

Shyam Biswal : MeDIP on 2.1M Nimblegen Array

Nitesh Turaga

January 15, 2014

Analysis

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

Data Description

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. 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. 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

Analysis using CHARM

Initial array quality assessment was done using `charm` and the `.xys` files.

Array quality scores were generated with `charm::qcReport`, and data quality was checked using the *Enriched* channel. Four of the six arrays show a large standard deviation in the signal strength and seem to have a problem in hybridization. `charm::pmQuality`, provides the array signal quality score. 3 of the samples have a signal strength above the cutoff(70).

Refer: `qcReport.pdf`

Charm fails to find DMRs both while taking into account surrogate variables(SV) and not while not accounting for SV's. This might be because the image quality is poor after hybridization with a lot of variability in 3 out of the 6 images. The case control being used is the H460_parent vs H460_knock-in. No other annotation is recorded in the experimental metrics `.xlsx` file given.

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.

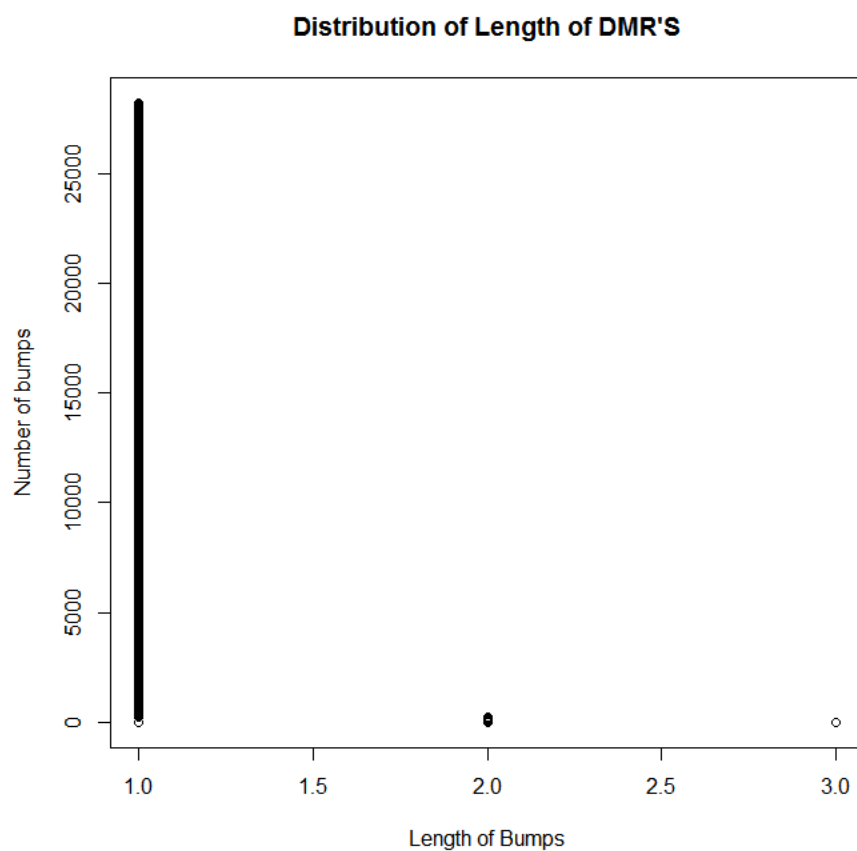
2. Run 2

Bumphunter, with the inbuilt argument to `pickCutoff` was run. The cutoff chosen by bumphunter was at 0.41, bumps found 28206. But just like other nimblegen data sets, it is unable to match genes, as the number of bumps of length greater than 4 is 0.

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

```
> table(bumps$L)
```

1	2	3
27981	222	3



Argument description for `bumphunter`,

cutoff:

A numeric value. Values of the estimate of the genomic profile above the cutoff or below the negative of the cutoff will be used as candidate regions. It is possible to give two separate values (upper and lower bounds). If one value is given, the lower bound is minus the value.

pickCutoff:

Should bumphunter attempt to pick a cutoff using the permutation distribution?

