Shyam Biswal : MeDIP on 2.1M Nimblegen Array

Nitesh Turaga

January 16, 2014

Data Description

Data from Shyam Biswal, used to compare the differential methylation regions of H460 Knock-in cell line and H460 parent cell line.

The data was generated from two color Nimblegen 2.1M oligonucleotide microarrays. The raw Nimblegen data was in the form of `.tif` images and the corresponding array design files were given for building the annotation package `100929_HG19_Deluxe_Prom_Meth_HX1` by Nimblegen.

There were a total of 6 samples, 3 for each cell line(H460 knock-in and H460 parent). The 2.1M Platform refers to the high-density Roche NimbleGen 2.1 million-feature arrays, i.e, 2.1M probes on each array

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. `.xys` 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.

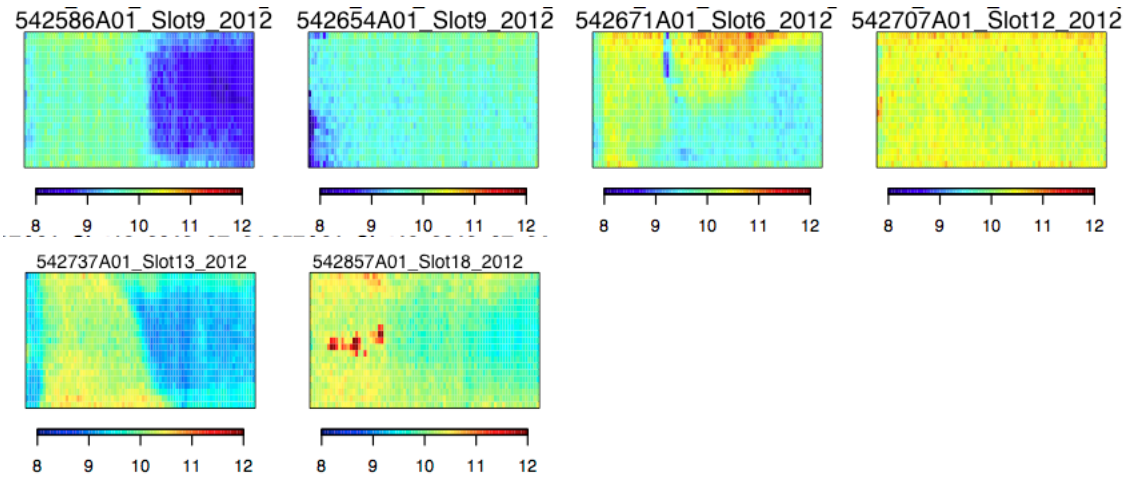
The TIF files were also processed with DEVA using the DNA methylation work flow to identify peaks and generate a result for each sample.

Preliminary Assessments

General QC-analyses brought to light that the signal intensities and local enrichment at methylated sites were not really large. Initial array quality assessment was done using the Bioconductor `charm-package` on the `.xys` files. The data quality can be assessed by looking at the Enriched channel in the MeDIP array, where we expect every probe to have a signal. Since, the enriched channel contains the only methylated DNA, a successful hybridization would indicate a strong 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.

Sample ID	Status	Quality Score
542586A01_Slot9_2012-07-24_H460	H460_parent	59.6804601609024
542654A01_Slot9_2012-07-25_H460	H460_knockin	74.9611413105766
542671A01_Slot6_2012-07-24_H460	H460_knockin	71.2294097946914
542707A01_Slot12_2012-07-25_H460	H460_knockin	76.519565668383
542737A01_Slot13_2012-07-24_H460	H460_parent	64.4412397468657
542857A01_Slot18_2012-07-24_H460	H460_parent	69.8646821319654

Table 1: Array quality scores



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

The CHARM Algorithm

Array quality scores were generated with the Bioconductor `charm - package`, and data quality was checked using the *Enriched* channel. Four of the six arrays show a large standard deviation in the signal strenth and seem to have a problem in hybridization.

To estimate the DNA methylation values, the background signal is removed before computing the log-ratios. A within sample normalization method - **loess** is used, and then a between array normalization - quantile is used. The control probes which are the non-CpG probes are excluded. This provides us a probe-level estimate of DNA methylation from the 2.1M oligonucliotide microarray.

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 by `charm`. This fails to find DMRs in the samples. Also, while not taking into account batch effects (surrogate variables), `charm` does not find any DMR's.

The Bumphunter Algorithm

Bumphunter is 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.

Analysis using Bumphunter

1. Run 1

   Initially ran `bumphunter` on the data, with cutoff value 1.0, this failed to find any bumps. The sensitivity of the cutoff was not enough to catch any DMRs in the sample set. Bumphunter is better used for large sample sets, for better performance.

