Published online 27 October 2016 *Nucleic Acids Research, 2017, Vol. 45, Database issue* ***D97–D103***
doi: 10.1093/nar/gkw996

NGSmethDB 2017: enhanced methylomes and differential methylation

Ricardo Lebrón[1,2,†], Cristina Gómez-Martín[1,2,†], Pedro Carpena[3], Pedro Bernaola-Galván[3], Guillermo Barturen[4], Michael Hackenberg[1,2,*] and José L. Oliver[1,2,*]

[1]Department of Genetics, Faculty of Science, University of Granada, Campus de Fuentenueva s/n, 18071-Granada, Spain, [2]Laboratory of Bioinformatics, Centro de Investigación Biomédica, PTS, Avda. del Conocimiento s/n, 18100-Granada, Spain, [3]Department of Applied Physics II, Universidad de Málaga, 29071 Málaga, Spain and [4]Genetics of Complex Diseases Group, GENyO, Pfizer-University of Granada-Junta de Andalucía Center for Genomics and Oncological Research, 18100-Granada, Spain

Received September 12, 2016; Revised October 08, 2016; Editorial Decision October 13, 2016; Accepted October 14, 2016

ABSTRACT

The 2017 update of *NGSmethDB* stores whole genome methylomes generated from short-read data sets obtained by bisulfite sequencing (WGBS) technology. To generate high-quality methylomes, stringent quality controls were integrated with third-part software, adding also a two-step mapping process to exploit the advantages of the new genome assembly models. The samples were all profiled under constant parameter settings, thus enabling comparative downstream analyses. Besides a significant increase in the number of samples, *NGSmethDB* now includes two additional data-types, which are a valuable resource for the discovery of methylation epigenetic biomarkers: (i) differentially methylated single-cytosines; and (ii) methylation segments (i.e. genome regions of homogeneous methylation). The *NGSmethDB* back-end is now based on *MongoDB*, a NoSQL hierarchical database using JSON-formatted documents and dynamic schemas, thus accelerating sample comparative analyses. Besides conventional database dumps, track hubs were implemented, which improved database access, visualization in genome browsers and comparative analyses to third-part annotations. In addition, the database can be also accessed through a *RESTful* API. Lastly, a *Python* client and a multiplatform virtual machine allow for program-driven access from user desktop. This way, private methylation data can be compared to *NGSmethDB* without the need to upload them to public servers. Database website: http://bioinfo2.ugr.es/NGSmethDB.

INTRODUCTION

DNA methylation at the cytosine carbon 5 position (5mC) is the main epigenetic mark, being able to modify gene expression patterns without entail DNA sequence changes (1,2). Furthermore, such modification is reversible during cell differentiation (3,4). CpG methylation is involved in cell differentiation in mammals, gene imprinting, the inactivation of X chromosome and genome stability (1,5–11). Differential methylation at CpG islands, many of which overlap promoters, can control tissue specificity of gene expression (12).

The short-read data sets from whole-genome shotgun bisulfite sequencing (WGBS) can be used (13,14) to generate whole-genome methylation maps or methylomes (15–17). When integrated with other genome maps (18,19), they may help to understand how the changes in DNA methylation (20–22) interact with other genetic and epigenetic marks (23,24) to control normal development or to provoke pathological dysregulations, as cancer (25,26).

Current problems with NGS methylation profiling (Barturen *et al*. 'Error Correction in Methylation Profiling From NGS Bisulfite Protocols', in 'Algorithms for Next-Generation Sequencing Data: Techniques, Approaches and Applications', Springer, in preparation) include: (i) the correct handling of the different error sources that might appear along the process (sequencing errors, clonal reads, sequence variation, bisulfite failure and miss-alignments) and that could lead to wrong methylation calling; and (ii) the diversity of protocols, methods and optional parameter-values in carrying out the alignment of the reads to a reference genome, as well as the read-out of the methylation levels from the alignment (27,28). Although other methylation databases exist (17,29–38), the above problems are not always properly addressed and therefore methylation data stored in them can be often unsuitable for compara-

*To whom correspondence should be addressed. Tel: +34 958243261; Fax: +34 958244073; Email: oliver@ugr.es
Correspondence may also be addressed to Michael Hackenberg. Tel: +34 958249695; Fax: +34 958244073; Email: hackenberg@ugr.es
†These authors contributed equally to this work as the first authors.

tive downstream analyses. Another common problem is the restricted range of samples stored in some of the databases, either focusing on specific gene loci (29–32), tissues (33,34) or diseases (35–38) and thus hampering large-scale comparative analyses.

Several years ago we initiated the development of integrated methylation pipelines (28,39) to minimize all the potential errors, at the same time unifying the different protocols. Our most recent development is *MethFlow* (Lebrón et al. 'MethFlowVM: a virtual machine for the integral analysis of bisulfite sequencing data', in preparation), a pipeline integrating the stringent quality controls built into *MethylExtract* (28) with third-part software in order to produce enhanced methylomes. We use *MethFlow* to populate and update *NGSmethDB*, thus profiling the samples under uniform conditions, which enables comparative, large-scale downstream analyses. Furthermore, *NGSmethDB* now includes two additional data sets, which may be a valuable resource for the discovery of methylation epigenetic biomarkers: (i) differentially methylated single-cytosines; and (ii) methylation segments (i.e. genome regions of homogeneous methylation).

