect CpG islands. This is not possible when using algorithms based on sliding windows to predict CGIs, as variations in genome G+C content, O/E ratio, or CpG density cannot be easily taken into account to guarantee an unbiased detection [39]. Second, by means of a specifically designed genome browser based on *JBrowse* [37, 38], we focus on those CGIs located within orthologous gene-contexts [26], thus ensuring that we are comparing CGIs from true homologous sequence segments. Finally, to study sequence divergence at base level between homologous CpG islands, we lifted genome coordinates between assemblies from different species by using the *LiftOver* utility [20].

The *CpGislandEVO* platform first offers the possibility to explore the CGI database obtained with the *CpGcluster* algorithm [21, 22, 39]. After selecting genome assembly and chromosome, the server offers links to (i) *CpGcluster* predictions, (ii) UCSC genome browser [20], and (iii) single-cytosine methylation data by means of a subset (http://bioinfo2.ugr.es/CpGislandEVO/methylation.php) of *NGSmethDB* [31, 32]. Summary statistics for the CGI database and CGI distribution in the orthologous gene bodies of the different species are shown on-line: http://bioinfo2.ugr.es/CpGislandEVO/statistics.php. Second, a coupled genome browser allows sequence comparisons between CGIs located on homologous segments from different species. The user can navigate the database in two ways: (1) by directly introducing a human gene/protein reference name or (2) by providing a chromosome and an approximate coordinate (and then the closest upstream and/or downstream human gene with at least one orthologous gene will be shown). The server first returns the orthologous genes (Table 1) for the query gene with links to *Ensembl* [40] and UCSC [20] genome browsers, as well as to a specific island viewer we have developed on the basis of the *JBrowse* next-generation browser [37, 38]. The *CpGislandEVO* viewer allows the comparative genomics of CGIs in different species.

As an example, we focus on the human query gene KDM1A for the lysine-specific histone demethylase 1A. Figure 1 shows the promoter region of this gene and the CGIs and methylation data for PBMC cells. The homologous CGIs from six other species are shown for comparison. The small methylated human CGI is conserved in the three primate species, while the larger unmethylated human CGI is conserved even in the mouse. On the other hand, Figure 2 uses two frames within the same window, to compare the CGIs in the promoter region of the gene KDM1A in human and

Figure 1: Promoter region of the human gene KDM1A showing CGIs and methylation data for PBMC cells. The lifted homologous CGIs from six other species are shown for comparison. The small methylated human CGI is conserved in three primate species, while the larger unmethylated human CGI is not only conserved in some primates but also in the mouse.

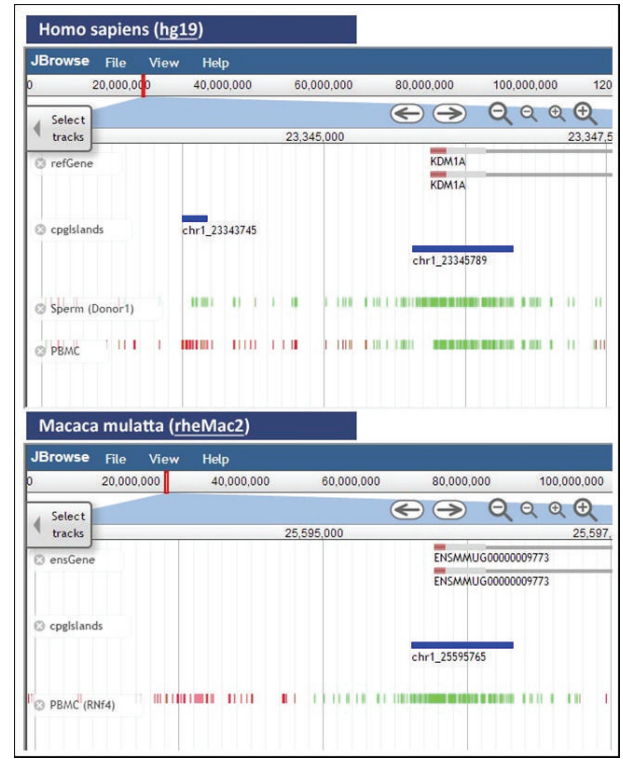

Figure 2: Comparison of the promoter region of the gene KDM1A in human and rhesus monkey using two frames within the same window. The unmethylated CGI is conserved between the two species, while the small human differentially methylated CGI is missing in the rhesus monkey.

rhesus monkey. The unmethylated CGI is conserved between the two species, while the small human differentially methylated CGI is missing in the rhesus monkey. In this way, *CpGislandEVO* put together in the same screen information scattered in diverse sources, or only attainable after running different computer programs, thus allowing evolutionary compositional comparisons as well as accurate sequence analyses between islands from different species, but located on homologous gene contexts.

