Genomics 99 (2012) 209–219

Contents lists available at SciVerse ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno

Characterization of DNA methylation and its association with other biological systems in lymphoblastoid cell lines

Zhe Zhang [a,*], Jinglan Liu [b], Maninder Kaur [c], Ian D. Krantz [c]

[a] Center for Biomedical Informatics, The Children's Hospital of Philadelphia, Philadelphia, PA 19104, USA
[b] Department of Pathology and Laboratory Medicine, St. Christopher's Hospital for Children and Drexel University College of Medicine, Philadelphia, PA 19134, USA
[c] Division of Human Genetics, The Children's Hospital of Philadelphia and The Perelman School of Medicine, The University of Pennsylvania, Philadelphia, PA 10104, USA

ARTICLE INFO

Article history:
Received 6 May 2011
Accepted 5 January 2012
Available online 15 January 2012

Keywords:
Lymphoblastoid cell line
DNA methylation
Epigenetics
Systems biology
Genetic diseases

ABSTRACT

Lymphoblastoid cell line (LCL) is a common tool to study genetic disorders. However, it has not been fully characterized to what degree LCLs preserve the in vivo status of non-genetic biological systems, such as DNA methylation and gene transcription. We previously reported that DNA methylation in LCLs is highly variable in a data set of ~27,000 CpG dinucleotide sites around transcription start site (TSS) and 63 human subjects including healthy controls and probands of genetic disorders. Disease-causing mutations are linked to differential methylation at some CpG sites, but account for a small proportion of the total variance. In this study, we repeated the experiments to ensure that the high variance is not due to technical error and scrutinized the characteristics of DNA methylation and its association with other biological systems. Using sequence information and ChIP-seq data, we conclude that local CpG density and histone modifications not only correlate to baseline methylation level, but also affect the direction of methylation change in LCLs. Integrative analysis of gene transcription and DNA methylation data of the same subjects shows that medium or high methylation around TSS blocks the transcription while low methylation is a necessary, but not sufficient condition of downstream gene transcription. We utilized epigenetic information around TSS to predict active gene transcription via logistic regression models. The multivariate model using DNA methylation, eight histone modifications, and two regulatory protein complexes (CTCF and cohesin) as predictors has better performance (accuracy = 95.1%) than any univariate models of single predictors. Linear regression analysis further shows that the transcriptional levels predicted by epigenetic markers have significant correlation to microarray measurements ($p = 2.2e-10$). This study provides new insights into the epigenetic systems of LCLs and suggests that more specifically designed experiments are needed to improve our understanding on this topic.

1. Introduction

Status of cytosine methylation at CpG dinucleotide sites is a key component of epigenetic regulation of gene activity. It influences gene transcription by adjusting the accessibility of chromosomal regions, and controls various biological processes such as X inactivation [1], gene imprinting [2], chromatin remodeling [3], and pathogenesis [4]. De novo methylation is an essential element of cell differentiation during development [5,6] while in somatic cells, methylation status is believed to remain stable throughout cell division, so studies of tissue-specific methylation often use cultured cells as source material [6–8].

The distribution of the ~30 million CpG sites in the human genome is not even. They are generally under-represented and hypermethylated in intergenic regions while a large number of unmethylated CpG sites, called CpG islands (CGIs), cluster around transcription start sites (TSS) [9,10]. Perturbed DNA methylation at gene promoters has been linked to a number of human disorders [11–15]. Technology for quantitative measurement of methylation at all CpG sites is available but costly. A recent study that investigated the dynamics of DNA methylation during cell differentiation used massively parallel sequencing technology to obtain 542 million sequencing reads on average from three samples [6]. Although the reads averagely cover the whole human genome by nine folds, more than 20% of the CpG sites are mapped by less than three reads, making them unsuitable for quantitative comparison between samples. Limiting measurement to regions of higher biological relevance, such as promoters and 5′-UTRs, would lower the experimental cost and increase the practical number of samples of individual studies. Weber et al. compared the methylation at >12,000 CpG islands between normal fibroblasts and SW48 cancer cells via microarray technology and identified over 200 hypermethylated loci in cancer cells [16]. Ehrich et al. used mass spectrometry technology to quantify methylation patterns of >400 cancer-related genes in 59 cancer cell lines and

[*] Corresponding author at: Center for Biomedical Informatics, The Children's Hospital of Philadelphia, 3615 Civic Center Blvd, Philadelphia, PA 19104, USA.
E-mail address: zhangz@email.chop.edu (Z. Zhang).

0888-7543/$ – see front matter © 2012 Elsevier Inc. All rights reserved.
doi:10.1016/j.ygeno.2012.01.002

discovered that a large portion of the tested genes have altered methylation in cancer cells [7]. Koga et al. utilized tiling microarrays to measure methylation at promoters of all RefSeq genes in normal melanocytes and eight melanoma cell strains and revealed the diagnostic value of DNA methylation information [17]. These studies demonstrate the usefulness and feasibility of identifying chromosomal regions that exhibit differential methylation under varying biological conditions.

Lymphoblastoid cell lines (LCLs) are established by transforming lymphoblasts with Epstein–Barr virus (EBV) [18]. It is a renewable source of genetic information and a common tool for studying human disorders [19–22]. GM12878, an LCL generated from a female donor, is a model cell line used by the International HapMap [23] and ENCODE [24] Projects. ChIP-seq data of histone modifications in GM12878 are publically available [25], along with DNA methylation data generated by both microarray and Methyl-seq technologies [8].

LCLs usually have normal diploid karyotypes and stable DNA sequences. Its estimated mutation rate is $2–30\times10^{-7}$ mutations per cell division [26]. However, the viral transformation and continuous cell culturing and storage may lead to more substantial alterations in the epigenetic, transcriptional, and translational systems. Altered and destabilized DNA methylation was recently reported in LCLs [27–29], which suggests that the in vivo status of epigenetic systems is not fully preserved. Nevertheless, LCLs are a valuable resource for investigating mutationally defined genetic disorders. It is easier to isolate and evaluate the consequence of DNA mutations in cultured cells grown under a controlled environment than in fresh cells whose status can be confounded by numerous environmental and clinical factors. Therefore, a comprehensive characterization of non-genetic biological systems and their association with each other in LCLs would provide valuable information for future studies.

We previously used microarray-based technology to quantify DNA methylation at 27,578 CpG sites in LCLs generated from 22 healthy controls, two Roberts syndrome (RBS) probands and 39 Cornelia de Lange syndrome (CdLS) probands [30]. CdLS is a dominant congenital multisystem disorder with craniofacial, cardiac, gastrointestinal, genitourinary, skin and other system involvement as well as delays in growth and intellectual development. Disease-causing mutations of CdLS have been identified in genes *NIPBL*, *SMC1A*, and *SMC3*, all of which have been associated with cohesin complex [31]. Our comparative analysis identified 152 CpG sites whose methylation was different between the control and the CdLS groups with a high degree of confidence ($p<0.001$). This number is much smaller than the number of genes whose transcription is significantly altered in CdLS as measured by gene expression microarray studies [19]. Therefore, the disease state is unlikely a major factor of between-sample variation in this data set.

In this study, we used the same 63 samples for a generalized characterization of DNA methylation in LCLs. We repeated the microarray experiments to estimate the contribution of measurement errors to the total variance. Instead of comparing the control and CdLS samples, data analysis in this study is focused on between-sample variation independent of disease state, age, gender and other known clinical variables. Integrative analysis of various existing and new data sets identified distinctive patterns of associations between histone modifications, DNA methylation, and gene transcription. The results of this study will lead to a better understanding of the non-genetic biological systems in LCLs.

2. Material and methods

2.1. Sample preparation of methylation assays

Cell culture, DNA isolation, and bisulfite treatment were performed as described previously [30]. In summary, lymphoblastoid cell lines (LCLs) of 63 human subjects were cultured anonymously and processed in random order; DNA was isolated using the DNA purification kit from Gentra Systems; and 500 ng purified DNA from each sample was conversed using the EZ DNA methylation kit from Zymo Research. The bisulfate conversion changed the unmethylated C to T, but made no change at the methylated CpG sites. The 63 prepared LCLs as well as 3 universally methylated and 6 universally unmethylated controls were randomly assigned to 6 Infinium HumanMethylation27 BeadChips (Illumina, Inc.). All human subjects were included in this study under an IRB-approved protocol of informed consent at The Children's Hospital of Philadelphia and the Misakaenosono Mutsumi Developmental, Medical, and Welfare Center and their detailed description is available as GSE18458 series within Gene Expression Omnibus database.

2.2. Processing of methylation data

Each HumanMethylation27 BeadChip carries beads measuring DNA methylation at 27,578 CpG sites located around 14,495 unique Entrez genes. Each site is measured by two types of beads; one measures the methylated (M) allele and the other measures the unmethylated (U) allele. After the prepared DNA samples were hybridized to the beads and fluorescently stained, the BeadChips were scanned by BeadArray Reader (Illumina, Inc.) and the scanned data were processed by BeadStudio Methylation Module (Illumina, Inc.). Background-subtracted signal intensities of both alleles and a detection p value of each CpG site were exported from BeadStudio and imported into R statistical environment (http://www.r-project.org) for further processing and statistical analysis. The intensities of methylated and unmethylated alleles were normalized separately across 63 LCLs using the quantile spline method [32] of *affy* package in R. Since CpG sites on X and Y chromosomes can have very different methylation between males and females, those sites were normalized separately in male and female groups. The methylation level at each CpG site in each LCL is represented as $\beta=M/(M+U)$; where M and U are normalized intensities of methylated and unmethylated alleles. Therefore, β value indicates the fraction of methylated alleles in a cell population. We considered β value less than 0.1 or greater than 0.9 as low or high methylation level corresponding, and the β value between 0.1 and 0.9 as medium methylation level. The whole processed data set is a 63 by 27,578 data matrix of β values ranging between 0 and 1.

