JIlumina - A comprehensive Java-based API for statistical Illumina Infinium HumanMethylation450 and Infinium MethylationEPIC BeadChip data processing

Diogo Almeida[1 5], Ida Skov[2], Jesper Lund[2], Afsaneh Mohammadnejad [2], Artur Silva[5], Fabio Vandin[6 2], Qihua Tan[3 4], Jan Baumbach[2] and Richard Röttger[2,*]

[1]Department of Biochemistry and Molecular Biology, University of Southern Denmark, Campusvej 55, 5230 Odense, Denmark

[2]Department of Mathematics and Computer Science, University of Southern Denmark, Campusvej 55, 5230 Odense, Denmark

[3]Unit of Human Genetics, Department of Clinical Research, Faculty of Health Science, University of Southern Denmark, 5000 Odense, Denmark

[4]Epidemiology, Biostatistics and Biodemography, Department of Public Health, Faculty of Health Science, University of Southern Denmark, 5000 Odense, Denmark

[5]Laboratory for Genomics and Bioinformatic, Institute of Biological Sciences, Federal University of Pará, 66075110, Belém, Brazil

[6]Departament of Information Engineering, University of Padova, Via Gradenigo 6/B, I-35131 Padova, Italy

Summary

Measuring differential methylation of the DNA is the nowadays most common approach to linking epigenetic modifications to diseases (called epigenome-wide association studies, EWAS). For its low cost, its efficiency and easy handling, the Illumina HumanMethylation450 BeadChip and its successor, the Infinium MethylationEPIC BeadChip, is the by far most popular techniques for conduction EWAS in large patient cohorts. Despite the popularity of this chip technology, raw data processing and statistical analysis of the array data remains far from trivial and still lacks dedicated software libraries enabling high quality and statistically sound downstream analyses.

As of yet, only R-based solutions are freely available for low-level processing of the Illumina chip data. However, the lack of alternative libraries poses a hurdle for the development of new bioinformatic tools, in particular when it comes to web services or applications where run time and memory consumption matter, or EWAS data analysis is an integrative part of a bigger framework or data analysis pipeline. We have therefore developed and implemented JIlumina, an open-source Java library for raw data manipulation of Illumina Infinium HumanMethylation450 and Infinium MethylationEPIC BeadChip data, supporting the developer with Java functions covering reading and preprocessing the raw data, down to statistical assessment, permutation tests, and identification of differentially methylated loci. JIlumina is fully parallelizable and publicly available at `http://dimmer.compbio.sdu.dk/download.html`

*To whom correspondence should be addressed. Email: roettger@imada.sdu.dk

1 Introduction

DNA methylation is an epigenetic process associated to genomic imprinting, inactivation of the X chromosome in females, and cellular specialization [1]. DNA methylation patterns are subject to naturally occurring changes due to aging but are also influenced by the development of diseases such as cancer and diabetes [2]. Differentially methylated DNA loci can be identified by using DNA bisulftite treatment and quantified by microarray technologies such as the Illumina Infinium HumanMethylation450 and Infinium MethylationEPIC BeadChip, which cover around 450,000 and 850,000 CpG sites respectively along the human genome [3, 4]. Even though this is only a very small fraction of all (estimated 20 millions) CpG sites in the human genome, the Illumina probes are designed to assess CpG positions of 99% of the RefSeq sequences. 72% of these CpG positions are distributed in promoter regions and gene body, the remaining in intergenic and 3'UTR regions [3, 4].

Due to the popularity and importance of the Illumina arrays it is crucial that the practitioners are equipped with easy-to-use, interactive and integrated software solutions, which aid with all necessary data analysis steps from raw data processing to the identification of significantly loci differentially methylated between two or more groups of patients, or correlated with clinical outcome parameters. Software developers, so far, face a set of central issues when working with Illumina array data: The raw data is only provided in an undocumented proprietary binary file format (IDAT). Accessing the data requires the usage of Illumina's closed-source GenomeStudio software in order to manually convert the IDAT file into a human-readable and parsable ASCII file [1]. Recently, essential parts of the file format have been re-engineered in the framework of the open source R-package Illuminaio [1], which now provides simple I/O operation functionality in R.

