Differntial Methylation Analysis of H460 Parent and H460 Knock-in cell lines on 2.1M Nimblegen Array

Nitesh Turaga

January 23, 2014

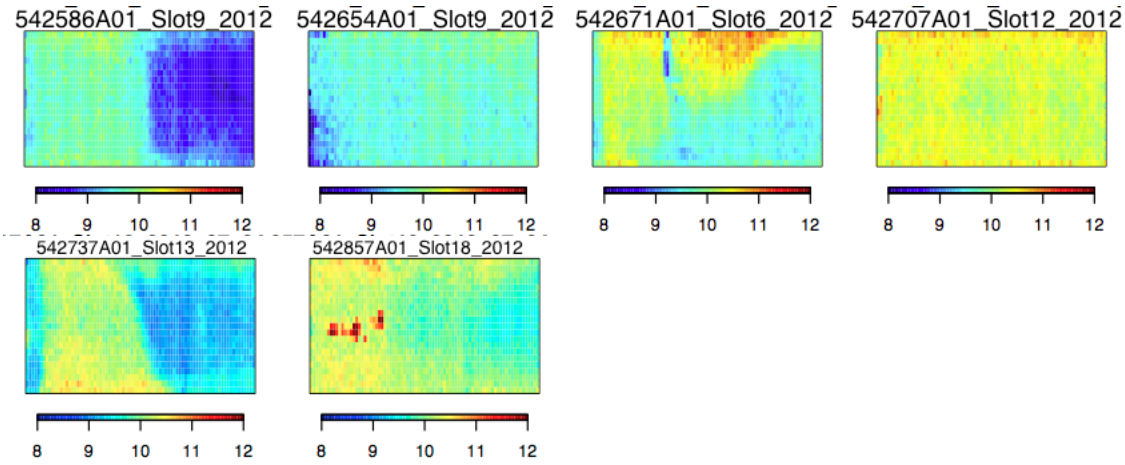
To compare the differential methylation regions of H460 Knock-in cell line and H460 parent cell line.The data was generated from Nimblegen 2.1M oligonucleotide microarrays. The raw Nimblegen data were `.tif` images. There were a total of 6 samples, 3 for each cell line(H460 knock-in and H460 parent).Array images were processed with DEVA-v1.2 (Nimblegen software for automated feature extraction and data analysis). The TIF files were processed and converted to **XYS** files for analysis. These files report the, **X** - coordinate of the feature on the image, **Y**-coordinate of the feature on the image, and the **S**ignal - the flourescence intensity of the pixels that make the feature. The TIF files were also processed with DEVA using the DNA methylation work flow to identify peaks and generate a result for each sample.

Preliminary Assessments

General QC-analyses brought to light that the signal intensities and local enrichment at methylated sites were not large. The data quality was assessed by looking at the Enriched channel in the MeDIP array, where we expect every probe to have a signal. Since, the enriched channel has methylated DNA, a successful hybridization would indicate a signal. The array signal is calculated as the average percentile rank of the signal probes among the background probes. The score ranges between 0 to 100, where 100 indicates the ideal scenario or perfect hybridization. This quality score is calculated before any kind of normalization is done on the arrays.

Table 1: Array quality scores

Sample ID	Status	Quality Score
542586A01_Slot9_2012-07-24_H460	H460_parent	59.6
542737A01_Slot13_2012-07-24_H460	H460_parent	64.4
542857A01_Slot18_2012-07-24_H460	H460_parent	69.8
542671A01_Slot6_2012-07-24_H460	H460_knockin	71.2
542654A01_Slot9_2012-07-25_H460	H460_knockin	74.9
542707A01_Slot12_2012-07-25_H460	H460_knockin	76.5



As we can in Table 1, the quality score of the H460 Parent arrays are pretty low.

The CHARM Algorithm

CHARM is specifically designed to maximize the number of assayed CpGs .This array design improves the detection strategy because it facilitates a smoothing strategy and assays many more CpGs. The basic measurement used to quantify methylation is the log-ratio of the intensities observed in the treated and control channels. To detect methylated regions in the CHARM method, the M-values were normalized and processed using genome-weighted smoothing.(5)

The normalization method uses genome sequence information and knowledge of the fragment selection method to select pseudo-housekeeping probes for which one can in fact assume M = 0. Then apply the Loess normalization procedure developed for expression arrays to the pseudo-housekeeping genes, obtain the correction curve, and use this curve to correct M-values for all probes. To obtain a smoothed M-value at any given genomic location, average all the M-values that were within a prespecified distance from the location in question. The interval providing the values that are averaged is referred to as the smoothing window and its length is referred to as the window size. (5)

After estimating the DNA methylation values in terms of percentage methylation, we use the regression based DMR-finding approach after correcting for batch effects, which is provided in the `charm-package`. This fails to find DMRs in the current samples.

The Bumphunter Algorithm

Bumphunter used to estimate regions for which the genomic profile deviates from a baseline value(cut off). It is implemented to detect differntially methylated genomic regions between two populations. It is also a regression based approach.It reports the result, with a table of candidate differentially methylated regions and the corresponding annotation for wach region.