2.3. Bioinformatics analysis

UCSC Genome Browser tracks were downloaded using the *Table Browser* tool. The “*HAIB Methyl27*” track provided the DNA methylation data of two technical replicates of GM12878. We called a CpG site in low quality if its β value was 0 in either replicate or the β value difference between replicates was greater than 0.05. Enrichment of pre-defined gene sets in significant genes was analyzed via the functional annotation tools of DAVID (Database for Annotation, Visualization and Integrated Discovery) [33]. The functions/packages used for statistical analysis are *cor/stats* for correlation analysis, *aov/stats* for ANOVA analysis, *t.test/stats* for Student's *t* test, *prcomp/stats* for principal components analysis, *performance/ROCR)* for ROC analysis, *glm/stats* for logistic regression and *lm/stats* for linear regression analysis. More information about the data analysis is available in Supplemental methods.

3. Results

3.1. General characterization of DNA methylation around TSS

We used the Infinium HumanMethylation27 microarray platform to measure the methylation levels of 27,587 CpG sites close to the TSS in 63 lymphoblastoid cell lines (LCLs) after treating extracted

DNA with bisulfite conversion. To evaluate the technical errors introduced by microarray experiments, all samples were measured twice at the same location (Wistar Institute). The detailed batch comparison is described in Supplemental File 1. In summary, we found that when the measurements have the best detection p value, 1) all samples except one have high correlation between the duplicated measurements (average Pearson's r = 0.991); 2) outliers are common, but they are mostly consistent between batches; 3) while about one-eighth of the duplicated measurements have β value difference greater than 0.05, the batch difference of a given CpG site is generally consistent across samples; and 4) CpG sites with medium methylation level ($0.1<\beta<0.9$) are less affected by batch effect than sites with very high or low methylation. We then concluded that measurements with the best detection p value are precise and repeatable. Since the proportion of measurements having the best quality is substantially higher in the second batch than in the first (97.7% vs. 45.9%), we only used the data of the second batch throughout the rest of this study.

Starting from 27,587 CpG sites and 63 samples, we filtered the data by removing eight samples with over 1000 less-than-the-best quality measurements, and further excluded CpG sites having any less-than-the-best quality measurements or extreme outliers (more than 20 interquartile ranges from the first or third quartile). The filtering substantially reduced the proportion of technical errors. The remaining data includes 24,952 CpG sites and 55 samples of 19 gender and race matched healthy controls, 2 Roberts syndrome probands, and 34 CdLS probands (21 severe and 8 mild cases with *NIPBL* mutations, 4 mild cases with *SMC1A* mutations, and 1 mild case with *SMC3* mutation [31]).

Since most CpG sites measured by the microarray platform are located in CpG islands (CGIs) around transcription start sites (TSSs), the distribution of their methylation is skewed to the hypomethylation end (Supplemental Fig. 1) and the global average of β values, which indicate the fraction of methylated alleles in all cells, is 0.22. The average β values of 59.5% CpG sites are less than 0.1 and only 20.3% sites are hypermethylated (average $\beta>0.5$). We associated each CpG site to its nearest TSS according to the "*UCSC Genes*" track of UCSC Genome Browser. Approximately 98% of sites are located within the −1.5 to 1.5 kb region of any TSS. Consistent with previous studies (Fig. 4A in [6] and Fig. 1E in [7]), sites closer to the TSS are generally less methylated, especially those in CGIs (Fig. 2A). About 43% of sites are located within CGIs according to the "*CpG Islands*" track of UCSC Genome Browser. The average β values of CGI and non-CGI sites are 0.077 and 0.334 respectively. CGI sites close to the ends of CGI generally have higher methylation than sites in the middle (Supplemental Fig. 2) and sites close to potential transcription factor binding sites (TFBS) have slightly lower methylation than sites in the flanking regions (Supplemental Fig. 3). The TFBS information was downloaded from the UCSC "*TFBS conserved*" track, which includes human–mouse–rat conserved loci matching to the consensus binding motifs of 258 transcription factors. Methylation at TFBSs differs dramatically between those motifs (Supplemental Table 2). For example, the average β values around loci matching to V$HNF1_01 and V$NRF2_01 are 0.340 and 0.052 respectively.

About two CpG sites were measured around each TSS on average. We queried whether individual sites act independently or adjacent sites are regulated concordantly. The answer to this query will tell us to what degree a single CpG site represents the overall methylation status of surrounding region. We identified 10,878 adjacent pairs of CpG sites located within 1 kb of each other and evaluated the differential methylation of paired sites. The β value difference of adjacent sites is 0.128 on average while the global average of β value differences between any two autosomal sites is 0.288. The average difference is further reduced to 0.041 when sites are within 10 bases of each other (Supplemental Fig. 3). We next evaluated the co-regulation of adjacent sites by calculating their correlation across all 55 samples. The average Pearson's r is only 0.114. Pairs located in the same CGI have slightly higher correlation (average r = 0.124). The correlation between pairs generally increases as their distance becomes shorter (Fig. 1B, blue line). Supplemental File 1 shows that the methylation measurements have low sensitivity to subtle change at highly methylated or unmethylated sites. Therefore, when we only use the pairs having medium methylation at both sites, the average r increases to 0.219. Furthermore, sites within 10 bases of each other have a much higher average r of 0.809. We compared the sequence features, such as GC content and TFBS frequency, around correlated and uncorrelated adjacent sites, but were unable to recognize notable differences between these two types of pairs. These results suggest that the methylation of most adjacent CpG sites is not closely co-regulated.

3.2. Differential methylation

DNA methylation at CpG sites around TSSs is highly variable between samples, especially when the β values are between 0.1 and 0.9 (Supplemental Fig. 5). To test whether such high variability is also present in other cell types, we performed a meta-analysis of twelve DNA methylation data sets generated from different cell types, but using the same microarray platform (Supplemental File 2). After calculating the correlation of methylation levels between each pair of samples in the same disease or treatment group, we compared the distribution of correlation coefficients of different cell types. LCLs have higher between-sample variation than all the other cell types (T cells, monocytes, whole blood, and colon mucosa) with the exception of colorectal cancer cells.

Outliers are common in the data set as 535 autosomal CpG sites have β values ranged from less than 0.1 to more than 0.9. Principal components analysis (PCA) using all autosomal sites was unable to clearly separate samples by their disease status (Fig. 2A). The top three principal components only account for less than 20% of the total variance, suggesting that DNA methylation in LCLs is affected by many factors, including, but not limited to, disease status, gender, genetic background, developmental stage and cell culture condition.

Two-way ANOVA analysis of gender and disease status as two interacting factors identified 283 CpG sites differentially methylated between 19 controls and 21 severe CdLS patients. These sites have ANOVA p values less than 0.01 and β value differences greater than 0.05. Among the 699 sites significantly different between females and males, 610 and 3 are located on chromosomes X and Y respectively, suggesting that gender has little impact on autosomal methylation. We took a closer look at the gender difference on the X chromosome, and noticed that sites in CpG islands are generally unmethylated in males but mostly have higher methylation in females (Supplemental Fig. 6). This result was indeed anticipated because increased methylation of CpG islands plays an essential role in X inactivation [34]. An unexpected observation is that the 193 X chromosome sites having greater than 0.5 β values in both genders are significantly more methylated in males than in females (average $\beta = 0.77$ vs. 0.71, p = 1.7e-20, paired *t* test). It is unlikely that such a dramatic difference was caused by technical bias or data processing. We postulate that it is more difficult for females to maintain hypermethylation at those sites on both X chromosomes due to the need to methylate CpG islands. However, a valid interpretation of this observation requires further investigation.