The advent of Illuminaio has lead to the development of several of sophisticated software packages for EWAS data analysis, such as Minfi [5], RNBeads [6] and WaterMelon [7]. As these packages are based on Illuminaio, they are implemented in R, which poses some limitations when considering the usability for non-computer experts. Even for experienced software developers, R may render problems: R versions and package incompatibilities, overhead when binding with libraries, lack of software design patters, performance and integration in non-R based frameworks, just to name a few.

In order to provide the community with a wider basis for future software implementations, we have developed Jllumina, a Java library for the handling and processing of Illumina raw data files. It is the backbone of the recently published DiMmer tool [8]. Besides the very I/O capability, we have additionally implemented popular post-processing functions like quality control, normalization, cell composition correction, and significance level calculations. The provided features are mostly parallelizable re-implementations based on the R packages Illuminaio [1] and Minfi [5]. The library can process both, Illumina Infinium HumanMethylation450 data and the new Infinium MethylationEPIC BeadChip data.

In the remainder of this manuscript, we first describe the general workflow of an EWAS study, and the raw data format. Afterwards, we describe the methods provided by Jllumina in detail. The technical documentation can be found online at `http://dimmer.compbio.sdu.dk/javadoc`.

	Loading set	Cell composition	Quantile normalization	CpG statistics
Jllumina	14 secs	66 secs	21 secs	89 secs
Minfi	68 secs	194 secs	32 secs	126 secs

Table 1: Jllumina and Minfi R package run times. All steps were performed on a standard Macbook Air with Intel core i5, 1.6 **Ghz and 8GB of RAM. We used 1000 permutations for Jllumina's CpG statistics.**

2 Methods

2.1 Overview

EWAS data analysis can be divided in two major parts: pre-processing and downstream analysis. The first one regards to the manipulation of the methylation signal distribution in order to remove artifacts, which can compromise the downstream analysis. Downstream analysis, relies on the discovery of differentially methylated positions/regions and its association to a specific phenotype/disease of interest, and is strongly dependent of the pre-processing step. Jlummina provides an API for the implementation of any step associated to the pre-processing of Illumina raw data: (i) Background correction, which is related to the source of noise intrinsic to the technology; (ii) Probe filtering: The quality of the signal from a signal probe can be checked using the control probes provided by Illumina. Removing low quality probes can maximize the correct identification of differentially methylated positions; (iii) Normalization removes cross-array artifacts, and integrates the distributions from Infinium I and Infinium II probe types. In addition, Jlummina also provide classes for downstream analyses: (i) statistical testing of single CpG sites using t-test (case-control), and (ii) linear regression models (multiple labels), as well as (iii) permutation test functions to retrieve empirical p-values, and (iv) methods for correcting p-values for multiple testing. A comparison table, for the steps mentioned before, with well known R-packages is available at `http://dimmer.compbio.sdu.dk/download.html`. In addition, we give run times for Illumina array data comparing the Minfi R package and Jllumina in Table 1, using the Polycystic ovarian syndrome (PCOS) dataset, consisting of a case-control cohort with 60 Chinese female patients [9].

2.2 Calculating the Methylation Levels

An IDAT file stores the methylation intensities for a large set of CpG sites of a single sample. Note that the Illumina Infinium HumanMethylation450 and Infinium MethylationEPIC BeadChip methylation platforms actually employs two different probe types simultaneously in one experiment: *Infinium I* (which uses two assays for a given locus to detect the methylated and unmethylated signal) and *Infinium II* (which uses only one assay to explore a locus pair but in different color channels). Each probe, independent of its assay type, results in two different intensity values, one for the methylated and one for the unmethylated channel.

The most important information about a CpG site is its methylation level, which is either provided as the so-called β-value

$$\beta\text{-value} = \frac{M}{U + M}$$

or the Logit transformed M-value

$$M\text{-value} = \log_2 \left(\frac{\beta}{1 - \beta} \right)$$

where M corresponds to the mean intensity of methylated locus and U to the mean intensity of the unmethylated locus [10].

2.3 Background Correction

Illumina provides control probes to generate a background signal. The noise associated to the Illumina signal is (by default) assumed to be 5% of the background signal. Hence, our library reads the methylation level at 5% of the methylation distribution of the methylation levels of the control probes, and then subtracts this value from each of the 450,000 (850,000) CpG methylation signals of the given array.

2.4 Probe Filtering

The background signals from the the control probes are also used to assess the quality of the CpG sites' measurements. For a given array, the proportion of control probes in which its signal is at least as strong as the methylation signal from a given CpG site is called *detection-P*. The cumulative probability associated to this proportion is estimated assuming that the signal follows a standard normal distribution. Now low-quality CpG sites can be filtered by applying a p-value threshold. A CpG is dropped if the proportion of low-quality CpG sites across all samples is greater than a second threshold, default: 5%.

