Analysis of Gene Expression Omnibus Leukaemia Data from the Illumina HumanMethylation450 Array

Alice Zhu, Rachel Edgar, Shaun Jackman and Nick Fishbane

Question

DNA methylation has recently been emerging as an important feature of cancer[1]. Changes in the methylation of genes and their promoters can cause a change of gene expression that may result in oncogenesis. Several types of leukaemia are known to have mutations in genes involved in DNA methylation[2,3]. Study of genome-wide DNA methylation patterns may lead to new insights into the development of leukaemia. Public availability of genomic data allows for meta studies with samples sizes that would not be possible without the sharing of data between research groups. The Gene Expression Omnibus (GEO) represents a vast quantity of data for large-scale research including many studies on DNA methylation and cancer, and specifically leukaemia.

For our study we will compare the methylation patterns in Acute Promyelocytic Leukemia (APL) and Acute Lymphoblastic Leukemia (ALL) to each other and to control samples. We hope to identify biologically relevant changes in CpG island level methylation that could relate to phenotypic differences seen between APL and ALL.

Data

Table 1. Samples available from Illumina 450K leukemia or B-cells studies

Series	Study	Cases	Controls
GSE39141[4]	ALL	29 (bone marrow)	4 (B-cells)
GSE42118[5]	APL	8 (4 primary diagnosis, 4 remission)	2 (1 bone marrow, 1 CD34+ cell line)
GSE42865[6]	Progeria and Werner syndrome	0	9 (3 blood mononuclear cells, 3 immortalized B-cells, 3 naive B-cells)
Total		**37**	**15**

Exploratory Analysis Pre-Normalization

- Hierarchical clustering
  - HBM removed from differential methylation analysis as the samples clustered with cases, potentially due to the methylation variability seen between tissue types

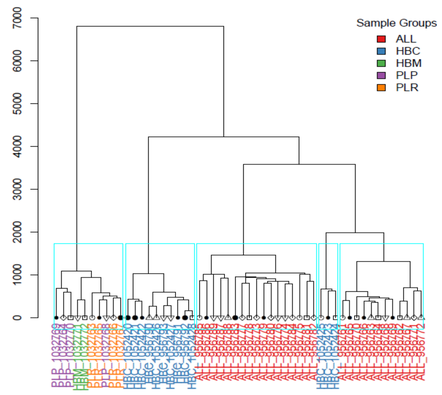


Fig 1. Unsupervised hierarchical clustering based on raw beta values from all probes. Colours below the branches represent the sample group.

Normalization

- Beta-Mixture Quantile Normalisation (BMIQ)
- Normalized each sample individually to shift type II probe beta distribution toward the type 1 probe beta distribution

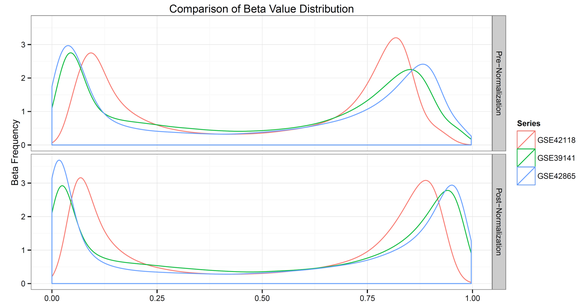
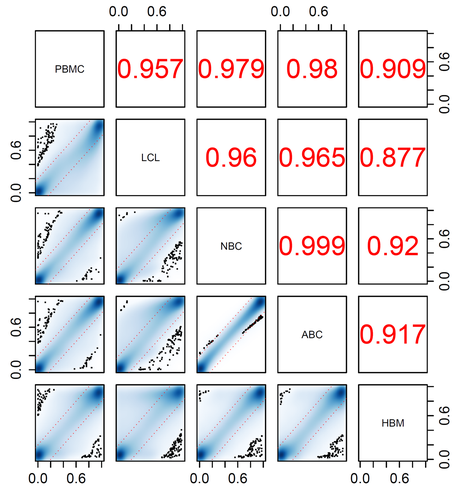