Among the 283 sites differentially methylated between control and severe CdLS samples, 177 are up-methylated and 106 are down-methylated in CdLS. The corresponding false discovery rate (FDR) is 0.24 according to a permutation procedure that shuffled the sample labeling of disease status, but not gender. According to DAVID functional annotation [33], a number of pre-defined gene sets are significantly enriched in the genes downstream to those differentially methylated sites. Some of those gene sets are evidently related to CdLS (Supplemental Table 1). For example, four genes (*TBX5*,

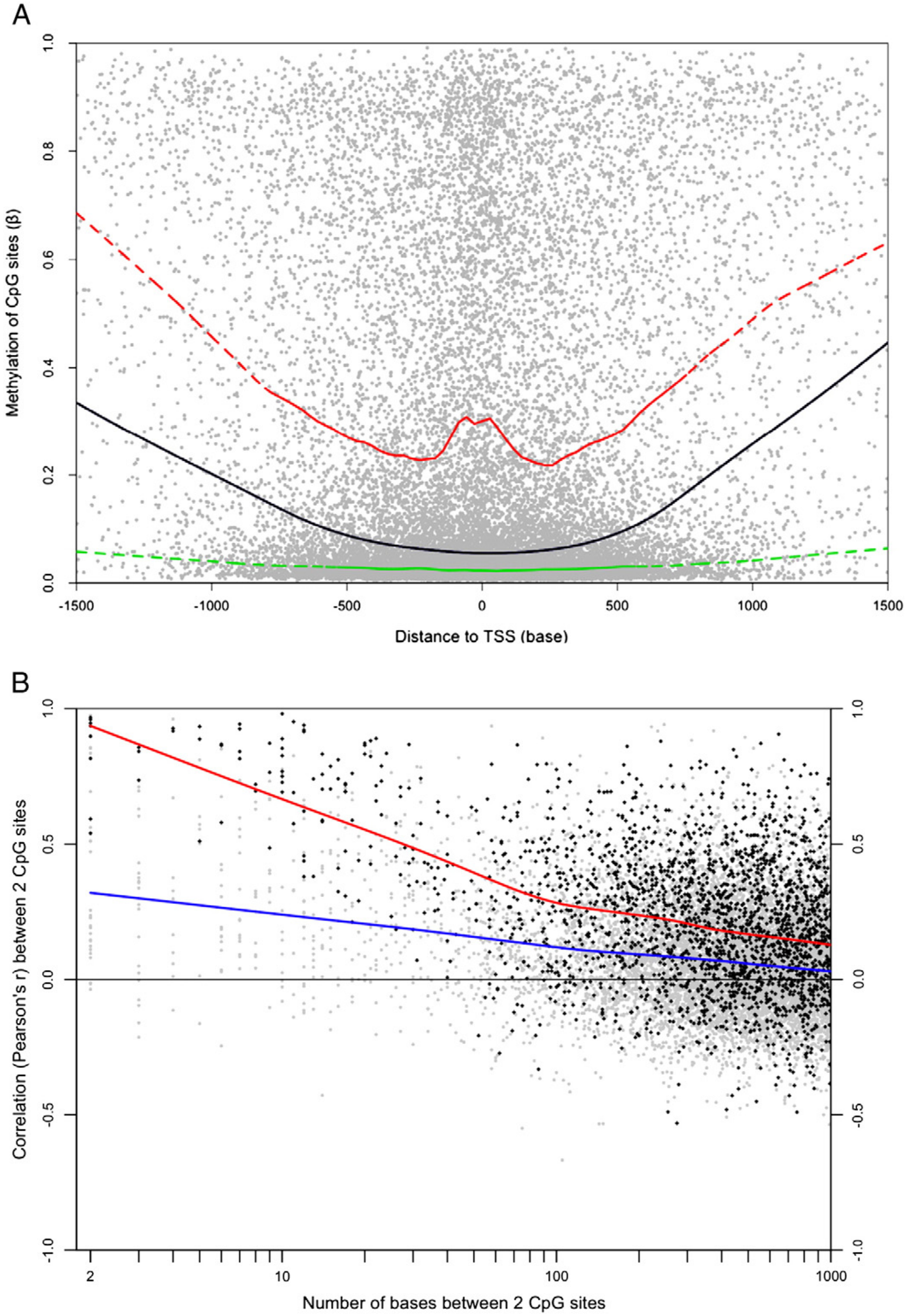

Fig. 1. Characteristics of DNA methylation in LCL. A) Each dot represents a CpG site on autosomes. The X-axis indicates the distance to the nearest TSS and the y-axis is the average β value of 55 LCLs. The lines were generated by Lowess smoothing (black: all sites; green: CGI sites; red: non-CGI sites). Non-CGI sites have higher methylation than CGI sites in general no matter their distance to TSS. B) Each dot represents a pair of CpG sites on autosomes. The x-axis is the distance between the two sites and the y-axis indicates their Pearson's correlation coefficient across 55 LCLs. The lines were generated by Lowess smoothing (blue: all pairs; red: pairs whose average β values are between 0.1 and 0.9 at both sites).

MSX1, *MBNL1*, and *SALL4*) involved in embryonic limb morphogenesis, which is one of the most common features of CdLS, have down-regulated CpG sites around their TSS.

The sites differentially methylated between controls and severe patients have two noteworthy features. First, all 283 sites have medium methylation ($0.1<\beta<0.9$) in at least one group although 57.8% of the total sites have low methylation ($\beta<0.1$) in both groups. We then limited the remaining analysis of this section to the 9776 sites having medium methylation. Second, the 177 sites up-methylated in CdLS include significantly lower percentage of CGI sites than the 106 down-methylated sites (5.6% vs. 49.1%, p = 4.6e-17, proportional test) while 18.7% of the unchanged sites are located in CGIs (Supplemental Fig. 7). This result suggests that the likelihood and direction of methylation change in CdLS are related to local CpG density.

We previously used transcriptional microarray data to generate a diagnostic index of CdLS by comparing healthy controls and severe patients. It was shown that this index could be used to discriminate controls and CdLS patients as well as CdLS subtypes [19]. We then evaluated if methylation information could be used for the same purpose via a combination of nearest centroid classification and leave-one-out validation (details in Supplemental methods). A methylation-based index classified control samples and severe patients with significant accuracy. In addition, this index can discriminate mild CdLS cases from both controls and severe patients (Fig. 2C). However, the leave-one-out validation misclassified six control samples and three severe patients (accuracy = 77.5%) while the index of transcriptional data correctly classified 90.1% testing samples (Fig. 1C in [19]). The area under ROC curves of methylation- and transcription-based prediction is respectively 0.860 and 0.985 (Supplemental Fig. 8). Therefore, DNA methylation pattern is a less powerful diagnostic index of CdLS than gene transcription pattern.

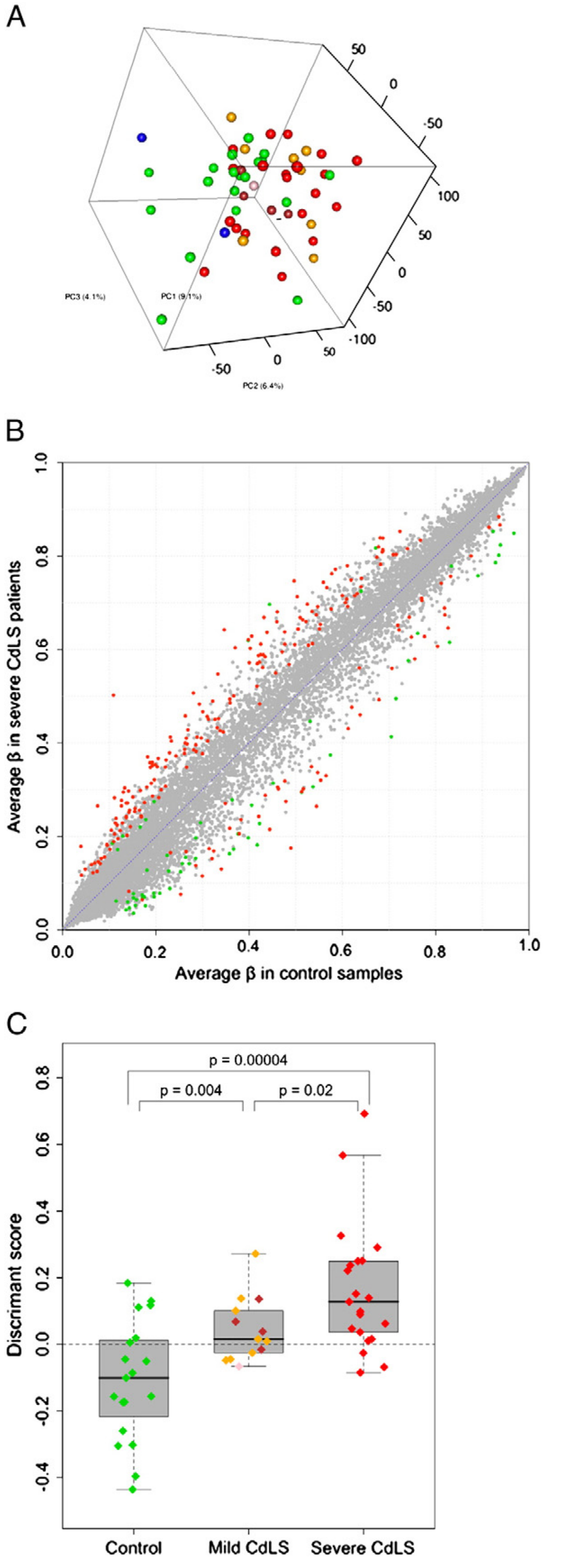

Fig. 2. Differential methylation. A) Principal components analysis of LCLs using all measured CpG sites on autosomes. Colors indicate disease status (green = control; red = severe CdLS; orange = mild CdLS with NIPBL mutation; brown = CdLS with SMC1A mutation; pink = CdLS with SMC3 mutation; and blue = Roberts Syndrome). B) Differential methylation between control and severe CdLS samples. Each dot represents one autosomal site. Significant sites ($p<0.01$ and $|\Delta\beta|>0.05$) were highlighted (green = CGI sites and red = non-CGI sites). C) The control, mild CdLS, and severe CdLS samples could be distinguished according to their methylation pattern. The y-axis indicates the discriminant score that is corresponding to the relative similarity of each sample to the centroids of control and severe CdLS groups. The scores of controls and severe patients were obtained via a leave-one-out procedure, and the scores of mild patients were based on the 283 sites differentially methylated between controls and severe patients (details in Materials and methods). By default, samples with score > 0 would be classified as CdLS. Each diamond represents a sample (colored as in Fig. 2A). The p values are the results of Student's *t* test.