2.5 Normalization

Jllumina comes with two types of normalization methods: the Illumina normalization and quantile normalization. Any other normalization function can be developed by extending the AbstractNormalization class (details in Jllumina javadoc).

2.5.1 Illumina Normalization

This method is based on the Illumina software GenomeStudio. Essentially, it uses the mean methylation values of the control probes to calculate a factor for each sample in order to correct the signal from intra/inter array artifacts. We provide a full Java re-implementation of this method adopted from the R-based Minfi package. The function *preprocessIllumina* is an implementation of the Illumina normalization and background correction. We tested and matched our library's output against the Minfi package on some test data sets.

2.5.2 Quantile Normalization

The (un)methylated signals are obtained by the two different assays (Infinium I and Infinium II). They differ in the overall methylation signal distribution, which renders a direct comparison impossible. With Jllumina, we implemented a classical (non-stratified) quantile normalization and a subset quantile normalization (stratified), taking into account the Illumina probe annotation of CpG sites as described by Touleimar & Jörg Tost [11]. The idea of quantile normalization is to make two different distributions similar. This is achieved iin Jllumina by: (i) computing the row means of the sorted methylation signal matrix and (ii) replacing the original methylation values by the computed mean values, according to the assigned ranks. This approach is straightforward for the non-stratified quantile normalization. The stratified quantile normalization follows the same two steps described before for Infinium type II probes, but a different approach for Infinium type I probes: the row means are computed by interpolating a reference distribution from the normalized values of Infinium type II probes and then the column-wise replacement is performed. We tested both, the stratified and the non-stratified methods, for correctness against the distributions computed by Minfi package [5]. On our test data, the results were exactly the same but achieved much faster due Jllumina's full parallelization.

2.6 Blood Cell Composition Estimation

One of the major goals in epigenetics is the association of differentially methylated sites to a phenotype of interest. In this context, the observed methylation signal has to take into account heterogeneity of cells in a given sample. Since blood samples are often used for methylation analysis, with Jllumina we provide an implementation for estimating the cell blood population distribuion for six cell types (NK, CD8T, CD4T, granulocytes and monocytes). This implementation follows the same strategy as the estimateCellCounts function of the Minfi package. The cell composition estimation uses a linear regression calibration step, in which the methylation contribution (coefficients) from the six cell types are computed. Given the surrogate target methylation values and the computed coefficients the proportion of cells can be estimated by minimizing a quadratic objective function [12].

2.7 Statistical Significance

The class CpGStatistics was implemented in order to compute p-values, and coefficients given a beta-value set and a statistical estimator (t-test or linear regression). The t-test is used for case-control studies and performed for paired (e.g. twin cohorts) or unpaired data sets, and can assume equal or non-equal variance between the two populations; likewise for linear regression models.

2.8 Permutation Tests

In order to assess p-value significance the CpGStatistics class also provides a parallelized permutation test function and a set of multiple testing correction functions. Here, we essentially

permute the patient labels and rerun the whole data analysis pipeline to calculate empirical p-values, the false discovery rate (FDR), the family wise error rate (FWER) and step-down min-p values. The empirical p-values can be understood as the frequency of permuted p-values at least as good as the original one. The FDR p-values (Beinjamini-Hochberg) [13] are computed as the proportion of permuted p-values, from all CpG sites, which are bigger than a given original p-value of a single CpG site. The FWER p-values (Bonferroni correction) is computed as the frequency of p-values, from a single CpG, which are at least as good as the smallest p-value found after running the permutations. Step-down minP [14] is a multiple test correction similar to the Bonferroni correction but less restrictive. Figure 1 plots the original test p-values against the corrected ones using the different methods described above using a Illumina Infinium HumanMethylation450 test data set available at http://dimmer.compbio.sdu.dk/download.html.

3 Results & Conclusion

With Jllumina, we present the first Java library providing for processing and manipulating Illumina IDAT files. Besides the capabilities for processing the raw input files, Jllumina also implements the most popular statistical analyses required for high-quality EWAS studies. Jllumina supports both, Illumina Infinium HumanMethylation450 and Infinium MethylationEPIC BeadChip data. Thus, Jllumina will enable Java developers to fully integrate standard EWAS data manipulation and statistics into their Java-based software frameworks. In the future, we will implement functions for NGS bisulfite data analysis and for integrating EWAS results with biological networks using Cytoscape and enrichment methods [15, 16, 17]. This will allow us to run de novo network enrichment to identify subnetworks (rather than DMRs) that are affected by differential DNA methylation.

