

Merging Gene Expression and Methylation Data

Gene expression profiling data can be integrated with DNA methylation data in BeadStudio, providing researchers a powerful approach to studying gene expression regulation.

INTRODUCTION

DNA methylation plays an essential role in the regulation of gene expression during development and differentiation, and in diseases such as multiple sclerosis, diabetes, schizophrenia, and cancer. Understanding the nature of complex gene regulation in cells will be greatly facilitated by the ability to assess epigenetic information for a large number of genes across the entire genome. More widespread analyses should also shed light on the epigenetic mechanism of interactions between cells and the environment.

Illumina offers a broad portfolio of products for gene expression profiling and methylation analysis. The BeadStudio analysis software is designed to facilitate an integrated data analysis, allowing users to combine data from methylation and gene expression products. The Illumina BeadStudio Methylation (M) Module is a powerful software tool to analyze data produced using Illumina Methylation analysis products. Gene expression (GX) data can also be imported into the M Module, thus enabling side-by-side comparison of methylation and transcript levels in a given sample.

INTEGRATED DATA EXPLORATION

When gene expression data are successfully imported into the Methylation Module, a new combined table is created that displays both the AVG_Beta methylation values and AVG_Signal expression values for a given sample (Figure 1). In addition, new columns are created to display statistical information that aid in the interpretation of the combined dataset. These column headings are described in Table 1.

Sorting the Combined Methylation and Gene Expression Analysis Table based on these columns can quickly identify genes that have high correlations (or anti-correlations) between methylation and expression levels. These can be viewed graphically as a line plot (Figure 2) or scatterplot (Figure 3) by using the respective functions

FIGURE 1: EXAMPLE OF A COMBINED GENE EXPRESSION AND METHYLATION TABLE

TargetID	R Pearson	R2	Rs Spearman	AVG_Beta	AVG_Signal
MLF1_E243_F	-0.9977	0.9954	-0.9667	0.16471	25171
MLF1_P97_F	-0.9568	0.9155	-0.7833	0.17984	25171
MLH1_P381_F	-0.7052	0.4972	-0.6333	0.01492	11365
MLL3_E93_R	-0.8871	0.7869	-0.3667	0.95143	368
MLL4_P1400_F	-0.5198	0.2702	-0.3667	0.03385	10760
MLL6_P957_F	0.4020	0.1616	0.2667	0.09348	2920
MMP1_P397_R	-0.2376	0.0565	-0.4500	0.08386	2056
MMP1_P460_F	-0.2499	0.0625	-0.6000	0.12077	2056
MMP10_E136_R	0.5026	0.2526	0.1633	0.04322	949
MMP14_P13_F	-0.9850	0.9703	-0.9500	0.95007	7474

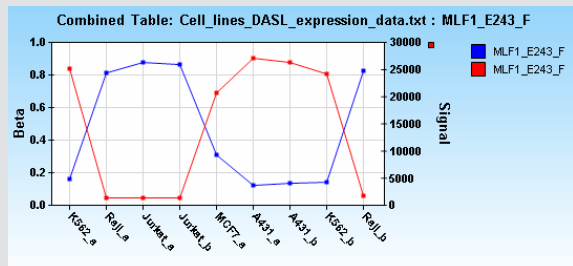
A screenshot showing a combined GX/Methylation Table in the BeadStudio M Module.

in BeadStudio. Click the button on the icon bar for a lineplot or the button for a scatterplot.

MERGING GENE EXPRESSION AND METHYLATION DATA

It is possible to merge any methylation data set with any gene expression data set using BeadStudio. Two files—a

FIGURE 2: LINE PLOT OF COMBINED TABLE



A line plot of MLF1 gene expression (red) signal intensity and methylation beta (blue) across several cell lines. Methylation and gene expression levels are inversely correlated for this gene.