3.3. The association of DNA methylation with other epigenetic features

GM12878 is a model LCL from a female donor. A variety of genomic data generated from GM12878 are available through the ENCODE (ENCyclopedia of DNA Element) project [24]. We downloaded three sets of GM12878 data from the UCSC Genome Browser tracks: "*HAIB Methyl27*" (DNA methylation data generated from Infinium microarrays), "*UW DNaseI HS*" (DNaseI hypersensitivity data generated by deep sequencing), and "*Broad Histone*" (CTCF binding and eight histone modification data generated by ChIP-seq experiments).

We were able to directly compare the average β values of two GM12878 replicates and the female samples in our data set since both were generated on the same microarray platform. The Pearson's r of all autosomal sites between the two vectors of average β values is 0.895 (Supplemental Fig. 9). After low quality measurements were removed, the between-data set correlation was improved to 0.947 while the average r value of all female sample pairs in our data set is 0.960. We concluded that GM12878 is compatible with our samples in terms of DNA methylation and it is possible to associate our methylation and gene expression data with data generated from GM12878.

DNaseI hypersensitivity is a sequence feature related to DNA accessibility [35]. More than 100,000 short DNaseI hypersensitivity regions were identified from each of two GM12878 replicates. We mapped the CpG sites in our data to those regions and found that sites located within those regions have significantly lower β values than the other sites (0.045 vs. 0.262, p < 1e-300). Since CGIs and DNaseI hypersensitivity regions are often overlapped, we asked whether CpG density and DNaseI hypersensitivity affect DNA methylation independently. Sites located in DNaseI hypersensitivity region only have slightly lower average β than sites located in CGIs only (0.075 vs. 0.089, p = 0.0001), and the average β of sites located in overlapping regions is further reduced to 0.028 (Supplemental Fig. 10). Therefore, CpG density and DNaseI hypersensitivity have additive effect on DNA methylation and DNaseI hypersensitivity is probably a stronger indicator of low methylation than CpG density.

Histone modifications are likely more involved in transcriptional regulation than DNA methylation. ChIP-seq data of eight histone

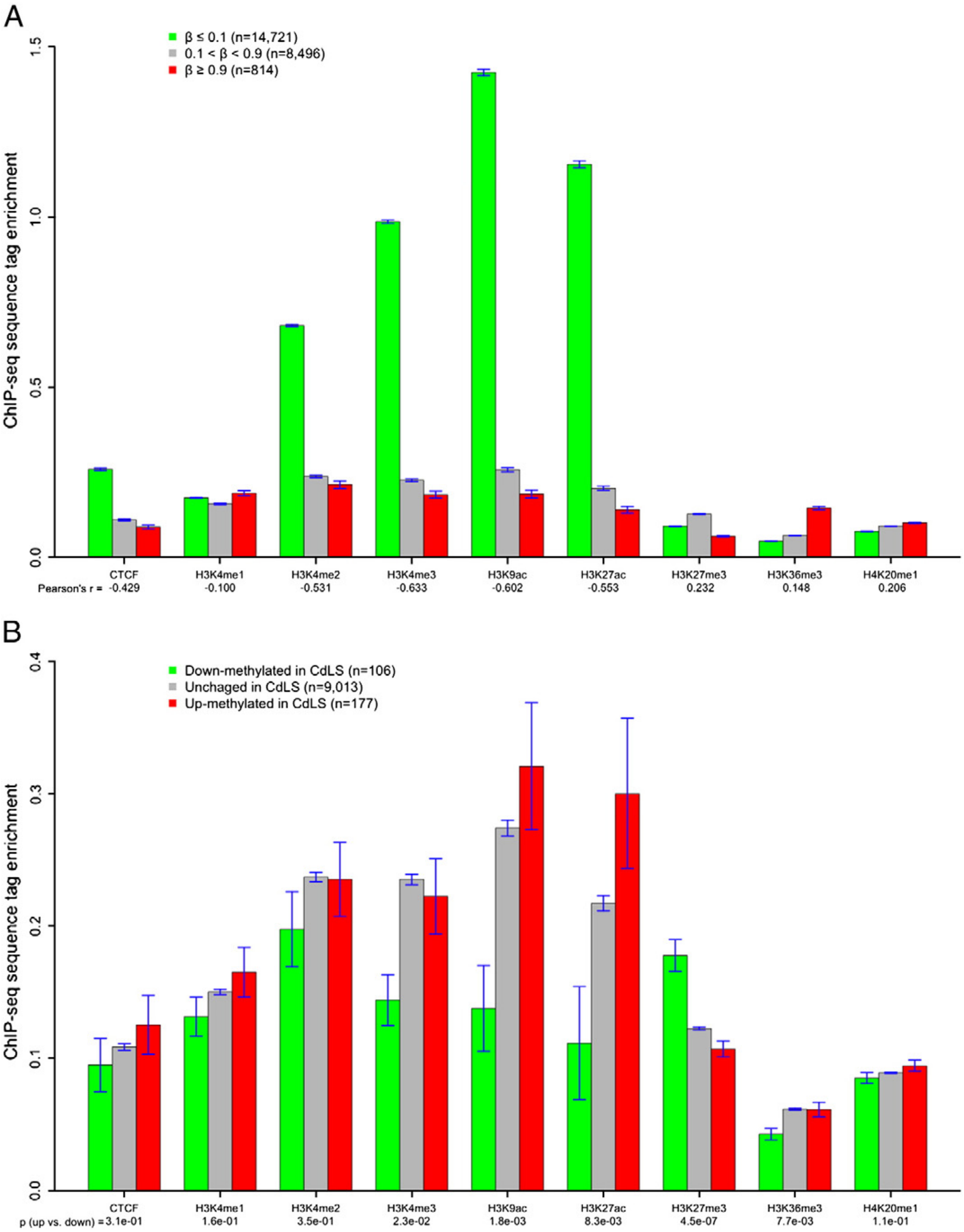

Fig. 3. The association between DNA methylation and histone modifications. The y-axis represents the average tag enrichment based on ChIP-seq data. A) The correlation of average methylation to CTCF binding and eight histone marks in GM12878 at all autosomal CpG sites. B) Histone status at CpG sites that were down-methylated, unchanged, and up-methylated in severe patients. The unchanged group include only sites with medium methylation ($0.1<\beta<0.9$). The p values are results of Student's *t* test comparing the down- and up-methylated sites.

modifications and CTCF binding of GM12878 are available at the 24,379 CpG sites measured by this study. We calculated the correlation between β value and tag enrichment at those sites (Fig. 3A) and found that DNA methylation has a negative and relatively stronger correlation with euchromatin (decondensed chromatin) marks such as H3K4me3 and H3K9ac [36], and positive but weak correlation with heterochromatin (condensed chromatin) marks such as H3K27me3 and H4K20me1 [37]. This observation is in agreement with previous studies. For example, Brunner et al. reported that H3K4me3 and H3K27me3 signals are respectively correlated to unmethylated and methylated status in human embryonic stem cells [8], and Wu et al. found that DNA methylation has a negative correlation to H3K9ac in a mouse leukemia cell line [38]. These results suggest that the association between DNA methylation and histone modifications has similar pattern in different cell types and species. The association between DNA methylation and histone modifications is not particularly affected by CpG density except H3K27me3 (Supplemental Fig. 11).

Histone modifications are also related to the direction of methylation alterations. CpG sites down-methylated in CdLS have lower tag enrichment of CTCF and all histone modifications except H3K27me3 than up-methylated and unchanged sites (Fig. 3B). The exception of H3K27me3 is probably caused by the fact that down-methylated sites include higher percentage of CGI sites than up-methylated and unchanged sites (Supplemental Fig. 7) while CGI sites have substantially higher H3K27me3 than non-CGI sites (Supplemental Fig. 11). However, CpG density has little effect on other histone modifications. For example, while CGI sites have slightly higher H3K9ac than non-

CGI sites, down- methylated sites have significantly lower H3K9ac than other sites. Altogether, these results suggest that histone modifications are not only correlated to baseline DNA methylation, but are also related to the likelihood and direction of methylation alterations in LCL.

3.4. The association between DNA methylation and gene transcription

We previously published a gene expression microarray data set that included LCLs of 39 human subjects: 18 controls, 17 severe CdLS probands, 2 Roberts syndrome probands and 2 Alagille syndrome probands [19]. The DNA methylation of 27 of those subjects (13 controls and 14 severe CdLS patients) was also measured by this study. The probes of Affymetrix U133 Plus 2.0 platform used for the expression experiments were remapped to the current version of NCBI Entrez genes [39] and grouped into 17,726 unique Entrez genes. 10,430 genes measured by at least six mRNA probes and one CpG probe (within −1.5 to 1.5 kb of TSS) were used in the remaining analyses of this section. According to MAS5.0 algorithm [40], 51.7% and 21.8% of these genes were respectively called present and absent in all samples. These genes were considered universally active or inactive in LCL regardless of gender, disease state, or other factors. The details about data processing, gene filtering and annotation mapping are available in Supplemental methods.