DATABASE CONTENT

Publicly available short-read data sets from WGBS bisulfite sequencing projects for different cell lines, primary tissues, pathological biopsies and autopsies were downloaded mainly from NCBI GEO (40) and the ROADMAP project (41). An updated list of the available methylomes, with detailed information on the source cell lines or tissues, is maintained online on the database website.

At the time of writing, the *NGSmethDB* includes 667 methylomes generated for CG, CHG (H = A, C, T) and CHH sequence contexts (Table 1), a significant increase over the 87 methylomes in the previous release. CG is the most spread methylation sequence context in mammals, while CHG and CHH have been recently found in almost all human tissues (3,21,42). The information stored for each sampled single-cytosine is detailed in Table 2. Of particular interest are the samples derived from primary tissues of three individuals from the ROADMAP project: 11 samples from STL001 (3-years-old healthy male), 11 samples from STL002 (30-years-old female, iron deficiency, bipolar disease) and 13 samples from STL003 (34-years-old male, polysubstance abuse).

Lastly, and as a novelty in this release of *NGSmethDB*, genome maps of differentially methylated cytosines (DMCs) and methylation segments for human (hg38) and tomato (sl2.50) were also included in the database. Both data sets are a valuable resource for the discovery of methylation epigenetic biomarkers. The information stored in *NGSmethDB* for each DMC is detailed in Supplementary Table S1, and that for methylation segments in Supplementary Table S2.

DATABASE BACK-END

Single-cytosine methylation, methylation segments and differential methylation data are stored hierarchically in *MongoDB* (https://www.mongodb.com/), a NoSQL database that avoids the traditional table-based relational database structure in favor of *JSON*-formatted documents with dynamic schemas. This makes the programmatic data comparison of different samples easier and faster. However, when querying the database CSV or TSV formatted files can be optionally generated, thus also allowing for downstream analyses by means of conventional spreadsheets. Each assembly is stored in a database and inside every database there is collection for each chromosome. Within the collection, each *JSON*-like document represents one cytosine and contains hierarchically all genotypes (i.e. the different individuals from which the samples were obtained), differential methylation and methylation data of all individuals and samples. The first level is the data type (genotype, methylation or differential methylation), the second is the individual, the third is the sample and the fourth are the data values themselves (i.e. the DNA methylation levels, the alleles, the methylation differences, etc.).

WHOLE-GENOME, SINGLE-CYTOSINE RESOLUTION METHYLOMES

The high-quality methylomes stored in *NGSmethDB* were produced by *MethylExtract* (28), a software for DNA methylation profiling and genotyping from the same sample. For the most recent genome assemblies, we used our improved methylation pipeline *MethFlow* (Lebrón *et al.* 'MethFlowVM: a virtual machine for the integral analysis of bisulfite sequencing data', in preparation), using optimized (default) values for all the samples. The core of *MethFlow* is *MethylExtract*, to which third-part software was added to improve its overall performance. The preprocessing improvement consists in the use of *Trimmomatic* (43) for adapter trimming and removing low quality 3′ ends. The alignment to a three letter genome is then performed by means of *Bismark* (44) that uses *Bowtie2* (45) as aligner. Next, BSeQC (46) is used for the elimination of known technical artefacts that may result in inaccurate methylation estimation. And finally, *MethylExtract* performs the methylation calling and genotyping combined step. This last software minimizes several important error sources like sequencing errors, bisulfite failure, clonal reads and single nucleotide variants. The result of the entire process is a high quality, whole-genome methylation map or methylome, as well as the genotypes at all single-cytosine positions.

Another important novelty of *MethFlow*, as compared with previous pipelines, is the incorporation of a two-step mapping process in order to (i) exploit the advantages of the new genome assembly models (47) and (ii) recover the useful information of multiple-mapped reads for the analysis (see Supplementary Figure S1 for details). First, the reads are mapped against a decoy assembly (canonical chromosomes + alternative loci + decoy sequences). This increases the number of correctly mapped reads but also the number of multiple-mapped reads (that are discarded). A certain percentage of those ambiguous reads are then recovered in a second mapping step against the canonical chromosomes.

DIFFERENTIALLY METHYLATED CYTOSINES

Given the increasing biological relevance of differential

Table 1. Number of methylomes by species and sequence context stored in *NGSmethDB*

Species	Reference genome assembly	Sequence context	No. of methylomes
Homo sapiens	hg19	CG	57
		CHG	54
Homo sapiens	hg38	CG	35
Pan troglodytes	panTro4	CG	5
		CHG	5
Macaca mulatta	rheMac3	CG	6
		CHG	6
Mus musculus	mm10	CG	41
		CHG	41
Solanum lycopersicum	sl2.50	CG	8
		CHG	8
		CHH	8
Solanum pimpinellifolium	sl2.50	CG	2
		CHG	2
		CHH	2
Arabidopsis thaliana	tair10	CG	129
		CHG	129
		CHH	129
		TOTAL	667

Table 2. Information stored in *NGSmethDB* for each single-cytosine. All fields are described as shown in the results of *NGSmethDB API client*. Each row corresponds to a single-cytosine and each column to a field. For more information, see the manual *NGSmethDB*.