TABLE 1: COMBINED DATASET COLUMN HEADINGS

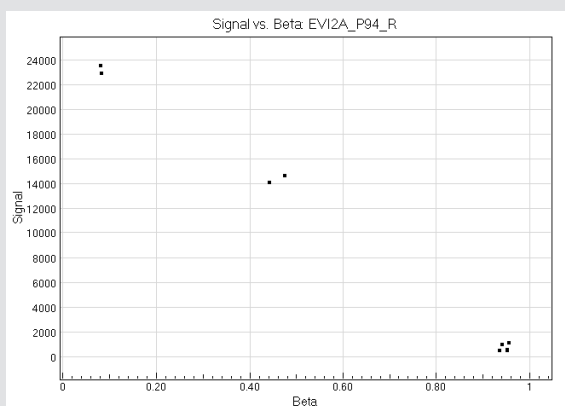
HEADING	DEFINITION	RELEVANCE
R Pearson	Pearson's correlation coefficient (r) reflects the degree of linear relationship between methylation beta and gene expression signal intensity. Pearson's correlation coefficient assumes a linear relationship between variables.	Correlation of +1 means that there is a perfect positive linear relationship between these variables, -1 means a perfect negative (inverse) relationship, and 0 means the complete absence of a relationship.
R2	R ² (the coefficient of determination or r squared) measures the association between beta and signal.	R2 can be interpreted as the proportion of variance in beta that is contained in the expression signal. 0 means no relationship; 1 means perfect correlation.
Rs Spearman	Spearman's rank correlation coefficient gives an estimate of monotone association between beta and signal and provides a measure of how closely the two sets of rankings agree with each other.	This coefficient should be used when the distribution of the data makes Pearson's correlation coefficient invalid (e.g., in the case of non-linear association). Linearity can be visualized with a scatterplot.

GX Data Table and a Gene Lookup Table—are required. Both of these files must be tab-delimited and saved as a plain-text file format (*.txt). All fields and matches are case insensitive.

GX Data Table

This file contains the GX intensity values to import into the Methylation Module (Figure 4).

FIGURE 3: SCATTERPLOT OF COMBINED TABLE



Each sample is plotted by its expression signal versus methylation beta for a particular gene. This gene exhibits a negative linear correlation between gene expression levels and methylation beta value.

Gene LookUp Table

This file contains the cross-mapping information between the GX and Methylation probes (Figure 5).

MAINTAINING DATA TABLES WHEN USING MICROSOFT EXCEL

The best method for creating or editing these data table files is by using Microsoft Excel, and saving the table as a **Text (Tab delimited) (*.txt)** file. An essential caveat is that tab-delimited tables exported from BeadStudio must be imported into Excel in a specific manner to prevent corruption.

1. In BeadStudio, click the **Export Displayed Data to File** button.
2. Save the data table as a *.txt file.
3. In Microsoft Excel, select **File | Open**.
4. In the Files of Type area, select **Text Files (*.prn; *.txt; *.csv)**.
5. Browse to, select, and open the *.txt file you exported from BeadStudio.
6. In Steps 1 and 2 of the Text Import Wizard, click **Next**.
7. In Step 3 of the Text Import Wizard, in the Data Preview area, shift-click to select all columns.
8. In the Column Data Format area, select **Text**.
9. Click **Finish**.

FIGURE 4: EXAMPLE OF A GX DATA TABLE

	A	B	C
1	TargetID	K562_a.AVG_SIGNAL	Jurkat_a.AVG_SIGNAL
2	15E1.2	252.3533	176.8368
3	2-PDE	269.7352	158.4017
4	76P	241.8085	220.6028
5	7A5	-3.841293	-3.284624
6	A1BG	-1.7452	4.899265
7	A2BP1	1.327652	4.203279
8	A2BP1	1.327652	4.203279
9	A2M	-5.313865	-3.323264
10	A2ML1	5.831338	2.402984
11	A3GALT2	8.342595	8.21016
12	A4GALT	114.346	112.329
13	A4GNT	7.393178	17.53191
14	AAA1	4.518675	9.116126
15	AAAS	107.7798	320.9568
16	AACS	385.1862	228.6798
17	AACSL	12.00363	4.446897
18	AADAC	4.736413	0.9942682
19	AADACL1	229.8904	412.8228
20	AADACL2	-1.915843	-2.923774

Part of a GX data table file is displayed in Excel. Row 1 contains the column header names. GX TargetID values are listed in Column A, and Columns B and C list the associated gene expression AVG_SIGNAL values for two different samples.

GX DATA TABLE

GX Data Table Requirements

- Intensity data column headers in the GX data file must take the form 'samplename.AVG_SIGNAL' (e.g., K562_a.AVG_SIGNAL). If AVG_SIGNAL is not at the end of the column name, the sample will not be imported.
- The sample name must exactly match the name in the methylation project.
- Not all samples in the methylation project need to have a corresponding sample in the GX data file. There can be samples in the GX data table that are not in the methylation project. These are ignored and not imported.
- There must be a column named TargetID. This is used by the lookup table.
- There can be additional columns (e.g., annotation), but these are not imported into the methylation project.