The methylation–expression association of 19,615 pairs of CpG sites and genes is summarized in Fig. 4A, which illustrates that the association is dependant on the relative location of CpG sites to the TSSs of downstream genes. In the boxed area of Fig. 5A, the overall methylation–expression correlation is negative and highly significant (Spearman's $\rho = -0.54$, $p = 1.8e\text{-}272$). However, the correlation was non-linear. When β value is higher than 0.1, methylation level has little impact on gene expression level ($\rho = 0.013$, $p = 0.67$). The overall pattern in Fig. 5A is consistent with the common perception of how DNA methylation regulates gene expression. Around the TSS and in the 5′-UTR, methylation regulates downstream transcription mainly through physical blocking, so high methylation represses gene expression. In the promoter region, DNA methylation indirectly regulates transcription by adjusting histone and transcription factor accessibility, so its correlation to gene expression could be in either direction. A previous study reported that exons usually have high methylation and the gene body of highly expressed genes is more methylated than the body of inactive genes [6]. This result explains the concurrence of high methylation and high expression in the region beginning approximately 1 kb downstream of the TSS where the coding region of some genes has started.

Consistent with earlier results demonstrating that methylation at TFBSs varies dramatically between transcription factors, the transcription of TF target genes also differs significantly in LCL (Supplemental Table 2). The TFBS motif having the lowest average methylation, V$NRF2_01, has the highest average expression of target genes, which is more than 250% of the global average of gene expression. On the other hand, V$HNF1_01 has the highest methylation and lowest target expression. The average methylation and expression have a negative correlation of 0.743 across TFs and change concordantly in a linear pattern (Fig. 4B). Interestingly, ten TFBS motifs have both lower methylation and expression than the global average, probably because lower methylation reduces the binding of those TFs or the increased binding of those TFs is repressive to gene expression. Overall, the combined analysis of genomic sequence, DNA methylation, and gene expression information provides a reference of TF activity in LCL. It also indicates that DNA methylation may affect gene expression through regulating TF binding.

We demonstrated in our previous study that changes in DNA methylation contributes little to the overall gene expression variation in CdLS [30]. Although more than one thousand genes demonstrate significantly changed expression in CdLS, the changes usually are of a small magnitude. A deficiency in cohesin or other transcription factors is more likely to be the cause of these subtle expression changes rather than differences in DNA methylation. The latter is probably involved in more dramatic events such as gene activation or inactivation. We hypothesized that a methylation–expression association is more evident when large between-sample variance exists. The methylation–expression correlation across 27 samples common to both studies were calculated for 1006 CpG-gene pairs on autosomes whose between-sample variance is at the top 25% in both methylation and transcription data. As shown in Fig. 4C, the overall correlation is skewed slightly towards the negative side (average Spearman's $\rho = -0.043$, $p = 2.5e\text{-}8$). There are 38 pairs having ρ values less than -0.5, corresponding to a permutation FDR of 0.12. Six genes, *C21orf56*, *CIDEB*, *DDX43*, *DENND2D*, *LDHC*, and *LOXL3*, have ρ values less than -0.5 with two CpG sites around their TSSs, indicating that the expression of these genes is more likely to be directly regulated by DNA methylation. There are 17 pairs having ρ values greater than 0.5 (FDR = 0.27). On average, CpG sites positively correlated to gene expression are located more upstream than sites negatively correlated to gene expression. (−250 vs. +155 bases of TSS, $p = 0.04$).

3.5. Prediction of gene transcription based on epigenetic status around the TSS

Fig. 4A shows that DNA methylation at TSSs and 5′-UTRs has a relatively consistent association with gene expression in LCLs. We queried whether DNA methylation in those regions could be used to predict the activation of downstream transcription. The analysis was first limited to 3048 genes that are unanimously active or inactive according to expression microarray data and have at least one measured CpG site located between 100 bases upstream of the TSS and 100 bases downstream of the TSS or the end of 5′-UTR (whichever comes first). These genes were split into two groups of 887 inactive and 2161 active genes. The vast majority (~95%) active genes have low methylation ($\beta < 0.1$), while only about two-thirds of inactive genes have non-low methylation ($\beta > 0.1$) around their TSS (Fig. 5A). Therefore, low methylation is a necessary, but not a sufficient condition of downstream transcription, which also involves histone modifications, TF binding, and polymerase activation. The X chromosome demonstrated an interesting pattern (Supplemental Fig. 12). Due to X inactivation in females, the majority of X-linked genes have non-low methylation regardless of whether they are expressed or not; while in males, almost all active genes have low methylation. Among the 102 genes that have active expression but non-low methylation, 79 have no CGI around their TSS, so inhibitory action of high methylation on downstream transcription is weakened when genes have low CpG density around TSS.

We applied a training–testing procedure during which 2048 genes were randomly selected to train a logistic regression model and the remaining 1000 genes were used to test model performance. This procedure was repeated 100 times to remove sampling bias. As expected, DNA methylation around the TSS is a highly sensitive, but non-specific, predictor of active transcription and its model outperforms the model using CpG density as predictor (Table 1A). The multivariate model using both CpG density and methylation as predictors has improved performance and predicted gene activation with 84% accuracy. Models using our previous cohesin binding data [19] as well as CTCF and histone data of GM12878 as predictors were also created and tested. Remarkably, some histone modifications are strong predictors of gene expression even although the data were generated independently from unrelated sample. For example, H3K9ac alone can predict gene expression with 93.7% accuracy, 93.6% sensitivity and 94.1% specificity. Finally, we integrated all available predictors into one multivariate model, which has better and more balanced performance than univariate models.

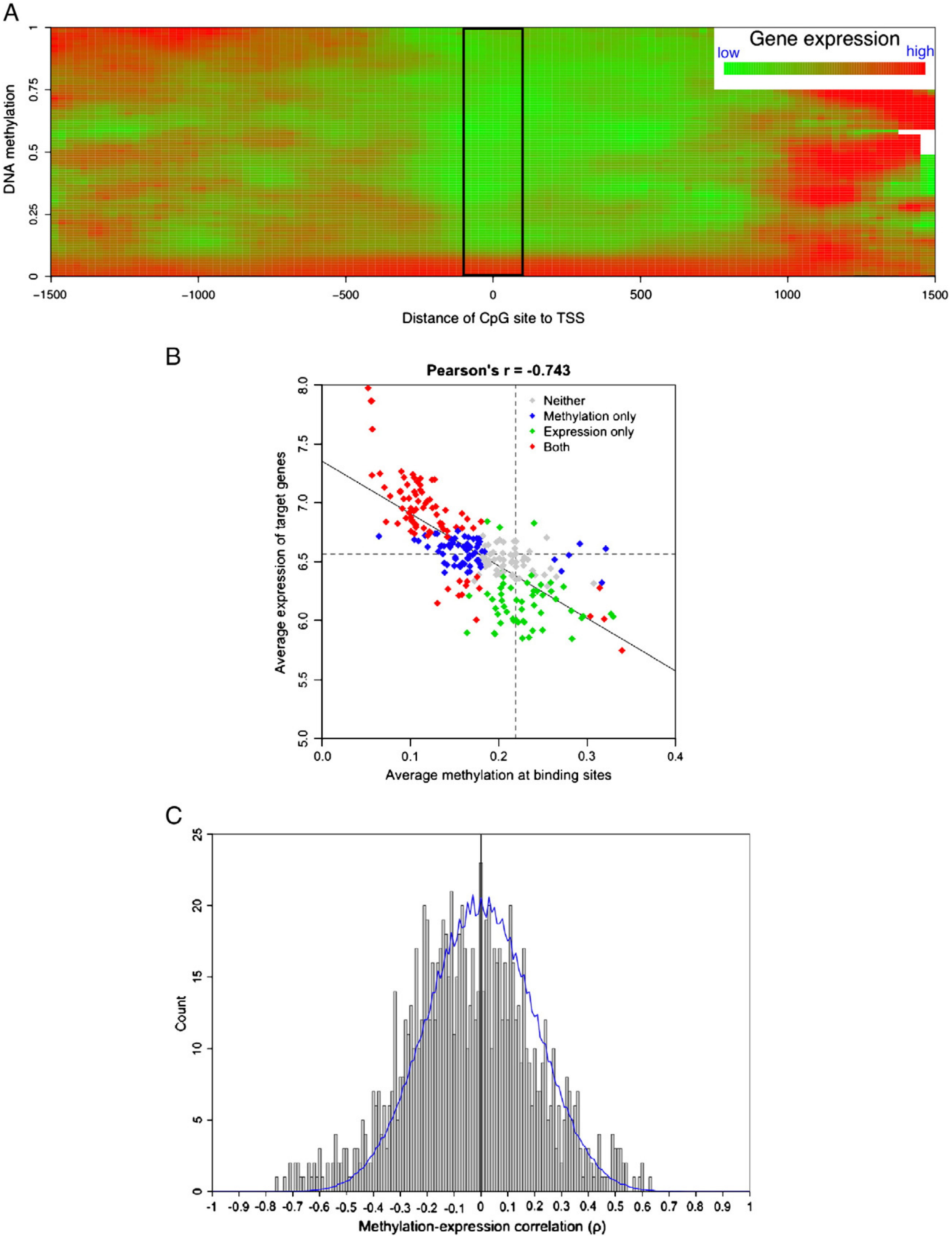

Fig. 4. The association between DNA methylation around TSS and downstream gene transcription. A) The distance of the CpG sites to TSS affected the methylation-transcription association based on 19,615 CpG-transcript pairs. Color indicates average transcription level after local smoothing. The black box shows a distinctive pattern of negative correlation when the CpG sites are located within [−100, 100] around TSSs. B) The linearly correlated methylation around TFBS and expression of target genes. Each diamond corresponds to one of 258 TFBS motifs. X-axis represents the average methylation of CpG sites within [−250, 250] around TFBS and y-axis represents average expression of genes with at least one TFBS within [−1500, 1500] of their TSS. Color indicates whether the averages are significantly ($p<0.01$) different from the global average. More details are available in Supplemental Table 1. C) The distribution of Spearman's ρ values of CpG-gene pairs. All pairs include CpG site and gene both having high variance across 27 common samples. Blue line indicates the background distribution generated by 1000 re-sampling permutations. Among a total of 1006 pairs, 568 (56.5%) have negative ρ values.