Field	Description	Example
chrom	Chromosome	chr22
pos	Chromosome position	25174338
genotype	Genotype of methylation context	YG
methContext	Methylation context where is the cytosine	CG
w.methylatedReads	Number of reads in which this cytosine is methylated (Watson-strand only)	22
c.methylatedReads	Number of reads in which this cytosine is methylated (Crick-strand only)	27
methylatedReads	Number of reads in which this cytosine is methylated (both strands)	49
w.coverage	Number of reads mapped at this chromosome position (Watson-strand only)	26
c.coverage	Number of reads mapped at this chromosome position (Crick-strand only)	33
coverage	Number of reads mapped at this chromosome position (both strands)	59
w.methRatio	Methylated reads ratio at this chromosome position (Watson-strand only)	0.85
c.methRatio	Methylated reads ratio at this chromosome position (Crick-strand only)	0.82
methRatio	Methylated reads ratio at this chromosome position (both strands)	0.83
w.phredScore	Average sequencing quality score at this chromosome position (Watson-strand only)	39
c.phredScore	Average sequencing quality score at this chromosome position (Crick-strand only)	37
phredScore	Average sequencing quality score at this chromosome position (both strands)	38

methylation (48–50), a section of the database is now dedicated to precomputed DMCs. Variation in methylation levels can be caused by many different factors like cell type, genotype (individual), environment, age, sex, etc. and therefore the naive comparison of any kind of samples is not a reliable method to obtain biologically meaningful DMCs. Here, we restricted ourselves to compute differential methylation controlling only for the two main factors: the variation found among tissues within the same individual (intra-individual DMCs), and that found in the same tissue from different individuals (inter-individual DMCs), a strategy allowing for a consistent comparison of these two important levels of epigenetic variability (Lebrón *et al*. 'Intra- and inter-individual variability of single-cytosine methylation', in preparation). Right now, the relevant samples for human intra- and inter-individual variability are all from the ROADMAP project (41), which include three individuals and multiple tissue types with different replicas, but more comparisons might be added as soon as appropriate data become available. In tomatoes, we generated a catalogue of intra-individual DMCs by comparing two leaves of different age from the same plant of *Solanum pimpinellifolium* accession TO-937 (Gómez-Martín *et al*. 'Differential methylation in tomato leaves', in preparation), as well as another list of intra-cultivar DMCs in *Solanum lycopersicum* cv. Ailsa Craig on the basis of the whole-genome bisulfite sequencing on fruit in four stages of development carried out by Giovannoni and coworkers (51).

To detect DMCs between sample pairs we used the CG sequence context in humans and the CG and CHG contexts in tomato. Three statistical tests were applied: the Fisher's exact test as implemented in *methylKit* (52) and *MOABS* (53), and the similarity test implemented in *MOABS* (53). To ensure statistical consistence, only those cytosines reaching statistical significance by all three tests (consensus) were considered as DMCs. On the other hand, pair-wise comparisons between all relevant samples were made; reaching consensus in anyone of the pair-wise comparisons suffices to consider such cytosine position as a DMC. See Supplementary Table S1 for details on the information stored in *NGSmethDB* on each detected DMC.

