Comprehensive analysis of DNA methylation data with RnBeads

Yassen Assenov, Fabian Müller, Pavlo Lutsik, Jörn Walter, Thomas Lengauer & Christoph Bock

Supplementary Note

As an example for RnBeads-based analysis of Infinium 450k data, we performed a reanalysis of a publicly available glioblastoma dataset generated by The Cancer Genome Atlas (TCGA) project (Weisenberger, 2014). Glioblastoma multiforme is an aggressive type of brain cancer with a median survival time of little more than a year and substantial variation between patients (Wen and Kesari, 2008). In an attempt to stratify patients according to the molecular characteristics of the tumors, recent research has identified a subtype that is characterized by elevated levels of DNA methylation, prolonged survival and high frequency of mutations in the IDH1 gene (Noushmehr et al., 2010). The discovery of this "glioblastoma CpG island methylator phenotype positive" (G-CIMP+) subtype was based on Illumina's Infinium 27k assay, prompting us to validate this observation using RnBeads and an extended dataset of Infinium 450k profiles for 124 glioblastoma patients.

We downloaded the raw microarray signal intensity files in IDAT format from the TCGA website (http://tcga-data.nci.nih.gov), created a sample annotation file that contains the available patient data – including IDH1 mutation status – and then launched RnBeads. The software identifies the data directory and input file format from the annotation file and normalizes the raw intensity data using SWAN (Makismovic et al., 2012) (other normalization algorithms are supported as well, as described in the Online Methods). CpG-specific DNA methylation levels are obtained from the normalized data and collected in an *RnBSet* object that is the basis for all subsequent analyses. During quality control, RnBeads performs clustering of all samples based on genotype fingerprinting probes included on the Infinium microarray (Supplementary Figure 1a), which is an effective method for identifying sample mix-ups and duplications. Here, we identified two samples with identical SNP patterns, in concordance with their TCGA annotation as primary and recurrent tumors from the same patient. All other samples were taken from genetically unrelated patients. RnBeads provides flexible features for data filtering as part of the preprocessing module (Supplementary Figure 1b), which are useful for excluding measurements that could bias the analysis (e.g., due to low signal quality, overlap with SNPs, or X-chromosome association in case of different sex ratios between cases and controls).

Based on the filtered and quality-controlled dataset, RnBeads performs hierarchical clustering to facilitate data exploration and outlier detection. In the clustered heatmap, we observe a small and distinct group of samples with increased promoter hypermethylation suggestive of the G-CIMP+ subtype (Supplementary Figure 1c). These putative G-CIMP+ samples indeed exhibit the characteristic enrichment of IDH1 mutations and a clear separation with respect to their global DNA methylation levels – patterns that are particularly evident from a low-dimensional projection of the entire dataset that has been annotated with IDH1 mutation status and G-CIMP subtype information (Supplementary Figure 1d). The significance of this association is also confirmed by pairwise statistical tests for associations that RnBeads performs between all sample annotations (Supplementary Figure 1e). Furthermore, RnBeads calculates groupwise comparisons between the mean DNA methylation levels in the G-CIMP positive versus negative samples for CpG islands and for genome-wide tiling regions (Supplementary Figure 1f). The resulting scatterplots show that the gain of DNA methylation among the G-CIMP+ samples is more pronounced in CpG islands than in genomic regions exhibiting low CpG content.

These automated, exploratory analyses provide a starting point for dissecting the patterns and mechanisms of epigenetic deregulation that may affect DNA methylation in G-CIMP+ tumors. Follow-up analyses can be performed directly in R, most conveniently by using the precalculated *RnBSet* data object that RnBeads prepares as part of the initial analysis. Furthermore, RnBeads makes it easy to export the data and results in a variety of formats and to hand them over to stand-alone or web-based bioinformatic tools for further analysis.

Supplementary Tables

Supplementary Table 1: Comparison between software tools for DNA methylation analysis

<Large table available as a separate file>

Supplementary Table 2: Performance benchmark for large DNA methylation analyses with RnBeads

Data type[1]	No. of Samples[2]	No. of CpGs[3]	No. of Annotations[4]	No. of Comparisons[5]	Runtime (node)[6]	Runtime (cluster)[7]
Infinium 450k	100	482,421	2	2	2h 12min	1h 9min
Infinium 450k	500	482,421	6	6	15h 2min	7h 29min
Infinium 450k	1000	482,421	10	10	1d 13h 51min	20h 15min
Infinium 450k	4034*	482,421	5	18	9d 7h 21min	6d 18h 40min
RRBS	10	1,804,103	2	2	1h 56min	49min
RRBS	50	2,169,859	6	6	5h 32min	1h 54min
RRBS	100	2,221,920	10	10	10h 13min	2h 57min
RRBS	216*	2,295,083	7	11	1d 8h 50min	14h 27min
WGBS	5	28,132,494	2	2	20h 43min	8h 23min
WGBS	10	28,150,344	6	6	2d 10h 23min	20h 5min
WGBS	20	28,154,125	10	10	4d 12h 21min	1d 15h 34min
WGBS	41*	28,158,385	5	6	3d 4h 54min	1d 9h 27min

[1] Data from the following sources were included in the analysis: TCGA (Infinium 450k), ENCODE (RRBS), Ziller et al. (WGBS)
[2] Subsets of the full datasets were randomly generated in order to assess the effect of sample size on runtime
[3] Number of CpG sites present in at least one sample. For RRBS/WGBS, low-coverage sites are removed prior to counting
[4] Adding more columns to the sample annotation table increases the complexity and runtime of the analysis
[5] Including more pairwise comparisons in the analysis strongly increases runtime but can be parallelized effectively
[6] Serial runtime measured on a scientific computing cluster (16 nodes), summing up the runtime of all contributing nodes
[7] Parallel runtime / time to completion on a scientific computing cluster (16 nodes) with optimal use of job parallelization
* The analysis results for the full datasets are available as part of the Methylome Resource on the RnBeads website