Analysis using Peaks files from DEVA

This analysis uses the Peaks files generated from DEVA-v1.2, using the standard Nimblegen algorithms to identify peaks which coincide with methylated regions.

Sequential intersections, based on the number of peaks identified in each sample, was done in decreasing order for each experimental status. This step results in all the genes within each sample which are intersecting.

We find the common genes between H460_knock-in and H460_parent", and remove these genes from the intersected lists. We then order them by distance from the transcription start site(TSS).

NOTE: Files are attached as knock-in_genes_distFromTSS.csv and parent_genes_distFromTSS.csv. The distance is ordered in decreasing order of since the Max distance is a smaller positive number.

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 status types. The window for the Distance from the TSS is measured by -5000 to +5000 , so window size is 10,000 bp (column name is SHORTEST_DISTANCE_FROM_FEATURE_TO_DATA_POINT).

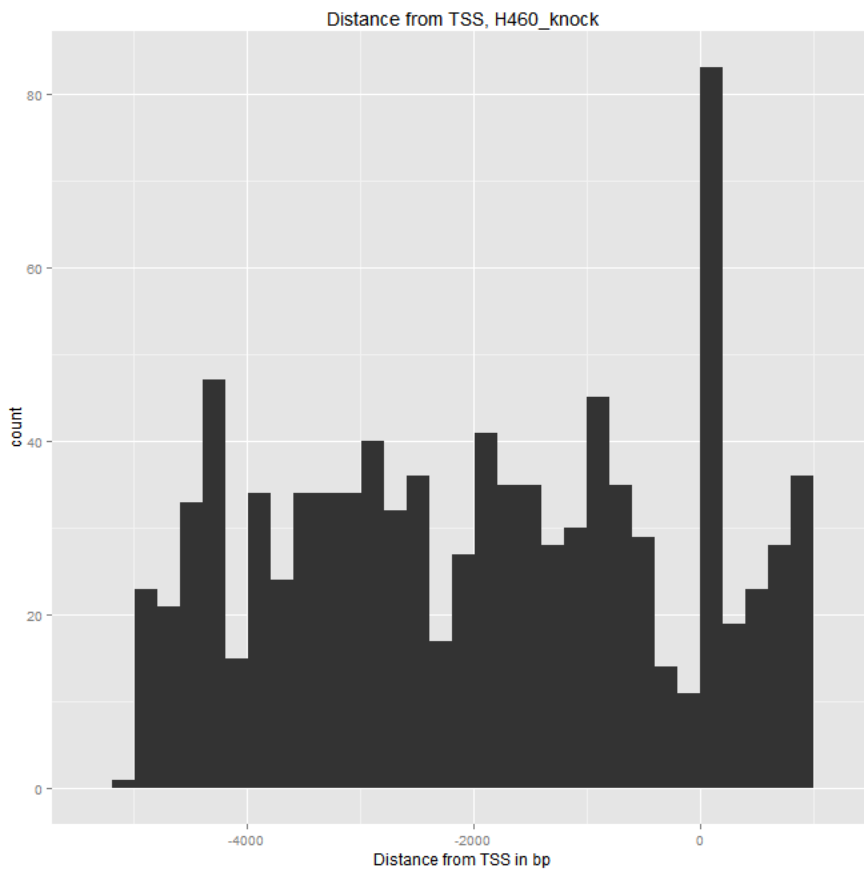
H460_knock-in

```
> table(annot_knock$FEATURE_TRACK)
```

```
transcription_start_site
                      944
```

```
> summary(annot_knock$SHORTEST_DISTANCE_FROM_FEATURE_TO_DATA_POINT)
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-4998.0	-3363.0	-1916.0	-1956.0	-569.8	995.0

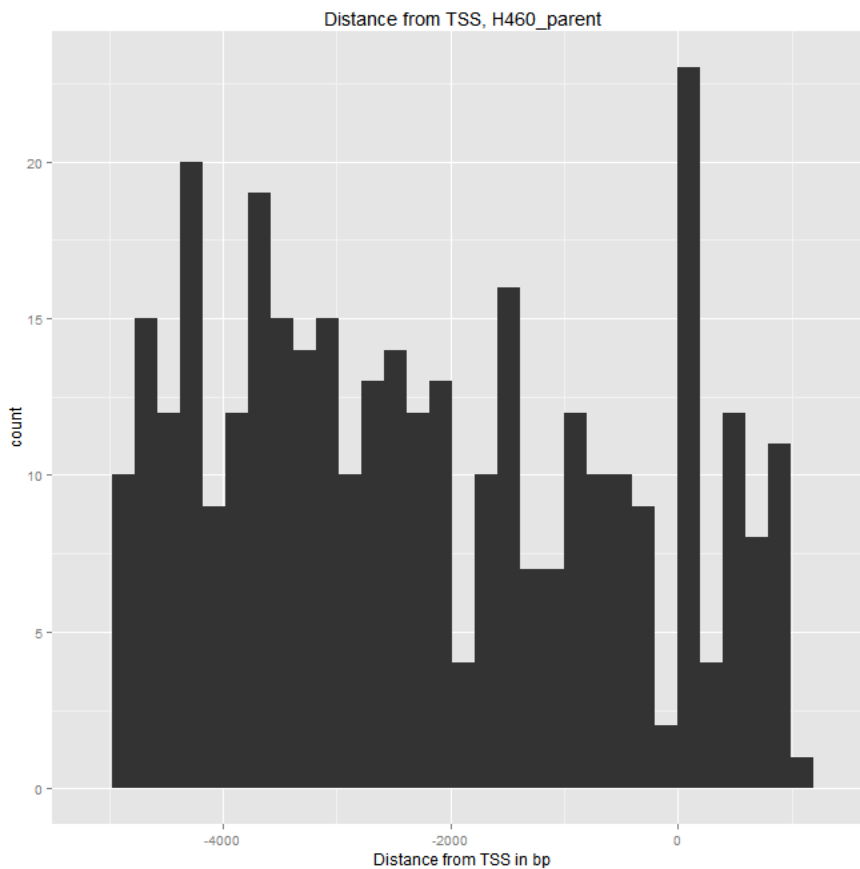


H460_parent