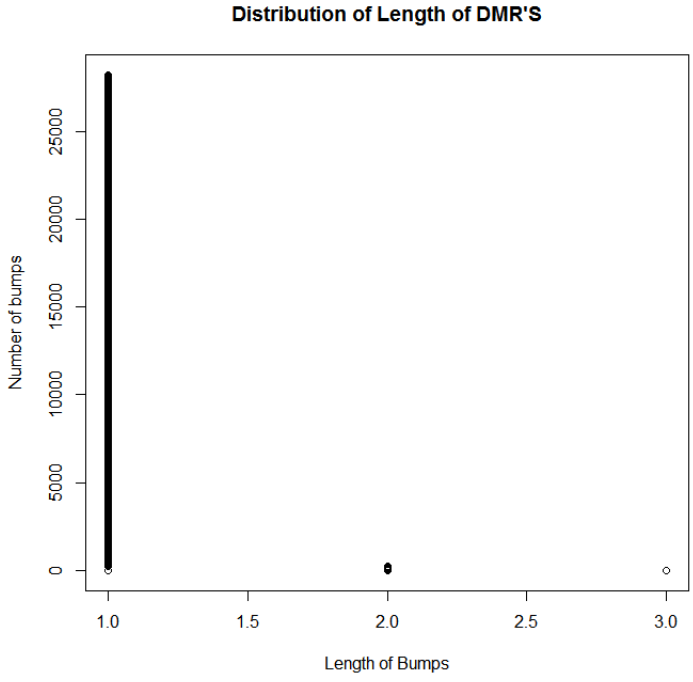
2. Run 2

   Bumphunter, with the inbuilt argument to pickCutoff was run. The cutoff chosen by bumphunter was at 0.41, bumps found 28206. The number of DMRs/bumps of length greater than four is 0. This result is not useful for analysis as the length of the DMRs are too small.

   The distribution of the length of DMRs found, is shown below. Most of them are single CpG sites of length 1. This does not work for a significant analysis.

Length(L)	1	2	3
Number of DMRs	27981	222	3

Table 2: Distribution of Bumps



The Nimblegen Algorithm

The standard Nimblegen algorithms were used to compute the normalized data and identify peaks of enrichment, coinciding with methylated regions. Next the data was transformed into a more usable format, i.e. the peaks near known transcription start sites (TSSs) were identied, according to 2 different cutoffs for the maximal distance between a peak and a TSS:

- -500 to +500, the default cutoff
- -5000 to +1500, a custom cutoff

Peaks were identified for each sample using both of these cut offs.

Analysis using Peaks files from DEVA

For each cell line, the genes which were common in each sample were identified, i.e, gene names which intersect. All the genes which were identified jointly in both cell lines were removed, making each gene set unique and exclusive to the specific cell line. Both sets of 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 1)

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

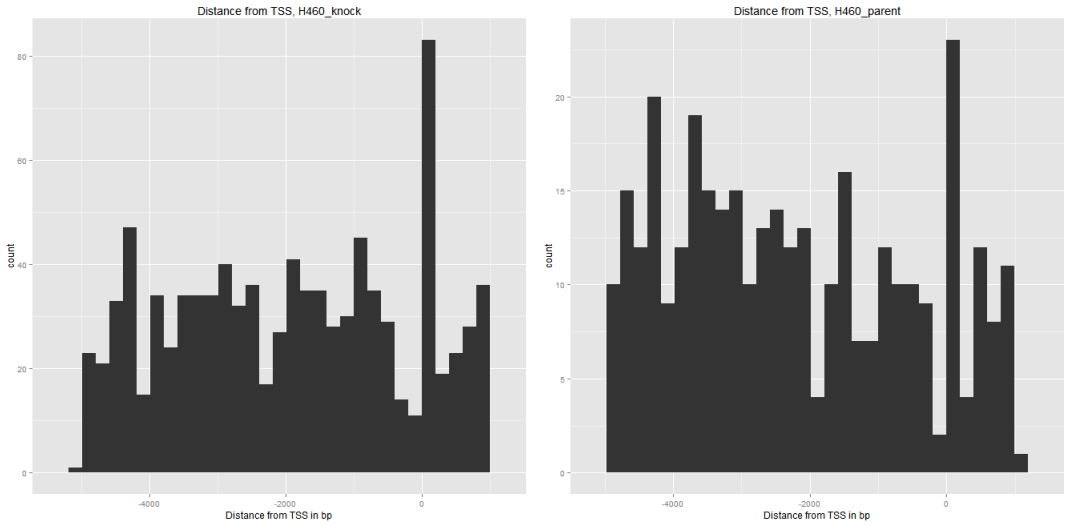
Table 3: Number of features in each cell line

`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.

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: Summary of Feature distance from data point



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

Figure 1: Distribution of Shortest distance from feature to data point

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".