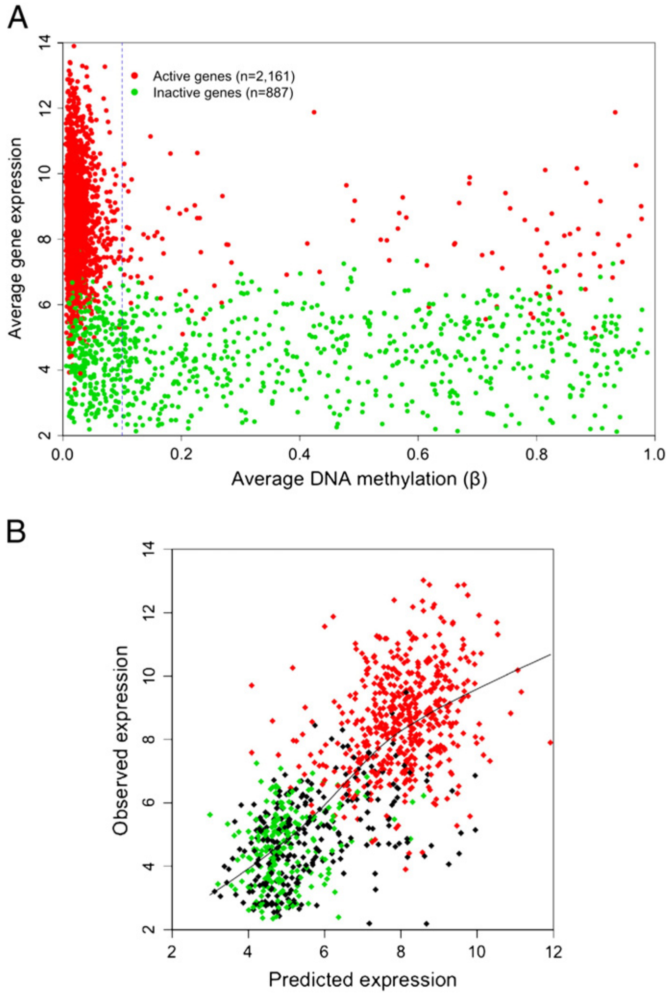

Fig. 5. The prediction of gene expression using epigenetic information. A) The average DNA methylation at TSS and 5′-UTR and the average expression of autosomal genes that are active (red) or inactive (green) in all samples. X chromosome genes are plotted separately in Supplemental Fig. 12. B) Predicted vs. observed expression level of 1000 testing genes (green: inactive, red: active; black partially active). The prediction is based on a linear regression model trained with DNA methylation, cohesin, CTCF and histone modification data of 3100 genes.

Predicting gene expression level is more challenging than predicting gene activation as epigenetic modifications are not the only regulators of gene expression. Downstream regulators such as transcription factors and miRNAs are probably more important to the tuning of gene expression levels. In addition, gene expression measurements are biased by hybridization efficiency of microarray probes, so they are not exactly correlated to mRNA abundance. To evaluate the predictive ability of epigenetic status on gene expression level, we applied the same training–testing/permutation procedure and used the same predictors to build a series of linear regression models during which 1052 poised genes (active in part of the 39 samples) were added to make a pool of 4100 genes in total. While 3100 genes were randomly selected for training models, the performance of the models was evaluated by the correlation between predicted and observed expression level of the remaining 1000 genes. The average Pearson's r of 100 permutations is 0.726 with the multivariate model of all available predictors (Table 1B). However, the correlation is mainly determined by the dramatic difference between inactive and active genes (Fig. 6B). When using active genes only, the correlation between expected and observed expression level was substantially decreased, but still significant (average r = 0.272, p = 2.2e-10). The involvement of H3K9ac in gene expression is supported by both types of models and its univariate model has performance close to the full model. The regulatory function of H3K9ac on gene expression has been reported in T cells [41].

4. Discussions

One of the most striking observations of this study is the high variability of methylation at CpG sites while known clinical factors, such as gender and disease-causing DNA mutations, only account for a small portion of the total variance between samples. The total variance is a composite of the following components: 1) measurement error of the microarray experiments; 2) bias introduced by bisulfate treatment or other DNA preparation steps; 3) methylation alteration caused by EBV transformation and cell culturing; and 4) variance inherited from the donors. Since the replicated measurements of the same samples have strong correlation (r = 0.99) between two microarray batches and the fact that we applied a strict filtering procedure to exclude questionable measurements and samples from data analysis, the contribution of measurement error to the total variance is minimized. Replicated samples in the same microarray batch, but processed separately through DNA preparation, also have much better correlation to each other than the correlation of any pairs of different samples (Supplemental File 1). We thus concluded that the majority of the total variance is not from the microarray experiments.

Virus transformation and continuous cell culturing and storage may contribute more to the total variance although our meta-analysis shows that LCLs are more similar to lymphocytes than other cell types (Supplemental File 2). Previous studies have reported altered methylation in LCLs at different chromosomal locations, questioning the fidelity of DNA methylation in LCLs to its in vivo status [27,29]. Furthermore, the meta-analysis demonstrates that LCLs have a much larger between-sample variance than most types of fresh cells. Similarly, Grafodatskaya et al. recently observed that LCLs have larger between-sample variance than white blood cells and suggested that methylation alteration in LCLs occurs at random locations [28]. Nevertheless, our results show that GM12878 and LCLs used in this study have highly correlated DNA methylation pattern although as a model cell line, GM12878 has been cultured for many generations (Supplemental Fig. 9). Conversely, if methylation alterations take place randomly and accumulatively, we would observe a reduced between-sample correlation over generations.

We postulate that while the biological systems such as the methylome and the transcriptome go through certain alterations during the establishment of LCLs, they will maintain a relatively stable status during cell culturing. Furthermore, the alterations are not random events, so LCLs generated separately from the same donor will have more similar methylation and transcription patterns than those generated from different donors. If proved true, this feature of LCLs will advocate its value in studying genetic disorders. Unlike fresh cells whose status is usually confounded by many uncontrollable factors, cultured cells are more homogeneous and grown under controlled environment. The effect of etiological mutations on biological systems is more isolated and recognizable in cell lines. For example, our previous study used LCLs to identify over one thousand genes significantly dysregulated in CdLS [19] while the magnitude of differential expression is mostly too small to be detected in fresh cells due to their lack of homogeneity. Therefore, LCLs are often a more practical experimental material for studying how etiologic mutations cause abnormalities in downstream systems although it cannot preserve the complexity of in vivo status as fresh or primary cells do. Future experiments that trace methylation alterations throughout LCL culturing and compare methylation patterns before and after EBV transformation in multiple sample groups will more conclusively test these hypotheses.

Although current data cannot directly validate that biological systems in different LCLs will reach and maintain a stable status, this

Table 1
Predict gene expression with regression models. Results in the tables are the summary of 100 re-sampling permutations of genes and each permutation randomly 1000 genes to test the models trained with the other genes. “NA” indicates that the model has no better performance than random prediction. A) Gene expression is represented as a binomial variable (whether a gene is expressed) and logistic regression is used for the modeling. The three best values are highlighted in bold. B). Gene expression is represented as continuous variable (how much a gene is expressed) and so linear regression is used instead. Numbers indicate the Pearson's correlation coefficients between predicted and observed expression level. Island = within or out of CpG island; Meth = average DNA methylation; AUC = area under ROC curve; ACC = accuracy of prediction; SENS = sensitivity; SPEC = specificity; PPV = positive predictive value; and NPV = negative predictive value. All genes include active, inactive, and partially active genes.