```
> table(annot_parent$FEATURE_TRACK)
```

```
transcription_start_site
349
```

```
> summary(annot_parent$SHORTEST_DISTANCE_FROM_FEATURE_TO_DATA_POINT)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-4975  -3669   -2474   -2223   -718     997
```



Algorithms and R-packages

List of R-packages used for analysis:

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

Extracting the percentage methylation values from the raw data was done using `charm::methp`, where the default arguments were used to normalize. The normalization methods included were spatial normalization(to correct for spatial artifacts), background subtraction(to estimate and remove the background signal before computing the log-ratios), loess within sample normalization and quantile between sample normalization.

These percentage methylation values are on a logit scale.

A regression based DMR-finding after correcting for batch effects pipeline was used in this analysis. Removing batch effects and using surrogate variables in finding DMRs (`refer sva-package`) have been shown to reduce dependence on unknown noise in the data set.

Pathway Comparisons

Both the list of Knocking genes and 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. The following result from R shows you which pathways and how many genes from that pathway. The result, for the names of the genes is stored in "pathway_genes_knock.csv".

```
> table(knock_df$pathway)
```

	Cell_Cycle	Complete_Homeobox_(HOX)_Genes
	1	6
Complete_Human_Tumor_Suppressor_Genes	Complete_Stem_Cell_Transcription_Factors	
	3	5
Complete_Stress_&_Toxicity	Cytokine_Production	
	6	1
DNA_Repair	Human_Notch_Signaling_Pathway	
	2	1
Human_Tumor_Suppressor_Genes	Polycomb_&_Trithorax_Complexes	
	1	1
Stem_Cell_Transcription_Factors	Toll-Like_Receptor_Signaling_Pathway	
	1	1

odd	odd	odd
even	even	even
odd	odd	odd
even	even	even

The H460 parent genes have only two genes in two pathways, and stored in "pathways_genes_parent.csv".