4 Availability

Jllumina is open source and publicly available at `http://dimmer.compbio.sdu.dk/download.html`. The website also provides API documentation (`http://dimmer.compbio.sdu.dk/javadoc/index.html`) and coding examples.

Acknowledgements

DA is grateful for financial support from CAPES Brazil and the SDU E-Science center. JB received financial support from the SDU2020 initiative and a Young Investigator Grant from the VILLUM Foundation.

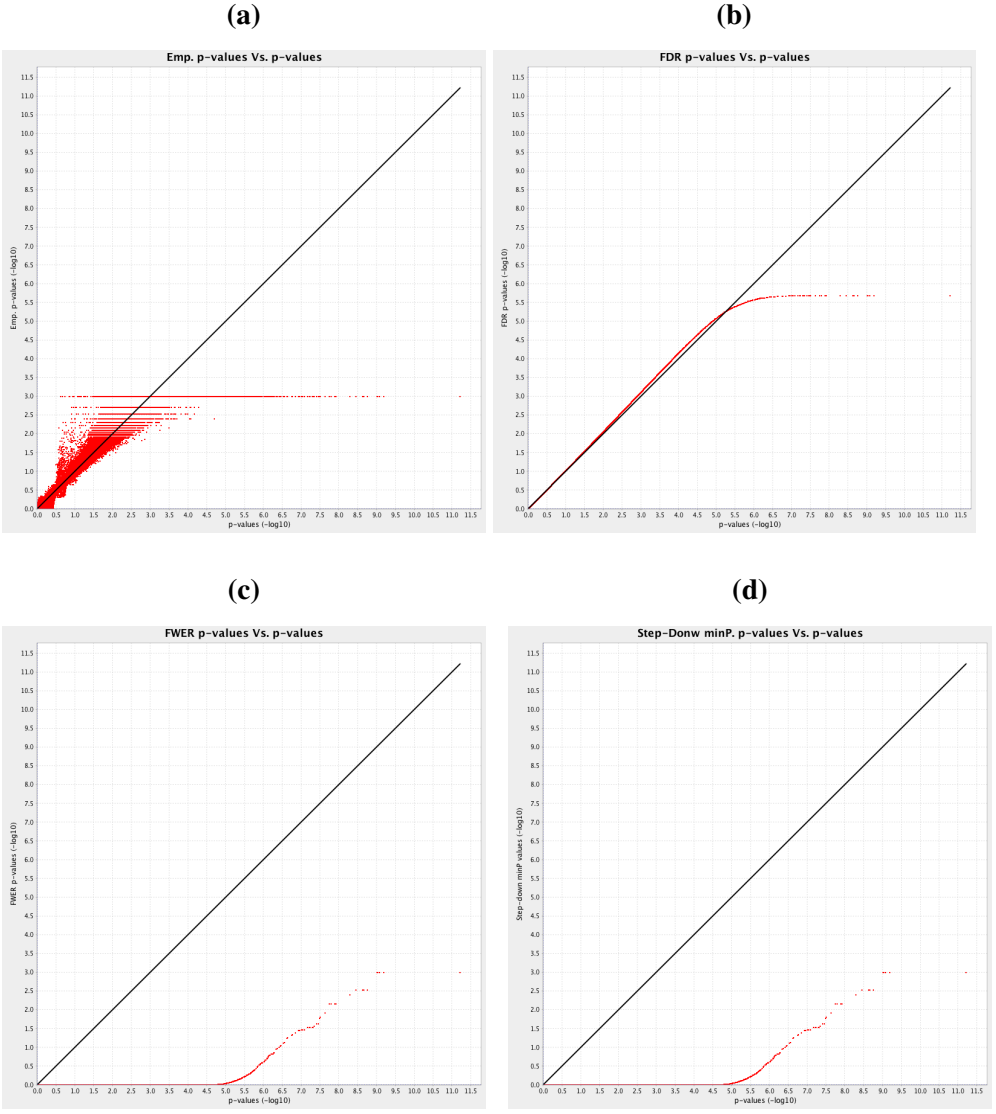

Figure 1: Original p-values plotted against a) empirical p-values, b) FDR-corrected p-values, c) FWER-corrected p-values, and d) Step-down minP corrected p-values, from the PCOS dataset using 1000 permutations. Apparently, Bonferoni corretion is too strict. We suggest to use FDR correction or empirical p-values.