4. Conclusions

We have compiled a database of statistically significant CGIs for the best assembled mammalian genomes using an improved version of the *CpGcluster* algorithm [21, 22, 39]. Then, by means of a specifically designed genome-browser based on *JBrowse* [37, 38], we focused on those CGIs located within orthologous gene-contexts [26], thus ensuring that we are comparing CGIs from true homologous genome segments. Finally, by lifting genome coordinates between assemblies from different species, the *CpGislandEVO* platform allows the direct comparison at base level between homologous CpG islands. The evolutionary comparative studies of CGIs can provide insights on their functional role in both health and disease, as well as on the evolutionary mechanisms generating and maintaining these important epigenome markers.

Authors' Contribution

Guillermo Barturen and Stefanie Geisen contributed equally to this work.

Acknowledgments

This work was supported by the Spanish Government [BIO2008-01353 to José L. Oliver and BIO2010-20219 to Michael Hackenberg], Basque country "AE" grant (to Guillermo Barturen) and Erasmus internships (to Stefanie Geisen, Francisco Dios, and E. J. Maarten Hamberg). The authors greatly acknowledge the continuous support by Robert Buels, Lead Developer of *JBrowse*.

References

[1] A. Bird, "DNA methylation patterns and epigenetic memory," *Genes and Development*, vol. 16, no. 1, pp. 6–21, 2002.

[2] A. P. Bird, "CpG-rich islands and the function of DNA methylation," *Nature*, vol. 321, no. 6067, pp. 209–213, 1986.

[3] D. Schubeler, "Molecular biology. Epigenetic islands in a genetic ocean," *Science*, vol. 338, no. 6108, pp. 756–757, 2012.

[4] J. Zhu, F. He, S. Hu, and J. Yu, "On the nature of human housekeeping genes," *Trends in Genetics*, vol. 24, no. 10, pp. 481–484, 2008.

[5] S. B. Baylin, M. Esteller, M. R. Rountree, K. E. Bachman, K. Schuebel, and J. G. Herman, "Abberant patterns of DNA methylation, chromatin formation and gene expression in cancer," *Human Molecular Genetics*, vol. 10, no. 7, pp. 687–692, 2001.

[6] C. de Smet, C. Lurquin, B. Lethé, V. Martelange, and T. Boon, "DNA methylation is the primary silencing mechanism for a set of germ line- and tumor-specific genes with a CpG-rich promoter," *Molecular and Cellular Biology*, vol. 19, no. 11, pp. 7327–7335, 1999.

[7] M. Esteller, P. G. Corn, S. B. Baylin, and J. G. Herman, "A gene hypermethylation profile of human cancer," *Cancer Research*, vol. 61, no. 8, pp. 3225–3229, 2001.

[8] J.-P. Issa, "CpG island methylator phenotype in cancer," *Nature Reviews Cancer*, vol. 4, no. 12, pp. 988–993, 2004.

[9] Y. Riazalhosseini and J. D. Hoheisel, "Do we use the appropriate controls for the identification of informative methylation markers for early cancer detection?" *Genome Biology*, vol. 9, no. 11, article 405, 2008.

[10] A. R. Krebs and D. Schubeler, "Tracking the evolution of cancer methylomes," *Nature Genetics*, vol. 44, no. 11, pp. 1173–1174, 2012.

[11] J. Zeng, G. Konopka, B. G. Hunt, T. M. Preuss, D. Geschwind, and S. V. Yi, "Divergent whole-genome methylation maps of human and chimpanzee brains reveal epigenetic basis of human regulatory evolution," *The American Journal of Human Genetics*, vol. 91, no. 3, pp. 455–465, 2012.

[12] C. G. Bell, G. A. Wilson, L. M. Butcher, C. Roos, L. Walter, and S. Beck, "Human-specific CpG, "beacons" identify loci associated with human-specific traits and disease," *Epigenetics*, vol. 7, no. 10, pp. 1188–1199, 2012.

[13] V. Negre and C. Grunau, "The MethDB DAS server: adding an epigenetic information layer to the human genome," *Epigenetics*, vol. 1, no. 2, pp. 101–105, 2006.

[14] J. Shi, J. Hu, Q. Zhou, Y. Du, and C. Jiang, "PEpiD: a prostate epigenetic database in mammals," *PLoS One*, vol. 8, no. 5, Article ID e64289, 2013.

[15] F. Gu, M. S. Doderer, Y.-W. Huang et al., "CMS: a web-based system for visualization and analysis of genome-wide methylation data of human cancers," *PLoS One*, vol. 8, no. 4, Article ID e60980, 2013.

[16] H.-C. Kuo, P.-Y. Lin, T.-C. Chung et al., "DBCAT: database of CpG islands and analytical tools for identifying comprehensive methylation profiles in cancer cells," *Journal of Computational Biology*, vol. 18, no. 8, pp. 1013–1017, 2011.

[17] B. Aissani and G. Bernardi, "CpG islands, genes and isochores in the genomes of vertebrates," *Gene*, vol. 106, no. 2, pp. 185–195, 1991.

[18] K. Jabbari and G. Bernardi, "CpG doublets, CpG islands and Alu repeats in long human DNA sequences from different isochore families," *Gene*, vol. 224, no. 1-2, pp. 123–128, 1998.