Creating a GX Data Table

The following steps describe a method for creating a GX data table based on the software requirements:

- Use the Sample Gene Profile table within the GX Module to produce the GX data table. This will ensure that there are no duplicated TargetIDs in the table. It is possible to use the Sample Probe Profile table, but this will result in non-unique TargetIDs and may produce

unintended results when the intensity values are imported into the Methylation Module.

- Export the data to a text file, ensuring that the TargetID and AVG_Signal columns are exported.

GENE LOOKUP TABLE

Lookup Table Requirements

The lookup table file consists of two required columns with a header name row. Column 1 should be named GENEEXPRESSION, and field values in this column correspond to the TargetID column in the GX data table. Column 2 should be named METHYLATION, and fields in this column correspond to the SYMBOL column in the methylation project.

- Each value in the GENEEXPRESSION column must be unique.
- METHYLATION column values should be unique. If there are more than one GX probe for one methylation symbol, a single GX probe should be chosen to include in the lookup table.
- Extra fields in the GENEEXPRESSION column that do not correspond to a field in the TargetID column of the GX data table are allowed, but are ignored by BeadStudio.
- Extra fields in the METHYLATION column that do not correspond to symbols in the methylation project are allowed, but are ignored by BeadStudio.

FIGURE 5: EXAMPLE OF A LOOKUP TABLE

	A	B
1	GENEEXPRESSION	METHYLATION
2	AATK	AATK
3	ABCA1	ABCA1
4	ABCB1	ABCB1
5	ABCB4	ABCB4
6	ABCC2	ABCC2
7	ABCC5	ABCC5
8	ABCG2	ABCG2
9	ABL1	ABL1
10	ABL2	ABL2
11	ABO	ABO
12	ACTG2	ACTG2
13	ACVR1	ACVR1
14	ACVR1B	ACVR1B
15	ACVR1C	ACVR1C
16	ACVR2B	ACVR2B
17	ADAMTS12	ADAMTS12
18	ADCYAP1	ADCYAP1
19	AFF3	AFF3
20	AFP	AFP
21	AGTR1	AGTR1

A lookup table file is displayed in Excel. Row 1 contains the column header names. GX TargetID values are listed in Column A, and Column B lists the associated Methylation project symbol values.

Creating a Lookup Table

The latest human whole genome expression arrays have a TargetID in the form of a gene symbol. When a lookup table is created for combining this type of whole genome GX data set and a custom methylation data set, the values in both columns of the lookup table are identical (Figure 5). Even though the entries are the same, a lookup file is still required. To create such a lookup table:

1. Export all the gene symbols from a methylation project.
2. Remove duplicate entries from the gene symbol list using Excel.
3. Rename the column header METHYLATION.
4. Duplicate the METHYLATION column in column B.
5. Rename Column A to form the GENEEXPRESSION column.

When the GX data file TargetIDs are not in the same format as the Symbols in the Methylation project, the two columns in the lookup table will be different. In this case, users must determine which GX TargetID values correspond to which symbols in the methylation project. Each pair of corresponding TargetID and Symbol should be listed in adjacent cells in a single row of the table.

SUMMARY

Illumina offers a broad portfolio of tools for genetic analysis, including gene expression and DNA methylation profiling. BeadStudio analysis software provides powerful methods for analysis and interpretation of results derived from Illumina products. Moreover, BeadStudio enables the integrated analysis of methylation and gene expression profiling by supporting the combination of data from separate assay products. As described in this technical note, this combination requires minimal data handling and preparation by researchers. This integrative approach provides a robust and powerful method to examine whole-genome gene expression, regulation, and epigenetics.

ADDITIONAL INFORMATION

Visit our website or contact us at the address below to learn more about Illumina DNA Methylation and RNA Analysis Solutions.

Illumina, Inc.
Customer Solutions
9885 Towne Centre Drive
San Diego, CA 92121-1975
1.800.809.4566 (toll free)
1.858.202.4566 (outside the U.S.)
techsupport@illumina.com
www.illumina.com

FOR RESEARCH USE ONLY

© 2008 Illumina, Inc. All rights reserved.
Illumina, Solexa, Making Sense Out of Life, Oligator, Sentrix, GoldenGate, DASL, BeadArray, Array of Arrays, Infinium, BeadXpress, VeraCode, IntelliHyb, iSelect, and CSeq are registered trademarks or trademarks of Illumina. All other brands and names contained herein are the property of their respective owners.
Pub. No. 470-2007-006 Current as of 16 January 2008