	Univariate												Multivariate	
	Island	Meth	Cohesin	CTCF	H3K4me1	H3K4me2	H3K4me3	H3K9ac	H3K27ac	H3K27me3	H3K36me3	H4K20me1	Island*Meth	Full
Table 1A Prediction of whether genes were active in LCLs by logistic regression models														
AUC	71.1	89.5	80.5	83.8	65.3	90.1	96.4	**97.0**	**97.1**	85.7	60.3	71.4	88.2	**97.4**
ACC	75.8	81.7	77.8	71.0	70.9	91.1	**93.7**	**93.7**	93.5	85.7	70.9	71.8	84.0	**95.1**
SENS	82.3	**96.8**	91.8	NA	NA	95.1	95.6	93.6	92.9	**97.0**	NA	94.2	95.8	**96.7**
SPEC	59.8	44.9	43.7	NA	NA	81.4	89.2	**94.1**	**94.8**	58.4	NA	17.2	55.3	**91.3**
PPV	83.3	81.1	80.0	NA	NA	92.6	95.6	**97.5**	**97.8**	85.1	NA	73.5	83.9	**96.4**
NPV	58.1	85.4	68.6	NA	NA	87.3	**89.2**	85.8	84.6	**88.8**	NA	55.1	84.4	**91.9**
Table 1B Prediction of how much genes were expressed by linear regression models														
All genes	NA	0.40	0.31	0.16	0.10	0.47	**0.65**	**0.66**	0.61	0.48	0.65	0.19	0.44	**0.73**
Active	NA	0.10	0.03	−0.03	−0.03	−0.02	0.13	**0.28**	**0.29**	0.08	0.00	−0.06	0.09	**0.28**
Inactive	NA	−0.04	0.01	−0.02	0.06	0.05	0.10	**0.13**	0.10	**0.13**	−0.04	−0.11	−0.02	**0.16**

hypothesis is strongly supported by the fact that histone modifications in GM12878 alone can predict gene expression in an unrelated sample set with 95% accuracy (Table 1B). This result also suggests that epigenetic status is the determinant factor of gene activation in LCL. The prediction accuracy is remarkably high considering the existence of a few technical difficulties, such as the small number of samples used for most predictive variables and the possible error of mapping epigenetic status around TSS to 3′-biased expression measurements. Therefore, the actual impact of epigenetic status on transcription activation could be even higher.

This study also suggests that a β of 0.1, or the methylation of 10% of alleles in a cell population, is enough to indicate gene inactivation (Fig. 4A). This is unlikely a consequence of biased Cy3/Cy5 measurements because the average β value of all three universally methylated controls is higher than 0.9. This observation brings up a series of questions. What is the cause of such heterogeneity? If a β of 0.1 is enough to inactivate transcription, is it necessary for cells to further increase methylation levels? If maintaining hyper-methylation status around the TSS requires extra energy, does it present an evolutionary disadvantage? In CdLS, there is a trend towards higher methylation levels at non-CGI sites (Fig. 2B), which usually has no effect on gene expression since most of the downstream genes are already silenced in control samples. Whether this represents a dysfunctional regulatory system of DNA methylation in CdLS will be one of the topics of our future studies.

Supplementary materials related to this article can be found online at doi:10.1016/j.ygeno.2012.01.002.

Acknowledgments

We are grateful for the participation of the children and families with CdLS and to the CdLS Foundation for their support. The authors also acknowledge support from Dr. Xiaowu Gai and Juan Perin of Center for Biomedical Informatics at CHOP. This study was supported by NIH/NICHD PO1HD052860 (IDK, ZZ) and NIH/NICHD R21HD050538 (IDK).

References

[1] T. Mohandas, R. Sparkes, L. Shapiro, Reactivation of an inactive human X chromosome: evidence for X inactivation by DNA methylation, Science 211 (1981) 393–396.

[2] E. Li, C. Beard, R. Jaenisch, Role for DNA methylation in genomic imprinting, Nature 366 (1993) 362–365.

[3] P. Jones, G. Veenstra, P. Wade, D. Vermaak, S. Kass, N. Landsberger, J. Strouboulis, A. Wolffe, Methylated DNA and MeCP2 recruit histone deacetylase to repress transcription, Nat. Genet. 19 (1998) 187–191.

[4] K. Robertson, DNA methylation and human disease, Nat. Rev. Genet. 6 (2005) 597–610.

[5] M. Okano, D. Bell, D. Haber, E. Li, DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development, Cell 99 (1999) 247–257.

[6] L. Laurent, E. Wong, G. Li, T. Huynh, A. Tsirigos, C. Ong, H. Low, K. Kin Sung, I. Rigoutsos, J. Loring, C. Wei, Dynamic changes in the human methylome during differentiation, Genome Res. 20 (2010) 320–331.

[7] M. Ehrich, J. Turner, P. Gibbs, L. Lipton, M. Giovanneti, C. Cantor, D. van den Boom, Cytosine methylation profiling of cancer cell lines, Proc. Natl. Acad. Sci. U. S. A. 105 (2008) 4844–4849.

[8] A. Brunner, D. Johnson, S. Kim, A. Valouev, T. Reddy, N. Neff, E. Anton, C. Medina, L. Nguyen, E. Chiao, C. Oyolu, G. Schroth, D. Absher, J. Baker, R. Myers, Distinct DNA methylation patterns characterize differentiated human embryonic stem cells and developing human fetal liver, Genome Res. 19 (2009) 1044–1056.

[9] M. Gardiner-Garden, M. Frommer, CpG islands in vertebrate genomes, J. Mol. Biol. 196 (1987) 261–282.

[10] F. Antequera, A. Bird, Number of CpG islands and genes in human and mouse, Proc. Natl. Acad. Sci. U. S. A. 90 (1993) 11995–11999.

[11] M. Esteller, P. Corn, S. Baylin, J. Herman, A gene hypermethylation profile of human cancer, Cancer Res. 61 (2001) 3225–3229.

[12] Y. Akiyama, C. Maesawa, S. Ogasawara, M. Terashima, T. Masuda, Cell-type-specific repression of the maspin gene is disrupted frequently by demethylation at the promoter region in gastric intestinal metaplasia and cancer cells, Am. J. Pathol. 163 (2003) 1911–1919.

[13] C. Tufarelli, J. Stanley, D. Garrick, J. Sharpe, H. Ayyub, W. Wood, D. Higgs, Transcription of antisense RNA leading to gene silencing and methylation as a novel cause of human genetic disease, Nat. Genet. 34 (2003) 157–165.

[14] H. Nakagawa, G. Nuovo, E. Zervos, E.J. Martin, R. Salovaara, L. Aaltonen, A. de la Chapelle, Age-related hypermethylation of the 5′ region of MLH1 in normal colonic mucosa is associated with microsatellite-unstable colorectal cancer development, Cancer Res. 61 (2001) 6991–6995.

[15] B. Javierre, A. Fernandez, J. Richter, F. Al-Shahrour, J. Martin-Subero, J. Rodriguez-Ubreva, M. Berdasco, M. Fraga, T. O'Hanlon, L. Rider, F. Jacinto, F. Lopez-Longo, J. Dopazo, M. Forn, M. Peinado, L. Carreño, A. Sawalha, J. Harley, R. Siebert, M. Esteller, F. Miller, E. Ballestar, Changes in the pattern of DNA methylation associate with twin discordance in systemic lupus erythematosus, Genome Res. 20 (2010) 170–179.

[16] M. Weber, J. Davies, D. Wittig, E. Oakeley, M. Haase, W. Lam, D. Schübeler, Chromosome-wide and promoter-specific analyses identify sites of differential DNA methylation in normal and transformed human cells, Nat. Genet. 37 (2005) 853–862.

[17] Y. Koga, M. Pelizzola, E. Cheng, M. Krauthammer, M. Sznol, S. Ariyan, D. Narayan, A. Molinaro, R. Halaban, S. Weissman, Genome-wide screen of promoter methylation identifies novel markers in melanoma, Genome Res. 19 (2009) 1462–1470.

[18] M. Sugimoto, H. Tahara, T. Ide, Y. Furuichi, Steps involved in immortalization and tumorigenesis in human B-lymphoblastoid cell lines transformed by Epstein–Barr virus, Cancer Res. 64 (2004) 3361–3364.

[19] J. Liu, Z. Zhang, M. Bando, T. Itoh, M. Deardorff, D. Clark, M. Kaur, S. Tandy, T. Kondoh, E. Rappaport, N. Spinner, H. Vega, L. Jackson, K. Shirahige, I. Krantz, Transcriptional dysregulation in NIPBL and cohesin mutant human cells, PLoS Biol. 7 (2009) e1000119.

[20] Y. Nishimura, C. Martin, A. Vazquez-Lopez, S. Spence, A. Alvarez-Retuerto, M. Sigman, C. Steindler, S. Pellegrini, N. Schanen, S. Warren, D. Geschwind, Genome-wide expression profiling of lymphoblastoid cell lines distinguishes different forms of autism and reveals shared pathways, Hum. Mol. Genet. 16 (2007) 1682–1698.

[21] E. Choy, R. Yelensky, S. Bonakdar, R. Plenge, R. Saxena, P. De Jager, S. Shaw, C. Wolfish, J. Slavik, C. Cotsapas, M. Rivas, E. Dermitzakis, E. Cahir-McFarland, E.

Kieff, D. Hafler, M. Daly, D. Altshuler, Genetic analysis of human traits in vitro: drug response and gene expression in lymphoblastoid cell lines, PLoS Genet. 4 (2008) e1000287.

[22] C. Tuck-Muller, A. Narayan, F. Tsien, D. Smeets, J. Sawyer, E. Fiala, O. Sohn, M. Ehrlich, DNA hypomethylation and unusual chromosome instability in cell lines from ICF syndrome patients, Cytogenet. Cell Genet. 89 (2000) 121–128.

