A framework for analysis of DNA methylation data

Stephan Flemming[1], Björn Andreas Grüning[1], Thomas Häupl[2], Stefan Günther[1]

[1] *Pharmaceutical Bioinformatics, Institute of Pharmaceutical Sciences, University of Freiburg*
[2] *Department of Rheumatology and Clinical Immunology, Charité University Hospital, Berlin*

stephan.flemming@pharmazie.uni-freiburg.de

Introduction
DNA Methylation data analysis may include raw data preprocessing, detection of significant differential methylation sites, network based approaches, or the analysis of correlation with gene expression data. Usage of several tools is frequently associated with the problem of incompatible file formats and the requirement of expert knowledge for each application. Therefore, advanced analyses are often omitted or cannot be sufficiently adapted to the user's needs. Furthermore, sharing of workflows, checking of intermediate results, and the reproducibility of complex calculations may be difficult.

Methods
To address these problems we have applied the workflow management system Galaxy [Bla10]. We present a comprehensive framework that provides a broad range of functionality for the analysis of DNA methylation data. All tools have been integrated as modules and can be linked together in reusable and shareable workflows. The framework can be used directly with a web browser, without any need for programming skills. All resulting data sets and appliances of tools are automatically logged, thus the reproducibility is further increased.
Our framework supports the analysis of Illumina Infinium Methylation Assays and provides in-house developed modules as well as widely used tools and methods for DNA methylation data evaluation, e.g. several libraries of the Bioconductor package. The following methods have been included:
- raw data preprocessing (e.g. data normalization)
- integration of GenomeStudio2011 (Illumina, Inc.)
- differential methylation detection (e.g. Fold change, linear model for log-transformation, F-Test, t-Test)
- comparative analysis of correlation networks
- ID mapping (e.g. GeneOntology)
- access to repositories for high-throughput data (e.g. NCBI-GEO)
- data visualization

The capability for data analysis is demonstrated by the evaluation of selected data sets.

Availability and Future Prospects
The framework will be available via Internet for third parties. For integration of Omics data the implementation of additional methods is planned, e.g. for correlation analysis of methylation and gene expression data.

References
[Bla10] Blankenberg D, Von Kuster G, Coraor N, Ananda G, Lazarus R, Mangan M, Nekrutenko A, Taylor J. "Galaxy: a web-based genome analysis tool for experimentalists". *Current Protocols in Molecular Biology*. 2010 Jan; Chapter 19:Unit 19.10.1-21.