DeepBlueR: Large-scale epigenomic analysis in R

Markus List[1], Felipe Albrecht[1,2], Christoph Bock[1,3,4] and Thomas Lengauer[1]
[1] *Max Planck Institute for Informatics, Saarland Informatics Campus, Saarbrücken, Germany.*
[2] *Graduate School of Computer Science, Saarland Informatics Campus, Saarbrücken.*
[3] *CeMM Research Center for Mol. Medicine of the Austrian Academy of Sciences, Vienna, Austria.*
[4] *Department of Laboratory Medicine, Medical University of Vienna, Vienna, Austria.*
markus.list@mpi-inf.mpg.de

The first complete draft of the human genome was expected to quickly advance our understanding of gene regulation and disease mechanisms. However, researchers have since realized that we have to look beyond the genome to understand the complex behavior observed between different cell types that all share the same genetic information. Epigenetic research focuses on understanding non-inheritable factors influencing gene regulation and covers various cellular mechanisms such as DNA methylation, histone modification, miRNA function and transcription factor binding sites. Recent advances in high-throughput profiling technologies allow for systematically collecting data on each of these mechanisms in large-scale experiments. These efforts are fostered and concerted by international collaborations, such as the International Human Epigenome Consortium (IHEC) [SHC+16] and its members. As a result of these collaborations, researchers can exploit massive amounts of publicly available epigenomic data on dozens of cell types, cell lines and tissues. Access to these data is streamlined by existing data portals [BdLMG+16, FdlTR+16] and, in principle, allows for answering important biomedical questions.

However, working with such data requires a suitable computational infrastructure not accessible ubiquitously. This creates a serious bottleneck in research and, as a result, data from these costly experiments are currently underused. To address this issue, we developed a new web resource, the DeepBlue Epigenomic Data Server [ALBL16] to provide access to more than 40,000 experimental files from four major epigenome projects: ENCODE [C+04], ROADMAP [KME+15], BLUEPRINT [AAA+12], the German Epigenome Program DEEP (`http://www.deutsches-epigenom-programm.de`), the Canadian CEEHRC (`http://www.epigenomes.ca/`), and the Japanese CREST (`http://crest-ihec.jp/english/database/index.html`).

A common challenge with this resources is that researchers are typically interested in a small fraction of the available epigenomic data to answer specific biomedical questions. Using a typical data repository to solve this task would require the user to download several files amounting to gigabytes of data that subsequently need to be filtered locally. In addition, it is often important to perform memory- and cpu-intensive operations to transform or aggregate these data, while the necessary computational resources are not accessible to every user. Therefore, the DeepBlue Data Server offers features beyond those of a centralized epigenomic data repository. It has a comprehensive programmatic interface (API) to enable users to perform complex data operations, such as searching, selecting, filtering, summarizing, and downloading of epigenomic data of interest. These operations can be combined into custom workflows, thus offering nearly the same degree of flexibility as a local programming environment.

Access to the DeepBlue API is possible through a widely used web standard and, thus, any of the major programming languages can be used to access the functionality. The R/Bioconductor environment is particularly popular for data analyses, which motivated us to develop an R/Bioconductor package that streamlines access to the DeepBlue API. Here, we present DeepBlueR [ALBL17], a new R/Bioconductor package that enables users to engage with the DeepBlue server in a seamless fashion from within the R environment. DeepBlueR mirrors all DeepBlue data operations as R commands and provides additional features for compressing, downloading and transforming aggregated epigenomic data into suitable R data structures. A mechanism for local caching guarantees that complex scripts can be executed with-

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3224v1 | CC BY 4.0 Open Access | rec: 5 Sep 2017, publ: 5 Sep 2017

out the need to download previously requested data from the server.

To demonstrate the power of this approach, we show how a few R commands suffice to perform a genome-wide analysis of DNA methylation across 212 samples from the BLUEPRINT consortium. DeepBlueR comes with extensive documentation, including a user manual, code examples and use cases. Access to DeepBlue is possible anonymously but registered users gain access to additional features such as private data and a history of their activity.

In summary, DeepBlueR is a new software package that enables R users to tackle large amounts of epigenomic data in their analysis through server-side processing while being able to exploit the rich R/Bioconductor ecosystem for subsequent analysis and visualization.

Availability:

DeepBlue project: `http://deepblue.mpi-inf.mpg.de`
DeepBlueR package: `http://bioconductor.org/packages/DeepBlueR/`

References

[AAA+12] David Adams, Lucia Altucci, Stylianos E Antonarakis, Juan Ballesteros, Stephan Beck, Adrian Bird, Christoph Bock, Bernhard Boehm, Elias Campo, Andrea Caricasole, et al. BLUEPRINT to decode the epigenetic signature written in blood. *Nature biotechnology*, 30(3):224, 2012.

[ALBL16] Felipe Albrecht, Markus List, Christoph Bock, and Thomas Lengauer. DeepBlue epigenomic data server: programmatic data retrieval and analysis of epigenome region sets. *Nucleic acids research*, 44(W1):W581–W586, 2016.

[ALBL17] Felipe Albrecht, Markus List, Christoph Bock, and Thomas Lengauer. DeepBlueR: large-scale epigenomic analysis in R. *Bioinformatics*, 2017.

[BdLMG+16] David Bujold, David Anderson de Lima Morais, Carol Gauthier, Catherine Côté, Maxime Caron, Tony Kwan, Kuang Chung Chen, Jonathan Laperle, Alexei Nordell Markovits, Tomi Pastinen, et al. The International Human Epigenome Consortium Data Portal. *Cell Systems*, 3(5):496–499, 2016.

[C+04] ENCODE Project Consortium et al. The ENCODE (ENCyclopedia of DNA elements) project. *Science*, 306(5696):636–640, 2004.

[FdlTR+16] José María Fernández, Victor de la Torre, David Richardson, Romina Royo, Montserrat Puiggròs, Valentí Moncunill, Stamatina Fragkogianni, Laura Clarke, Paul Flicek, Daniel Rico, et al. The BLUEPRINT Data Analysis Portal. *Cell systems*, 3(5):491–495, 2016.

[KME+15] Anshul Kundaje, Wouter Meuleman, Jason Ernst, Misha Bilenky, Angela Yen, Pouya Kheradpour, Zhizhuo Zhang, Alireza Heravi-Moussavi, Yaping Liu, Viren Amin, et al. Integrative analysis of 111 reference human epigenomes. *Nature*, 518(7539):317, 2015.

[SHC+16] Hendrik G Stunnenberg, Martin Hirst, International Human Epigenome Consortium, et al. The International Human Epigenome Consortium: A Blueprint for Scientific Collaboration and Discovery. *Cell*, 167(5):1145–1149, 2016.