[19] N. M. Cohen, E. Kenigsberg, and A. Tanay, "Primate CpG islands are maintained by heterogeneous evolutionary regimes involving minimal selection," *Cell*, vol. 145, no. 5, pp. 773–786, 2011.

[20] D. Karolchik, R. M. Kuhn, R. Baertsch et al., "The UCSC genome browser database: 2008 update," *Nucleic Acids Research*, vol. 36, no. 1, pp. D773–D779, 2008.

[21] M. Hackenberg, P. Carpena, P. Bernaola-Galván, G. Barturen, Á. M. Alganza, and J. L. Oliver, "WordCluster: detecting clusters of DNA words and genomic elements," *Algorithms for Molecular Biology*, vol. 6, no. 1, article 2, 2011.

[22] M. Hackenberg, C. Previti, P. L. Luque-Escamilla, P. Carpena, J. Martínez-Aroza, and J. L. Oliver, "CpGcluster: a distance-based algorithm for CpG-island detection," *BMC Bioinformatics*, vol. 7, article 446, 2006.

[23] D. Takai and P. A. Jones, "The CpG island searcher: a new WWW resource," *In Silico Biology*, vol. 3, no. 3, pp. 235–240, 2003.

[24] M. Weber, I. Hellmann, M. B. Stadler et al., "Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome," *Nature Genetics*, vol. 39, no. 4, pp. 457–466, 2007.

[25] R. Illingworth, A. Kerr, D. Desousa et al., "A novel CpG island set identifies tissue-specific methylation at developmental gene loci," *PLoS Biology*, vol. 6, no. 1, article e22, 2008.

[26] R. M. Waterhouse, E. M. Zdobnov, F. Tegenfeldt, J. Li, and E. V. Kriventseva, "OrthoDB: the hierarchical catalog of eukaryotic orthologs in 2011," *Nucleic Acids Research*, vol. 39, no. 1, pp. D283–D288, 2011.

[27] T. F. Smith and M. S. Waterman, "Identification of common molecular subsequences," *Journal of Molecular Biology*, vol. 147, no. 1, pp. 195–197, 1981.

[28] B. Giardine, C. Riemer, R. C. Hardison et al., "Galaxy: a platform for interactive large-scale genome analysis," *Genome Research*, vol. 15, no. 10, pp. 1451–1455, 2005.

[29] J. Goecks, A. Nekrutenko, and J. Taylor, "Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences," *Genome Biology*, vol. 11, no. 8, article R86, 2010.

[30] R. S. Illingworth and A. P. Bird, "CpG islands—'a rough guide'," *FEBS Letters*, vol. 583, no. 11, pp. 1713–1720, 2009.

[31] M. Hackenberg, G. Barturen, and J. L. Oliver, "NGSmethDB: a database for next-generation sequencing single-cytosine-resolution DNAmethylation data," *Nucleic Acids Research*, vol. 39, no. 1, pp. D75–D79, 2011.

[32] S. Geisen et al., "NGSmethDB: high-quality, single-cytosine resolution methylation maps," submitted to *Nucleic Acids Research*.

[33] M. Hackenberg, G. Barturen, and J. L. Oliver, "Methylation profiling from high-throughput sequencing data," in *DNA Methylation—From Genomics to Technology*, T. Tatarinova and O. Kerton, Eds., p. 27, In-Tech, 2012.

[34] G. Barturen, A. Rueda, J. L. Oliver, and M. Hackenberg, "MethylExtract: high-quality methylation maps and SNV calling from whole genome bisulfite sequencing data," submitted.

[35] K. A. Gray, L. C. Daugherty, S. M. Gordon, R. L. Seal, M. W. Wright, and E. A. Bruford, "Genenames.org: the HGNC resources in 2013," *Nucleic Acids Research*, vol. 41, pp. D545–D552, 2013.

[36] A. Kasprzyk, "BioMart: driving a paradigm change in biological data management," *Database*, vol. 2011, article bar049, 2011.

[37] M. E. Skinner and I. H. Holmes, "Setting up the JBrowse genome browser," *Current Protocols in Bioinformatics*, chapter 9, p. unit 9.13, 2010.

[38] M. E. Skinner, A. V. Uzilov, L. D. Stein, C. J. Mungall, and I. H. Holmes, "JBrowse: a next-generation genome browser," *Genome Research*, vol. 19, no. 9, pp. 1630–1638, 2009.

[39] M. Hackenberg, G. Barturen, P. Carpena, P. L. Luque-Escamilla, C. Previti, and J. L. Oliver, "Prediction of CpG-island function: CpG clustering vs. sliding-window methods," *BMC Genomics*, vol. 11, no. 1, article 327, 2010.

[40] T. J. Hubbard, B. L. Aken, S. Ayling et al., "Ensembl 2009," *Nucleic Acids Research*, vol. 37, database issue, pp. D690–D697, 2009.