References

[1] Z. D. Smith and A. Meissner. DNA methylation: roles in mammalian development. *Nature reviews. Genetics*, 14(3):204–20, 2013.

[2] K. D. Robertson. DNA methylation and human disease. *Nature reviews. Genetics*, 6(8):597–610, 2005.

[3] S. Moran, C. Arribas and M. Esteller. Validation of a DNA methylation microarray for 850,000 CpG sites of the human genome enriched in enhancer sequences. *Epigenomics*, 6(July 2015):epi.15.114, 2015.

[4] J. Sandoval, H. A. Heyn, S. Moran, J. Serra-Musach, M. A. Pujana, M. Bibikova and M. Esteller. Validation of a DNA methylation microarray for 450,000 CpG sites in the human genome. *Epigenetics*, 6(6):692–702, 2011.

[5] M. J. Aryee, A. E. Jaffe, H. Corrada-Bravo, C. Ladd-Acosta, A. P. Feinberg, K. D. Hansen and R. A. Irizarry. Minfi: A flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics*, 30(10):1363–1369, 2014.

[6] Y. Assenov, F. Müller, P. Lutsik, J. Walter, T. Lengauer and C. Bock. Comprehensive analysis of DNA methylation data with RnBeads. *Nature methods*, 11(11):1138–40, 2014.

[7] R. Pidsley, C. C. Y Wong, M. Volta, K. Lunnon, J. Mill and L. C. Schalkwyk. A data-driven approach to preprocessing Illumina 450K methylation array data. *BMC genomics*, 14(1):293, 2013.

[8] D. Almeida, I. Skov, A. Silva, F. Vandin, Q. Tan, R. Röttger and J. Baumbach. Efficient detection of differentially methylated regions using DiMmeR. *Bioinformatics*, (in press):2, 2016.

[9] S. Li, D. Zhu, H. Duan, A. Ren, D. Glintborg, M. Andersen, V. Skov, M. Thomassen, T. Kruse and Q. Tan. Differential DNA methylation patterns of polycystic ovarian syndrome in whole blood of Chinese women. *Oncotarget*, 5(0), 2016.

[10] C. S. Wilhelm-Benartzi, D. C. Koestler, M. R. Karagas, J. M. Flanagan, B. C. Christensen, K. T. Kelsey, C. J. Marsit, E. A. Houseman and R. Brown. Review of processing and analysis methods for DNA methylation array data. *British journal of cancer*, 109(6):1394–402, 2013.

[11] N. Touleimat and J. Tost. Complete pipeline for Infinium(®) Human Methylation 450K BeadChip data processing using subset quantile normalization for accurate DNA methylation estimation. *Epigenomics*, 4(3):325–41, 2012.

[12] E. A. Houseman, W. P. Accomando, D. C. Koestler, B. C. Christensen, C. J. Marsit, H. H. Nelson, J. K. Wiencke and K. T. Kelsey. DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC bioinformatics*, 13(1):86, 2012.

[13] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society B*, 57(1):289–300, 1995.

[14] S. Köhne and I. Pigeot. Resampling-Based Multiple Testing. Examples and Methods for p-Value Adjustment. *Comput Stat Data An*, 20:235–236, 1995.

[15] N. Alcaraz, H. Kücük, J. Weile, A. Wipat and J. Baumbach. KeyPathwayMiner: Detecting Case-Specific Biological Pathways Using Expression Data. *Internet Mathematics*, 7(4):299–313, 2011.

[16] J. Baumbach and L. Apeltsin. Linking Cytoscape and the corynebacterial reference database CoryneRegNet. *BMC genomics*, 9:184, 2008.

[17] N. Alcaraz, T. Friedrich, T. Kötzing, A. Krohmer, J. Müller, J. Pauling and J. Baumbach. Efficient key pathway mining: combining networks and OMICS data. *Integrative Biology*, 4(7):756–764, 2012.

[18] A. E. Jaffe and R. A. Irizarry. Accounting for cellular heterogeneity is critical in epigenome-wide association studies. *Genome biology*, 15(2):R31, 2014.