The Nimblegen Algorithm

The standard Nimblegen algorithms were used to compute the normalized data and identify peaks of enrichment, coinciding with methylated regions. Peaks near known transcription start sites

(TSS) were identified, with different cutoffs between maximal distance between a peak and TSS. In each cell line, the genes which were intersecting among all the samples were identified. All the genes which were identified jointly in both cell lines were removed, making each gene set unique and exclusive for a specific cell line. Both sets of unique and exclusive genes, were ordered by distance from the Transcription start site. 944 genes were identified in H460 knock-in cell line and 349 in H460 parent. (Table 2)

Table 2: Number of features in each cell line

Cell line	Feature Track	Number of features
H460 Knock in	transcription start site	944
H460 Parent	transcription start site	349

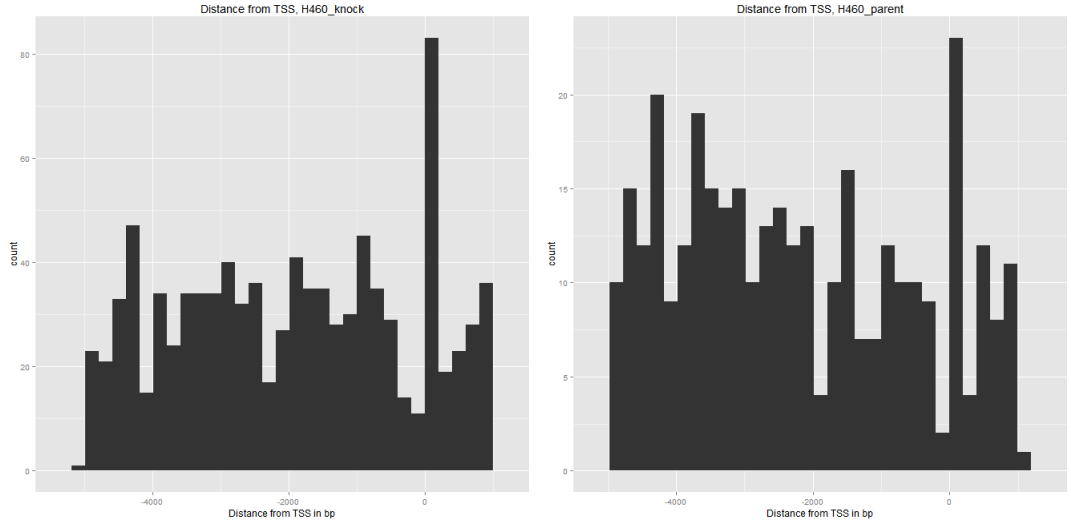
NOTE: Files are attached as knock-in_genes_distFromTSS.csv and parent_genes_distFromTSS.csv. The regions are ordered in a decreasing order based on the distance from TSS.

The only features left in the result are transcription start sites, other possible feature is CpG Island (column name is FEATURE TRACK). The distribution of the distance from the TSS is also shown for both cell lines. Column name is SHORTEST_DISTANCE_FROM_FEATURE_TO_DATA_POINT.

Table 3: Summary of Feature distance from data point

Cell line	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
H460 Knock in	-4998.0	-3363.0	-1916.0	-1956.0	-569.8	995.0
H460 Parent	-4975	-3669	-2474	-2223	-718	997

Table 4: Distribution of Shortest distance from feature to data point



(a) H460 knock-in distance from TSS **(b)** H460 parent distance from TSS

Pathway Comparisons

Both the list of H460 Knock-in genes and H460 parent genes were compared with a list of selected pathways.

1. Cell Cycle
2. Complete Homeobox (HOX) Genes
3. Complete Human Inflammatory Response and Autoimmunity
4. Complete Human Tumor Suppressor Genes
5. Complete Stem Cell Transcription Factors
6. Complete Stress and Toxicity
7. Cytokine Production
8. DNA Repair
9. Human Epithelial to Mesenchymal Transition (EMT)
10. Human Notch Signaling Pathway
11. Human T-Cell B-Cell Activation Methylation
12. Human T Helper Cell Differentiation
13. Human Tumor Suppressor Genes
14. Inflammatory Response and Autoimmunity
15. Polycomb and Trithorax Complexes
16. Stem Cell Transcription Factors
17. TGF f BMP Signaling Pathway
18. Toll-Like Receptor Signaling Pathway
19. WNT Signaling Pathway

Very few pathway genes matched with the H460 knock-in genes and the H460 parent genes.
The H460 knock-in genes have only 29 genes in common with the pathways(listed in Table 5). The following result from shows you the pathways and the genes along with gene description. The result, for the names of the genes is stored in "pathway_genes_knock.csv".