```
> table(parent_df$pathway)
```

Complete_Stem_Cell_Transcription_Factors	Stem_Cell_Transcription_Factors
1	1

Plots and Results

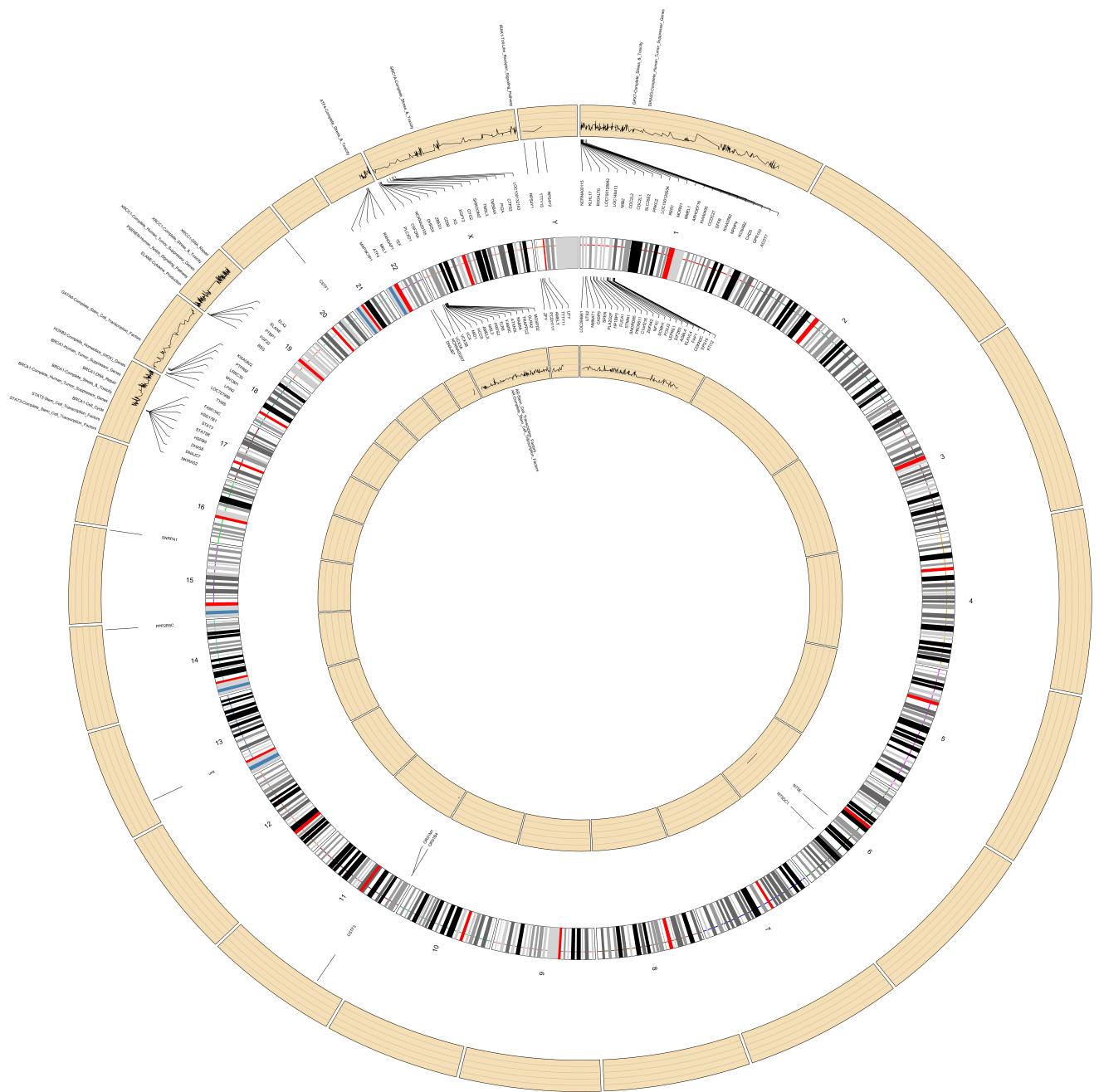
Circos plot was made using RCircos, the inner track refers to the H460 parent genes and the outer track refers to the H460 knock-in genes.

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, random 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



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