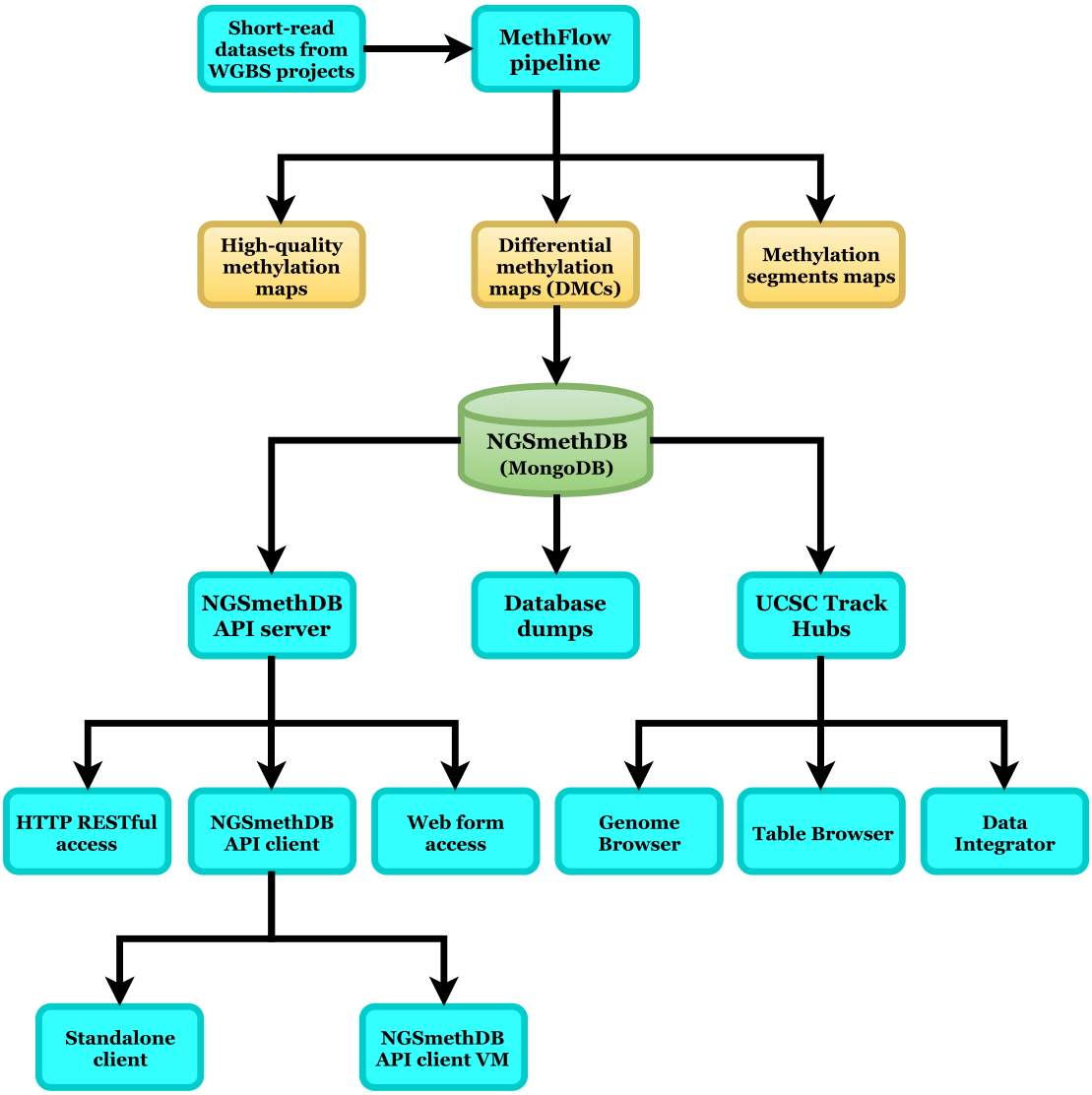

Figure 1. Data flow diagram for *NGSmethDB* indicating the source of primary data, the different types of extracted data and the different ways of data access.

METHYLATION SEGMENTS

Segmentation algorithms (54–57) can divide a DNA sequence into segments of homogeneous nucleotide composition, at a given significance level. In the same way, an array of single-cytosine methylation levels (in the CpG sequence context) along a chromosome sequence can be decomposed into segments of homogeneous methylation, thus revealing the regional variation of methylation levels. We have adapted our recursive segmentation algorithm to handle the methylation levels obtained by *MethFlow*. The details of the method will be given elsewhere (Carpena *et al.*, 'Segmenting whole-genome methylation maps', in preparation), but in essence the algorithm maximizes the difference of the mean values of adjacent segments by computing the *t*-statistic. Given the non-Gaussian nature of the distribution of methylation levels, the statistical significance of a given *t-value* was then obtained by a special randomization process, which takes into account both the methylation probability distribution in the sample (often bimodal) and the correlations among neighbour methylation values. See Supplementary Table S2 for details on the information stored in *NGSmethDB* on each methylation segment.

DATA SHARING, PROGRAMMATIC ACCESS AND VISUALIZATION

The data stored in *NGSmethDB* can be accessed in a variety of ways, depending on the data type or the access method (Figure 1). Single-cytosine methylation, differentially methylated cytosines and methylation segments can be accessed by downloading database dumps, querying the database through a custom web-form, using track hubs and Table Browser at UCSC, issuing HTTP simple queries on a web browser or through programmatic access (either using the *NGSmethDB API Client* or the *NGSmethDB API Virtual Machine*).