Fig 2. Density distributions of mean beta values for each probe across all samples in a series. Distributions in the top plot are before BMIQ normalization and distributions in the bottom plot are post BMIQ normalization

Control Sample Variability

- Our samples were from a variety of tissue types and from 3 different studies, to control for variability due to tissue or series treatment, probes that vary between 4 pairs of controls controls were filtered (probes considered variable had δ beta value > 0.2)
  - While HBM tissue samples were excluded from differential methylation analysis HBM was included here to control for variability between bone marrow and B cells

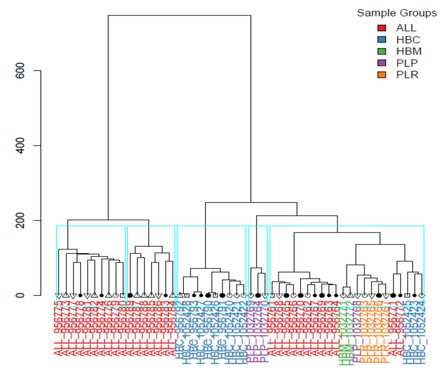


- 49,829 probes were removed based on variability between control sample tissue type, and 9 probes which had NA values in all samples of a control group. Leaving 435,739 probes for differential methylation analysis

PBMC=Blood mononuclear cells (n=3)
LCL= Immortalized B-cells (n=3)
NBC=Normal B-cells (n=3)
ABC=Normal B-cells (n=4)
HBM-healthy bone marrow (n=2)

Fig 4. Filtration of probes based on variability between different control sample tissue types. Correlation coefficient (calculated by Pearson's correlation analysis) between tissue types is show in the upper half of the plot. Broken red lines shown a δ average beta of 0.2.

Hierarchical clustering Post-Normalization, Post Filtering



- Remission APL cases do not cluster with primary samples have not been used in subsequent analysis
- APL now n=4

Fig 4. Unsupervised hierarchical clustering based on beta values from all 289532 probes after normalization and filtering. Node names are color-coded according to their group. Data is plotted using ward method.

Differential Methylation Analysis

- We used a Per-Island Mixed Effect Model to examine Island Level Differences between groups using logit transformed beta values (M values)

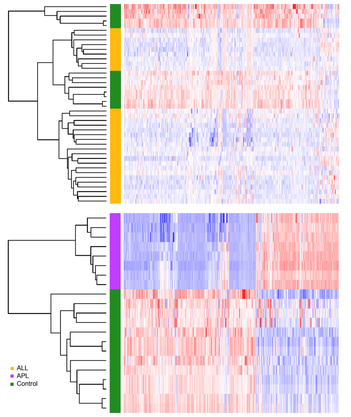


Fig 5. Heat map of M values of probes located in the top 10 differentially methylated islands. a) Probes (661) located in differentially methylated islands between ALL and controls. a) Probes (780) located in differentially methylated islands between APL and controls. Heat maps are clustered by sample and probe value.

$$y_{ij}=\mu+\beta_i+\tau_j+\varepsilon_{ij}$$
$$j=1\ldots52 \text{ (samples)}$$
$$i=1,\ldots M \text{ (M probes in island)}$$
$$\beta_i \sim N(0,\sigma^2)$$
$$\tau=\text{group of subjects } j$$
$$\varepsilon_{ij}=\text{noise } {}^{iid}\sim N(0,\sigma^2)$$

Equation 1. Per-Island Mixed Effect Model with probe (P) as a random fixed effect.

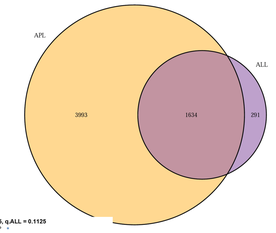


Fig 6. Venn diagram illustrating the overlap of differentially methylated islands ($q<1x10^{-25}$) between two leukaemia types (ALL and APL).