[23] K. Frazer, D. Ballinger, D. Cox, D. Hinds, L. Stuve, R. Gibbs, J. Belmont, A. Boudreau, P. Hardenbol, S. Leal, S. Pasternak, D. Wheeler, T. Willis, F. Yu, H. Yang, C. Zeng, Y. Gao, H. Hu, W. Hu, C. Li, W. Lin, S. Liu, H. Pan, X. Tang, J. Wang, W. Wang, J. Yu, B. Zhang, Q. Zhang, H. Zhao, J. Zhou, S. Gabriel, R. Barry, B. Blumenstiel, A. Camargo, M. Defelice, M. Faggart, M. Goyette, S. Gupta, J. Moore, H. Nguyen, R. Onofrio, M. Parkin, J. Roy, E. Stahl, E. Winchester, L. Ziaugra, D. Altshuler, Y. Shen, Z. Yao, W. Huang, X. Chu, Y. He, L. Jin, Y. Liu, W. Sun, H. Wang, Y. Wang, X. Xiong, L. Xu, M. Waye, S. Tsui, H. Xue, J. Wong, L. Galver, J. Fan, K. Gunderson, S. Murray, A. Oliphant, M. Chee, A. Montpetit, F. Chagnon, V. Ferretti, M. Leboeuf, J. Olivier, M. Phillips, S. Roumy, C. Sallée, A. Verner, T. Hudson, P. Kwok, D. Cai, D. Koboldt, R. Miller, L. Pawlikowska, P. Taillon-Miller, M. Xiao, L. Tsui, W. Mak, Y. Song, P. Tam, Y. Nakamura, T. Kawaguchi, T. Kitamoto, T. Morizono, A. Nagashima, Y. Ohnishi, A. Sekine, T. Tanaka, T. Tsunoda, P. Deloukas, C. Bird, M. Delgado, E. Dermitzakis, R. Gwilliam, S. Hunt, J. Morrison, D. Powell, B. Stranger, P. Whittaker, D. Bentley, M. Daly, P. de Bakker, J. Barrett, Y. Chretien, J. Maller, S. McCarroll, N. Patterson, I. Pe'er, A. Price, S. Purcell, D. Richter, P. Sabeti, R. Saxena, S. Schaffner, P. Sham, P. Varilly, L. Stein, L. Krishnan, A. Smith, M. Tello-Ruiz, G. Thorisson, A. Chakravarti, P. Chen, D. Cutler, C. Kashuk, S. Lin, G. Abecasis, W. Guan, Y. Li, H. Munro, Z. Qin, D. Thomas, G. McVean, A. Auton, L. Bottolo, N. Cardin, S. Eyheramendy, C. Freeman, J. Marchini, S. Myers, C. Spencer, M. Stephens, P. Donnelly, L. Cardon, G. Clarke, D. Evans, A. Morris, B. Weir, J. Mullikin, S. Sherry, M. Feolo, A. Skol, H. Zhang, I. Matsuda, Y. Fukushima, D. Macer, E. Suda, C. Rotimi, C. Adebamowo, I. Ajayi, T. Aniagwu, P. Marshall, C. Nkwodimmah, C. Royal, M. Leppert, M. Dixon, A. Peiffer, R. Qiu, A. Kent, K. Kato, N. Niikawa, I. Adewole, B. Knoppers, M. Foster, E. Clayton, J. Watkin, D. Muzny, L. Nazareth, E. Sodergren, G. Weinstock, I. Yakub, B. Birren, R. Wilson, L. Fulton, J. Rogers, J. Burton, N. Carter, C. Clee, M. Griffiths, M. Jones, K. McLay, R. Plumb, M. Ross, S. Sims, D. Willey, Z. Chen, H. Han, L. Kang, M. Godbout, J. Wallenburg, P. L'Archevêque, G. Bellemare, K. Saeki, D. An, H. Fu, Q. Li, Z. Wang, R. Wang, A. Holden, L. Brooks, J. McEwen, M. Guyer, V. Wang, J. Peterson, M. Shi, J. Spiegel, L. Sung, L. Zacharia, F. Collins, K. Kennedy, R. Jamieson, J. Stewart, I.H. Consortium, A second generation human haplotype map of over 3.1 million SNPs, Nature 449 (2007) 851–861.

[24] G. Weinstock, ENCODE: more genomic empowerment, Genome Res. 17 (2007) 667–668.

[25] T. Mikkelsen, M. Ku, D. Jaffe, B. Issac, E. Lieberman, G. Giannoukos, P. Alvarez, W. Brockman, T. Kim, R. Koche, W. Lee, E. Mendenhall, A. O'Donovan, A. Presser, C. Russ, X. Xie, A. Meissner, M. Wernig, R. Jaenisch, C. Nusbaum, E. Lander, B. Bernstein, Genome-wide maps of chromatin state in pluripotent and lineage-committed cells, Nature 448 (2007) 553–560.

[26] D. Araten, D. Golde, R. Zhang, H. Thaler, L. Gargiulo, R. Notaro, L. Luzzatto, A quantitative measurement of the human somatic mutation rate, Cancer Res. 65 (2005) 8111–8117.

[27] E. Brennan, M. Ehrich, D. Brazil, J. Crean, M. Murphy, D. Sadlier, F. Martin, C. Godson, A. McKnight, D. van den Boom, A. Maxwell, D. Savage, Comparative analysis of DNA methylation profiles in peripheral blood leukocytes versus lymphoblastoid cell lines, Epigenetics 4 (2009) 159–164.

[28] D. Grafodatskaya, S. Choufani, J. Ferreira, D. Butcher, Y. Lou, C. Zhao, S. Scherer, R. Weksberg, EBV transformation and cell culturing destabilizes DNA methylation in human lymphoblastoid cell lines, Genomics 95 (2010) 73–83.

[29] A. Saferali, E. Grundberg, S. Berlivet, H. Beauchemin, L. Morcos, C. Polychronakos, T. Pastinen, J. Graham, B. McNeney, A. Naumova, Cell culture-induced aberrant methylation of the imprinted IG DMR in human lymphoblastoid cell lines, Epigenetics 5 (2010) 50–60.

[30] J. Liu, Z. Zhang, M. Bando, T. Itoh, M. Deardorff, J. Li, D. Clark, M. Kaur, K. Tatsuro, A. Kline, C. Chang, H. Vega, L. Jackson, N. Spinner, K. Shirahige, I. Krantz, Genome-wide DNA methylation analysis in cohesin mutant human cell lines, Nucleic Acids Res. (2010).

[31] J. Liu, I.D. Krantz, Cornelia de Lange syndrome, cohesin, and beyond, Clin. Genet. 76 (2009) 303–314.

[32] C. Workman, L. Jensen, H. Jarmer, R. Berka, L. Gautier, H. Nielser, H. Saxild, C. Nielsen, S. Brunak, S. Knudsen, A new non-linear normalization method for reducing variability in DNA microarray experiments, Genome Biol. 3 (2002) research0048.

[33] d.W. Huang, B. Sherman, R. Lempicki, Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources, Nat. Protoc. 4 (2009) 44–57.

[34] A. Bird, DNA methylation patterns and epigenetic memory, Genes Dev. 16 (2002) 6–21.

[35] P. Sabo, M. Kuehn, R. Thurman, B. Johnson, E. Johnson, H. Cao, M. Yu, E. Rosenzweig, J. Goldy, A. Haydock, M. Weaver, A. Shafer, K. Lee, F. Neri, R. Humbert, M. Singer, T. Richmond, M. Dorschner, M. McArthur, M. Hawrylycz, R. Green, P. Navas, W. Noble, J. Stamatoyannopoulos, Genome-scale mapping of DNase I sensitivity in vivo using tiling DNA microarrays, Nat. Methods 3 (2006) 511–518.

[36] S. Chambeyron, W. Bickmore, Chromatin decondensation and nuclear reorganization of the HoxB locus upon induction of transcription, Genes Dev. 18 (2004) 1119–1130.

[37] T. Kouzarides, Chromatin modifications and their function, Cell 128 (2007) 693–705.

[38] J. Wu, S. Wang, D. Potter, J. Liu, L. Smith, Y. Wu, T. Huang, C. Plass, Diverse histone modifications on histone 3 lysine 9 and their relation to DNA methylation in specifying gene silencing, BMC Genomics 8 (2007) 131.

[39] M. Dai, P. Wang, A. Boyd, G. Kostov, B. Athey, E. Jones, W. Bunney, R. Myers, T. Speed, H. Akil, S. Watson, F. Meng, Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data, Nucleic Acids Res. 33 (2005) e175.

[40] W. Liu, R. Mei, X. Di, T. Ryder, E. Hubbell, S. Dee, T. Webster, C. Harrington, M. Ho, J. Baid, S. Smeekens, Analysis of high density expression microarrays with signed-rank call algorithms, Bioinformatics 18 (2002) 1593–1599.

[41] M. Fann, J. Godlove, M. Catalfamo, W.r. Wood, F. Chrest, N. Chun, L. Granger, R. Wersto, K. Madara, K. Becker, P. Henkart, N. Weng, Histone acetylation is associated with differential gene expression in the rapid and robust memory CD8(+) T-cell response, Blood 108 (2006) 3363–3370.