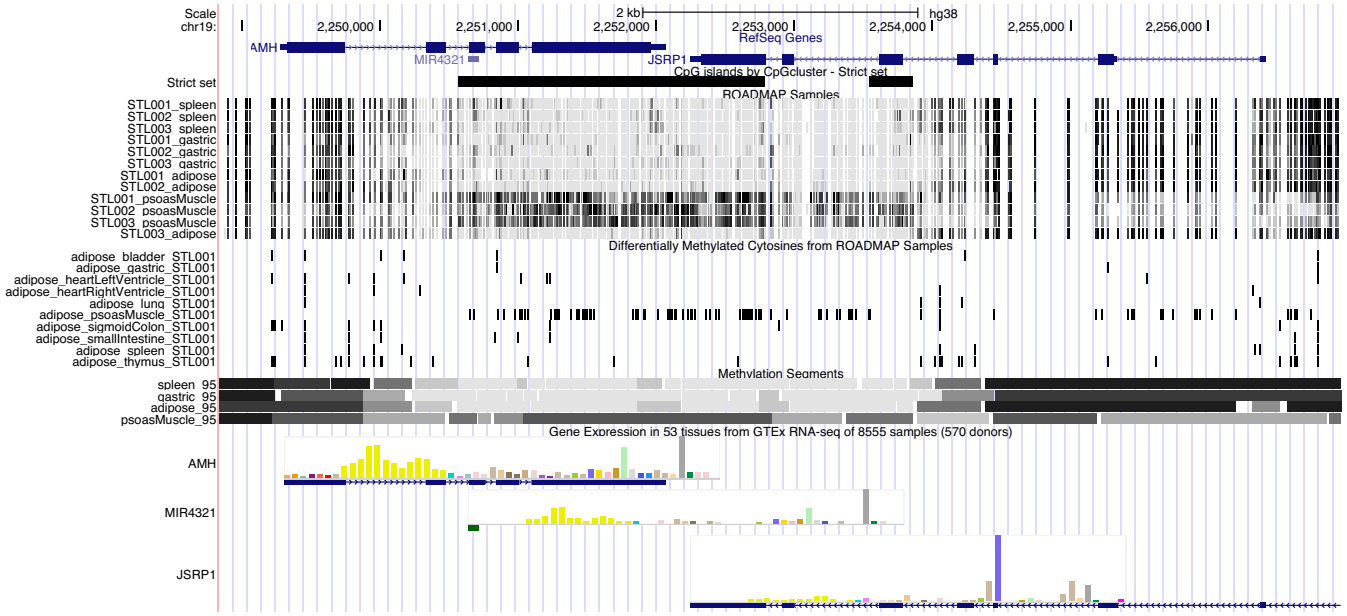

Figure 2. *NGSmethDB* data shown at the UCSC Genome Browser. A genome region of chromosome 19 (chr19:2, 248, 838-2, 256, 966) encompassing three genes (AMH, encoding for the anti-Mullerian hormone; MIR4321, encoding for the microRNA 4321; and JSRP1 encoding for a junctional sarcoplasmic reticulum protein) is shown. The three main types of *NGSmethDB* data are shown for different tissues: (i) methylation levels at single-cytosines; (ii) differentially methylated cytosines; and (iii) methylation segments. Third-part annotations are also shown: genes from the *Refseq* database (63), the strict set of CpG-islands predicted by *CpGcluster* (64,65) and gene expression levels from the NIH Genotype-Tissue Expression (GTEx) project (66). Online image: https://goo.gl/ElXE4t.

Database dumps are the easiest way to access *NGSmethDB* data. Complete zipped methylomes can be downloaded from the 'Content and dumps' page at the *NGSmethDB* website. Once unzipped, you get a tab-delimited file that can be directly opened in a spreadsheet for downstream analyses.

Another mode to access *NGSmethDB* data is through track hubs (58), which provide a standard and efficient mechanism for visualizing remotely hosted, Internet-accessible collections of genome annotations. Hub data sets can then be fully integrated into the University of California Santa Cruz (UCSC) Genome Browser (18). In this way, *NGSmethDB* data can be visualized and compared to a plethora of third-part annotations (Figure 2). In addition, UCSC tools, as *Table Browser* or *Data Integrator*, provide ways to (i) retrieve detailed *NGSmethDB* tab-delimited data sets from any genome, chromosome, genome region, gene, SNP or whatever other genome marker; (ii) combine methylation data and any other third-part annotation into a single set of data based on a specific join criteria, e.g. this can be used to find the methylation state of cytosines that intersect with CpG islands; and (iii) directly upload *NGSmethDB* data sets to public bioinformatics platforms as *Galaxy* (59), *GenomeSpace* (60) or *GREAT* (61) for further downstream, genome-wide analyses.

As a novelty for this release, and using Node.js (https://nodejs.org/en/), a *NGSmethDB API server* has been implemented on our server, which provides access to the entire *MongoDB* database via a *RESTful API* (62). This allows for additional web or programmatic ways to access *NGSmethDB* data:

i) Through a custom web-form; e.g. to retrieve single-cytosine methylation levels data from a specific genome region.
ii) HTTP access, i.e. to retrieve single-cytosine methylation data by directly issuing simple HTTP queries on the navigation bar of the user browser (see the *NGSmethDB* manual for details). Web-form and HTTP access methods are recommended only to retrieve data on a single position, or regions of moderate size as exons, genes, etc. Larger regions are better analyzed by means of track hubs or using the programmatic access.
iii) Program-driven access to *NGSmethDB*
   a) Through the *NGSmethDB Standalone API Client*. This is a multiplatform Python script (http://bioinfo2.ugr.es/NGSmethDB_API/NGSmethDB_API_client.py) running on Linux, Mac OS X and other UNIX systems. It permits to select assembly and samples, download methylation and differential methylation data and computing statistics for different genomic regions based on the coordinates provided as a BED formatted file —the BED format is described in UCSC FAQ (https://genome.ucsc.edu/FAQ/FAQformat.html#format1).
   b) Through the *NGSmethDB API Virtual Machine*. The *NGSmethDB API Client* has been encapsulated in a *VirtualBox* (https://www.virtualbox.org/) preconfigured Virtual Machine on which all the dependencies are already installed. The *NGSmethDB API VM* (http://bioinfo2.ugr.es/NGSmethDB_API/NGSmethDB_API.ova) is platform independent, being able to run on Windows, Linux or Mac desktops.

CONFIDENTIALITY ISSUES

It is increasingly frequent the use of private data that should not abandon the user desktop, particularly in biomedical and biotechnological genome research; therefore, uploading these data to a public server is frequently prohibitive. This is why we implemented the *NGSmethDB API server*, and developed also a *Python* standalone client able to programmatically access all the *NGSmethDB* data while running on the user desktop. The *NGSmethDB API client* has been already implemented within two multiplatform preconfigured virtual machines coming with all the dependencies installed: (i) the *NGSmethDB API virtual machine* (http://bioinfo2.ugr.es/NGSmethDB_API/NGSmethDB_API.ova), described in this paper, which allow downloading all the *NGSmethDB* data (single-cytosine methylation, methylation segments and differentially methylated cytosines); and (ii) *MethFlow*^{VM} (http://bioinfo2.ugr.es:8080/MethFlow/) that can be used to obtain methylomes from private WGBS short-read data sets and compare them to the downloaded *NGSmethDB* methylomes. With these two tools, the user will no longer need to upload private data to any public server to carry out comparative analyses against *NGSmethDB* data.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

Spanish Government [AGL2013-49090-C2-2-R to J.L.O. and M.H. and FIS2012-36282 to P.C. and P.B.]; Ministry of Education of Spain [FPU13/05662 to R.L.]. Funding for open access charge: Department of Genetics, University of Granada; Spanish Government [AGL2013-49090-C2-2-R to J.L.O. and M.H. and FIS2012-36282].
Conflict of interest statement. None declared.