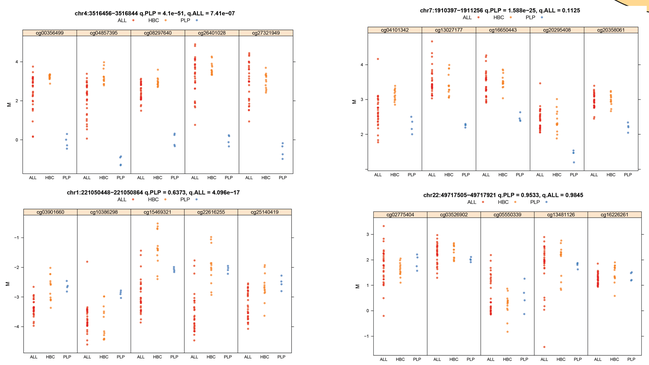


Fig 7. Scatter plot of beta values for four example islands. Individual probe beta values are averaged across all samples in a given group (ALL, APL, or control) for each probe in an island chr1:221050448−221050864 and chr7:1910397−1911256, chr4:3516456−3516844 are significantly differentially methylated (t>15) from control samples in both cancers, APL and ALL and, respectively. chr22:49717505−49717921 is a representative "boring" island which is not differently methylated in either cancer type.

Gene Set Enrichment Analysis

- Enrichment of the transmembrane receptor protein tyrosine phosphatase in both cancers is relevant to the cancer phenotype as the associated genes are involved in cell growth and have been shown as differentially methylated in liver cancer[4]
- SMAD binding genes are bone growth factors and implicated in head and neck cancer[5]
- Major histocompatibility complex has been previously linked to cervical cancer[6]
- Enrichment of molecular function in both cancer types likely relates to its high number of annotation and is not meaningful for our study

Table 2. Enrichment of GO terms in differentially methylated island list, according to Fishers exact and Kolmogorov-Smirnov tests.

GO.ID	Term	Annotated	Significant	Expected	Rank in classic Fisher	Classic Fisher	classicKS	Sample
GO:0003674	molecular_function	14813	160	363.23	463	1	< 1e-30	ALL
GO:0003674	molecular_function	14813	1226	1808.06	1214	1	< 1e-30	APL
GO:0005001	transmembrane receptor protein tyrosine ...	98	0	2.4	467	1	2.10E-12	ALL
GO:0005001	transmembrane receptor protein tyrosine ...	98	4	11.96	1100	1	3.30E-13	APL
GO:0046332	SMAD binding	114	1	2.8	319	0.94	9.50E-10	ALL
GO:0032395	MHC class II receptor activity	4	0	0.49	1217	1	0.016	APL

References

1. Esteller M, *et al.* (2001) A gene hypermethylation profile of human cancer. Cancer Res. 61(8):3225-9
2. Ley TJ, *et al.* (2010) DNMT3A mutations in acute myeloid leukemia. N Engl J Med. 363(25):2424-33
3. Wiemels JL, *et al.* (2001) Methyleneterahydrofolate reductase (MTHFR) polymorphisms and risk of molecularly defined subtypes of childhood acute leukemia. Proc Natl Acad Sci U S A. 98(7):4004-9
6. Chen, D. *et al.* Genome-wide Association Study of Susceptibility Loci for Cervical Cancer. *J. Natl. Cancer Inst.* (2013).
4. Hsu, S. H. *et al.* Methylation of gene encoding the growth suppressor protein tyrosine phosphatase receptor-type O (PTPRO) in human hepatocellular carcinoma and identification of VCP as its bona fide substrate.*J. Cell. Biochem.* (2013).
5. Yang, W. H., Lan, H. Y., Tai, S. K. & Yang, M. H. Repression of bone morphogenetic protein 4 by let-7i attenuates mesenchymal migration of head and neck cancer cells. *Biochem. Biophys. Res. Commun.* 433, 24-30 (2013).