Table 5: H460 Knock-in gene with corresponding pathways

Gene name	Description	Pathway
BRCA1	breast cancer 1, early onset	Cell_Cycle
HOXB2	homeobox B2	Complete_Homeobox_(HOX)_Genes
HOXB3	homeobox B3	Complete_Homeobox_(HOX)_Genes
HOXB4	homeobox B4	Complete_Homeobox_(HOX)_Genes
HOXB6	homeobox B6	Complete_Homeobox_(HOX)_Genes
HOXB7	homeobox B7	Complete_Homeobox_(HOX)_Genes
HOXB8	homeobox B8	Complete_Homeobox_(HOX)_Genes
BRCA1	breast cancer 1, early onset	Complete_Human_Tumor_Suppressor_Genes
DIRAS3	DIRAS family, GTP-binding RAS-like 3	Complete_Human_Tumor_Suppressor_Genes
XRCC1	X-ray repair complementing defective repair in Chinese hamster cells 1	Complete_Human_Tumor_Suppressor_Genes
GATA6	GATA binding protein 6	Complete_Stem_Cell_Transcription_Factors
HOXB3	homeobox B3	Complete_Stem_Cell_Transcription_Factors
HOXB8	homeobox B8	Complete_Stem_Cell_Transcription_Factors
SOX9	SRY (sex determining region Y)-box 9	Complete_Stem_Cell_Transcription_Factors
STAT3	signal transducer and activator of transcription 3 (acute-phase response factor)	Complete_Stem_Cell_Transcription_Factors
ATF4	activating transcription factor 4 (tax-responsive enhancer element B67)	Complete_Stress_&_Toxicity
BRCA1	breast cancer 1, early onset	Complete_Stress_&_Toxicity
ERCC1	excision repair cross-complementing rodent repair deficiency, complementation group 1	Complete_Stress_&_Toxicity
GPX7	glutathione peroxidase 7	Complete_Stress_&_Toxicity
SMC1A	structural maintenance ofchromosomes 1A	Complete_Stress_&_Toxicity
XRCC1	X-ray repair complementing defective repair in Chinese hamster cells 1	Complete_Stress_&_Toxicity
ELANE	elastase, neutrophil expressed	Cytokine_Production
BRCA1	breast cancer 1, early onset	DNA_Repair
XRCC1	X-ray repair complementing defective repair in Chinese hamster cells 1	DNA_Repair
PSENEN	presenilin enhancer 2 homolog (C. elegans)	Human_Notch_Signaling_Pathway
BRCA1	breast cancer 1, early onset	Human_Tumor_Suppressor_Genes
CBX2	chromobox homolog 2 (Pc class homolog, Drosophila)	Polycomb_&_Trithorax_Complexes
STAT3	signal transducer and activator of transcription 3 (acute-phase response factor)	Stem_Cell_Transcription_Factors
IRAK1	interleukin-1 receptor-associated kinase 1	Toll-Like_Receptor_Signaling_Pathway

The H460 parent genes have only 2 genes in two pathways(Table 6) , and stored in "pathways_genes_parent.csv".

Table 6: H460 parent gene with corresponding pathways

Gene name	Description	Pathway
AR	androgen receptor	Complete_Stem_Cell_Transcription_Factors
AR	androgen receptor	Stem_Cell_Transcription_Factors

DEVA was also run with a changed cutoff of -5000 to +1500 to produce peak files. A similar pipeline was run as with the new peak files as described above for pathway analysis. The results were different as expected because of a bigger window size.

Table 7: Number of features in each cell line with extended window

Cell line	Feature Track	Number of features
H460 Knock in	transcription start site	3877
H460 Parent	transcription start site	2639

The pathway analysis showed, in the H460 Knock-in gene set 189 genes were observed in 18 pathways and in the H460 parent gene set 70 genes were observed in 19 pathways.
The links are provided for both the tables below.

1. H460 knock-in genes with corresponding pathways, with changed cutoff of -5000 to +1500.(https://app.box.com/s/7xr01qkv3xx41oo5izew)
2. H460 parent genes with corresponding pathways, with changed cutoff of -5000 to +1500.(https://app.box.com/s/pu0n89gusqlf1umjwwhh)

R-packages Used

List of R-packages used for analysis:

1. Charm
2. Bioconductor
3. BiocGenerics
4. plyr
5. RCircos

References

1. Aryee MJ et al., Accurate genome-scale percentage DNA methylation estimates from microarray data, Biostatistics (2011) 12(2): 197-210
2. Seth Falcon, Benilton Carvalho with contributions by Vince Carey, Matt Settles and Kristof de Beuf. pdInfoBuilder: Platform Design Information Package Builder. R package version 1.24.0.
3. Rafael A. Irizarry, Martin Aryee, Hector Corrada Bravo, Kasper D. Hansen and Harris A. Jaffee (). bumphunter: Bump Hunter. R package version 1.0.0.
4. RCircos: an R package for Circos 2D track plots
5. Irizarry RA, Ladd-Acosta C, Carvalho B, et al. Comprehensive high-throughput arrays for relative methylation (CHARM) Genome Res. 2008;18(5):780790.
6. Martin J. Aryee, Zhijin Wu, Christine Ladd-Acosta, Brian Herb, Andrew P. Feinberg, Srinivasan Yegnasubramanian, and Rafael A. Irizarry. Accurate genome-scale percentage dna methylation estimates from microarray data. Biostatistics, 12(2):197-210, 2011.