REFERENCES

1. Bird,A. (2011) Putting the DNA back into DNA methylation. *Nat. Genet.*, **43**, 1050–1051.
2. Bird,A. (2011) The dinucleotide CG as a genomic signalling module. *J. Mol. Biol.*, **409**, 47–53.
3. 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**, 315–22.
4. Laurent,L., Wong,E., Li,G., Huynh,T., Tsirigos,A., Ong,C.T., Low,H.M., Sung,K.W.K., Rigoutsos,I., Loring,J. *et al.* (2010) Dynamic changes in the human methylome during differentiation. *Genome Res.*, **20**, 320–331.
5. Bird,A. (1986) CpG-rich islands and the function of DNA methylation. *Nature*, **321**, 209–213.
6. Antequera,F. (2003) Structure, function and evolution of CpG island promoters. *Cell Mol. Life Sci.*, **60**, 1647–1658.
7. Antequera,F., Boyes,J. and Bird,A. (1990) High levels of de novo methylation and altered chromatin structure at CpG islands in cell lines. *Cell*, **62**, 503–514.
8. Schübeler,D. (2015) Function and information content of DNA methylation. *Nature*, **517**, 321–326.
9. Putiri,E.L. and Robertson,K.D. (2011) Epigenetic mechanisms and genome stability. *Clin. Epigenet.*, **2**, 299–314.
10. Jones,P.A. and Gonzalgo,M.L. (1997) Altered DNA methylation and genome instability: a new pathway to cancer? *Proc. Natl. Acad. Sci. U.S.A.*, **94**, 2103–2105.
11. Deaton,A. and Bird,A. (2011) CpG islands and the regulation of transcription. *Genes and Development*, **25**, 1010–1022.
12. Song,F., Smith,J.F., Kimura,M.T., Morrow,A.D., Matsuyama,T., Nagase,H. and Held,W.A. (2005) Association of tissue-specific differentially methylated regions (TDMs) with differential gene expression. *Proc. Natl. Acad. Sci. U.S.A.*, **102**, 3336–3341.
13. Pomraning,K.R., Smith,K.M. and Freitag,M. (2009) Genome-wide high throughput analysis of DNA methylation in eukaryotes. *Methods*, **47**, 142–150.
14. Laird,P.W. (2010) Principles and challenges of genomewide DNA methylation analysis. *Nat. Rev. Genet.*, **11**, 191–203.
15. Hackenberg,M., Barturen,G. and Oliver,J.L. (2011) NGSmethDB: a database for next-generation sequencing single-cytosine-resolution DNA methylation data. *Nucleic Acids Res.*, **39**, D75–D79.
16. Geisen,S., Barturen,G., Alganza,Á.M., Hackenberg,M. and Oliver,J.L. (2014) NGSmethDB: an updated genome resource for high quality, single-cytosine resolution methylomes. *Nucleic Acids Res.*, **42**, D53–D59.
17. Song,Q., Decato,B., Hong,E.E., Zhou,M., Fang,F., Qu,J., Garvin,T., Kessler,M., Zhou,J. and Smith,A.D. (2013) A reference methylome database and analysis pipeline to facilitate integrative and comparative epigenomics. *PloS One*, **8**, e81148.
18. Kent,W.J., Sugnet,C.W., Furey,T.S., Roskin,K.M., Pringle,T.H., Zahler,A.M. and Haussler,a. D. (2002) The Human Genome Browser at UCSC. *Genome Res.*, **12**, 996–1006.
19. ENCODE Project Consortium, T.E.P. (2004) The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science*, **306**, 636–640.
20. Heyn,H., Moran,S., Hernando-Herraez,I., Sayols,S., Gomez,A., Sandoval,J., Monk,D., Hata,K., Marques-Bonet,T., Wang,L. *et al.* (2013) DNA methylation contributes to natural human variation. *Genome Res.*, **23**, 1363–1672.
21. Schultz,M.D., He,Y., Whitaker,J.W., Hariharan,M., Mukamel,E. A., Leung,D., Rajagopal,N., Nery,J.R., Urich,M. A., Chen,H. *et al.* (2015) Human body epigenome maps reveal noncanonical DNA methylation variation. *Nature*, **523**, 212–216.
22. Xie,H., Wang,M., de Andrade,A., Bonaldo,M.D.F., Galat,V., Arndt,K., Rajaram,V., Goldman,S., Tomita,T. and Soares,M.B. (2011) Genome-wide quantitative assessment of variation in DNA methylation patterns. *Nucleic Acids Res.*, **39**, 4099–108.
23. Romanoski,C.E., Glass,C.K., Stunnenberg,H.G., Wilson,L. and Almouzni,G. (2015) Epigenomics: roadmap for regulation. *Nature*, **518**, 314–316.
24. Chambers,J.C., Loh,M., Lehne,B., Drong,A., Kriebel,J., Motta,V., Wahl,S., Elliott,H.R., Rota,F., Scott,W.R. *et al.* (2015) Epigenome-wide association of DNA methylation markers in peripheral blood from Indian Asians and Europeans with incident type 2 diabetes: a nested case-control study. *Lancet Diabetes Endocrinol.*, **3**, 526–534.
25. Weisenberger,D.J. and Liang,G. (2015) Contributions of DNA methylation aberrancies in shaping the cancer epigenome. *Transl. Cancer Res.*, **4**, 219–234.
26. Plass,C., Pfister,S.M., Lindroth,A.M., Bogatyrova,O., Claus,R. and Lichter,P. (2013) Mutations in regulators of the epigenome and their connections to global chromatin patterns in cancer. *Nat. Rev. Genet.*, **14**, 765–780.
27. Krueger,F., Kreck,B., Franke,A. and Andrews,S.R. (2012) DNA methylome analysis using short bisulfite sequencing data. *Nat. Methods*, **9**, 145–151.
28. Barturen,G., Rueda,A., Oliver,J.L. and Hackenberg,M. (2013) MethylExtract: High-Quality methylation maps and SNV calling from whole genome bisulfite sequencing data. *F1000Research*, **2**, 217.
29. Amoreira,C., Hindermann,W. and Grunau,C. (2003) An improved version of the DNA Methylation database (MethDB). *Nucleic Acids Res.*, **31**, 75–77.
30. Grunau,C., Renault,E., Rosenthal,A. and Roizes,G. (2001) MethDB–a public database for DNA methylation data. *Nucleic Acids Res.*, **29**, 270–274.
31. Negre,V. and Grunau,C. (2006) The MethDB DAS Server: adding an epigenetic information layer to the human genome. *Epigenetics*, **1**, 101–105.
32. 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.