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

Table 5: H460 Knock-in gene with corresponding pathways

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

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

Table 6: H460 parent gene with corresponding pathways

Plots and Results

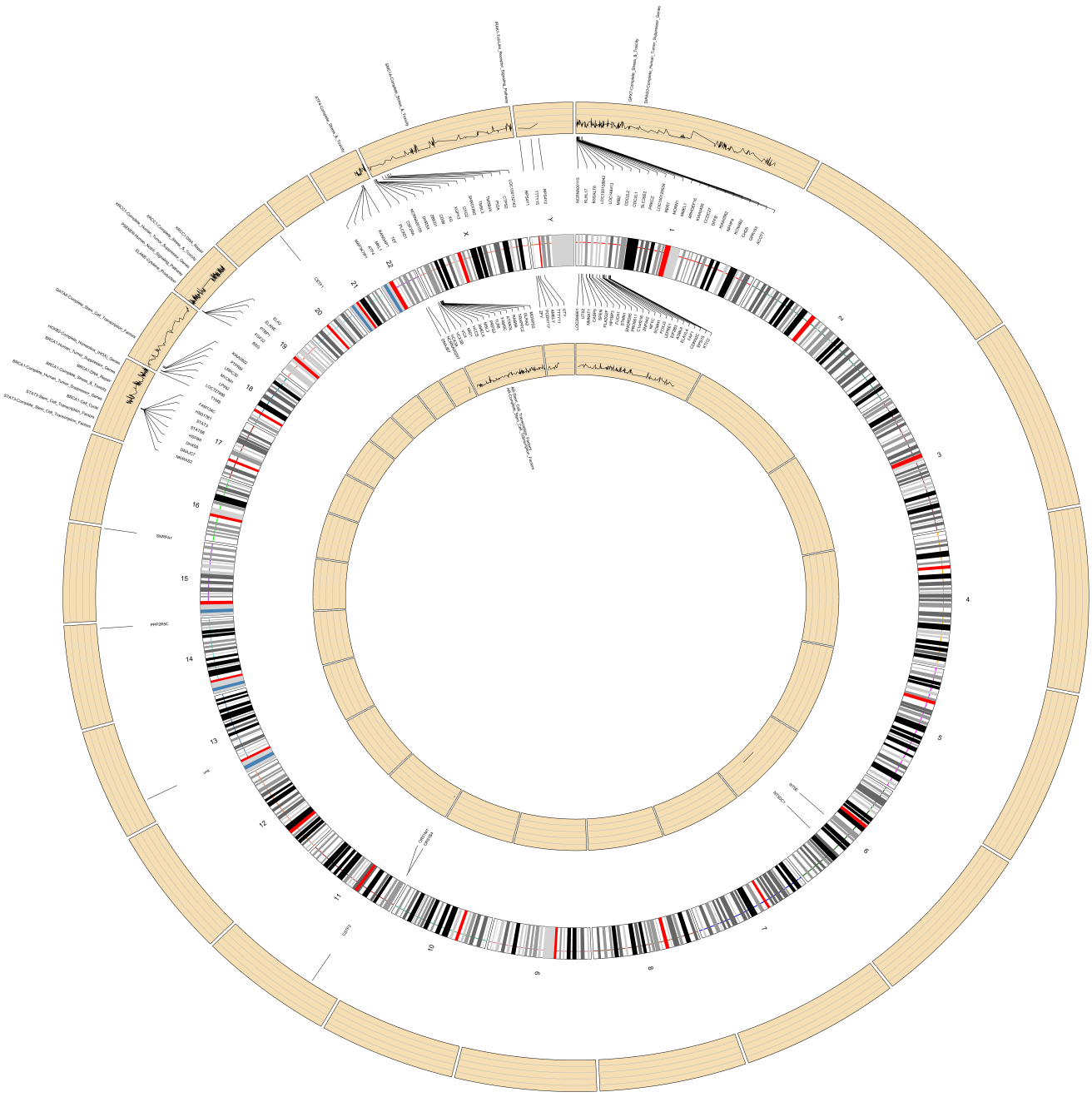
Circos plot was made using RCircos, the inner track refers to the H460 parent genes and the outer track refers to the H460 knock-in genes.

NOTE: Everything on the inside of the chromosome ideogram is the inner track. Everything on the outside of the ideogram is the outer track. Due to scaling issues, few genes have been omitted. Please refer the data files for more accurate results.

As it can be seen from the circos plot, very few chromosomes contribute to the methylated regions, including the sex chromosomes. The plot next to the gene names correspond to the peak values. It is easy to infer from this plot where there is a high frequency of methylated regions.

The pathways and the corresponding gene in that particular pathway are also plotted in the Circos plot in the inner and the outer tracks respectively.

REFER: BiswalCircosMain.png



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