33. Xin,Y., Chanrion,B., O'Donnell,A.H., Milekic,M., Costa,R., Ge,Y. and Haghighi,F.G. (2012) MethylomeDB: a database of DNA methylation profiles of the brain. *Nucleic Acids Res.*, **40**, D1245–D1249.
34. Shi,J., Hu,J., Zhou,Q., Du,Y. and Jiang,C. (2013) PEpiD: a Prostate Epigenetic Database in Mammals. *PloS One*, **8**, e64289.
35. Lv,J., Liu,H., Su,J., Wu,X., Liu,H., Li,B., Xiao,X., Wang,F., Wu,Q. and Zhang,Y. (2012) DiseaseMeth: a human disease methylation database. *Nucleic Acids Res.*, **40**, D1030–D1035.
36. Vaca-Paniagua,F., Oliver,J., Nogueira da Costa,A., Merle,P., McKay,J., Herceg,Z. and Holmila,R. (2015) Targeted deep DNA methylation analysis of circulating cell-free DNA in plasma using massively parallel semiconductor sequencing. *Epigenomics*, **7**, 353–362.
37. He,X., Chang,S., Zhang,J., Zhao,Q., Xiang,H., Kusonmano,K., Yang,L., Sun,Z.S., Yang,H. and Wang,J. (2008) MethyCancer: the database of human DNA methylation and cancer. *Nucleic Acids Res.*, **36**, D836–D841.
38. Gu,F., Doderer,M.S., Huang,Y.-W., Roa,J.C., Goodfellow,P.J., Kizer,E.L., Huang,T.H.M., Chen,Y., Noushmehr,H., Weisenberger,D. *et al.* (2013) CMS: a web-based system for visualization and analysis of genome-wide methylation data of human cancers. *PLoS One*, **8**, e60980.
39. Hackenberg,M., Barturen,G. and Oliver,J.L. (2012) DNA methylation profiling from high-throughput sequencing data. In: Tatarinova,T and Kerton,O (eds). *DNA Methylation - From Genomics to Technology*. InTech.
40. Barrett,T., Wilhite,S.E., Ledoux,P., Evangelista,C., Kim,I.F., Tomashevsky,M., Marshall,K.A., Phillippy,K.H., Sherman,P.M., Holko,M. *et al.* (2013) NCBI GEO: archive for functional genomics data sets–update. *Nucleic Acids Res.*, **41**, D991–D995.
41. Consortium,R.E., Kundaje,A., Meuleman,W., Ernst,J., Bilenky,M., Yen,A., Heravi-Moussavi,A., Kheradpour,P., Zhang,Z., Wang,J. *et al.* (2015) Integrative analysis of 111 reference human epigenomes. *Nature*, **518**, 317–330.
42. Xie,W., Barr,C.L., Kim,A., Yue,F., Lee,A.Y., Eubanks,J., Dempster,E.L. and Ren,B. (2012) Base-resolution analyses of sequence and parent-of-origin dependent DNA methylation in the mouse genome. *Cell*, **148**, 816–831.
43. Bolger,A.M., Lohse,M. and Usadel,B. (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, **30**, 2114–2120.
44. Krueger,F. and Andrews,S.R. (2011) Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics*, **27**, 1571–1572.
45. Langmead,B. and Salzberg,S.L. (2012) Fast gapped-read alignment with Bowtie 2. *Nat. Methods*, **9**, 357–359.
46. Lin,X., Sun,D., Rodriguez,B., Zhao,Q., Sun,H., Zhang,Y. and Li,W. (2013) BSeQC: quality control of bisulfite sequencing experiments. *Bioinformatics*, **29**, 3227–3229.
47. Church,D.M., Schneider,V.A., Steinberg,K., Schatz,M.C., Quinlan,A.R., Chin,C.-S., Kitts,P.A., Aken,B., Marth,G.T., Hoffman,M.M. *et al.* (2015) Extending reference assembly models. *Genome Biol.*, **16**, 13.
48. Esteller,M., Corn,P.G., Baylin,S.B. and Herman,J.G. (2001) A gene hypermethylation profile of human cancer. *Cancer Res.*, **61**, 3225–3229.
49. Doi,A., Park,I.-H., Wen,B., Murakami,P., Aryee,M.J., Irizarry,R., Herb,B., Ladd-Acosta,C., Rho,J., Loewer,S. *et al.* (2009) Differential methylation of tissue- and cancer-specific CpG island shores distinguishes human induced pluripotent stem cells, embryonic stem cells and fibroblasts. *Nat. Genet.*, **41**, 1350–1353.
50. Gu,J., Stevens,M., Xing,X., Li,D., Zhang,B., Payton,J.E., Oltz,E.M., Jarvis,J.N., Jiang,K., Cicero,T. *et al.* (2016) Mapping of variable DNA methylation across multiple cell types defines a dynamic regulatory landscape of the human genome. *G3 (Bethesda)*, **6**, 973–986.
51. Zhong,S., Fei,Z., Chen,Y.-R., Zheng,Y., Huang,M., Vrebalov,J., McQuinn,R., Gapper,N., Liu,B., Xiang,J. *et al.* (2013) Single-base resolution methylomes of tomato fruit development reveal epigenome modifications associated with ripening. *Nat. Biotechnol.*, **31**, 154–159.
52. Akalin,A., Kormaksson,M., Li,S., Garrett-Bakelman,F.E., Figueroa,M.E., Melnick,A. and Mason,C.E. (2012) methylKit: a comprehensive R package for the analysis of genome-wide DNA methylation profiles. *Genome Biol.*, **13**, R87.
53. Sun,D., Xi,Y., Rodriguez,B., Park,H.J., Tong,P., Meong,M., Goodell,M. a and Li,W. (2014) MOABS: model based analysis of bisulfite sequencing data. *Genome Biol.*, **15**, R38.
54. Bernaola-Galván,P., Román-Roldán,R. and Oliver,J.L. (1996) Compositional segmentation and long-range fractal correlations in DNA sequences. *Phys. Rev. E. Stat. Phys. Plasmas Fluids Relat. Interdiscip. Topics*, **53**, 5181–5189.
55. Bernaola-Galván,P., Oliver,J.L., Hackenberg,M., Coronado,a. V., Ivanov,P.C. and Carpena,P. (2012) Segmentation of time series with long-range fractal correlations. *Eur. Phys. J. B*, **85**, 211.
56. Oliver,J.L., Román-Roldán,R., Pérez,J. and Bernaola-Galván,P. (1999) SEGMENT: identifying compositional domains in DNA sequences. *Bioinformatics*, **15**, 974–979.
57. Carpena,P., Oliver,J.L., Hackenberg,M., Coronado,a. V., Barturen,G. and Bernaola-Galván,P. (2011) High-level organization of isochores into gigantic superstructures in the human genome. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.*, **83**, 031908.
58. Raney,B.J., Dreszer,T.R., Barber,G.P., Clawson,H., Fujita,P.A., Wang,T., Nguyen,N., Paten,B., Zweig,A.S., Karolchik,D. *et al.* (2014) Track data hubs enable visualization of user-defined genome-wide annotations on the UCSC Genome Browser. *Bioinformatics*, **30**, 1003–1005.
59. Afgan,E., Baker,D., van den Beek,M., Blankenberg,D., Bouvier,D., Čech,M., Chilton,J., Clements,D., Coraor,N., Eberhard,C. *et al.* (2016) The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. *Nucleic Acids Res.*, **44**, gkw343.
60. Qu,K., Garamszegi,S., Wu,F., Thorvaldsdottir,H., Liefeld,T., Ocana,M., Borges-Rivera,D., Pochet,N., Robinson,J.T., Demchak,B. *et al.* (2016) Integrative genomic analysis by interoperation of bioinformatics tools in GenomeSpace. *Nat. Methods*, **13**, 245–247.
61. McLean,C.Y., Bristor,D., Hiller,M., Clarke,S.L., Schaar,B.T., Lowe,C.B., Wenger,A.M. and Bejerano,G. (2010) GREAT improves functional interpretation of cis-regulatory regions. *Nat. Biotechnol.*, **28**, 495–501.
62. Fielding,R.T. (2000) Architectural styles and the design of network-based software architectures. *Doctoral Dissertation*. University of California, Irvine.
63. O'Leary,N.A., Wright,M.W., Brister,J.R., Ciufo,S., Haddad,D., McVeigh,R., Rajput,B., Robbertse,B., Smith-White,B., Ako-Adjei,D. *et al.* (2016) Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.*, **44**, D733–D745.
64. Hackenberg,M., Previti,C., Luque-Escamilla,P., Carpena,P., Martínez-Aroza,J. and Oliver,J. (2006) CpGcluster: a distance-based algorithm for CpG-island detection. *BMC Bioinformatics*, **7**, 446.
65. 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.
66. Lonsdale,J., Thomas,J., Salvatore,M., Phillips,R., Lo,E., Shad,S., Hasz,R., Walters,G., Garcia,F., Young,N. *et al.* (2013) The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.*, **45**, 580–585.