OXFORD

Briefings in Bioinformatics, 2016, 1–12

doi: 10.1093/bib/bbw098
Paper

The framework for population epigenetic study

Linna Zhao*, Di Liu*, Jing Xu*, Zhaoyang Wang*, Yang Chen, Changgui Lei, Ying Li, Guiyou Liu and Yongshuai Jiang

Corresponding authors: Yongshuai Jiang, College of Bioinformatics Science and Technology, Harbin Medical University, Harbin, China, and Training Center for Students Innovation and Entrepreneurship Education, Harbin Medical University, Harbin, China. E-mail: jiangyongshuai@gmail.com or jiangyongshuai@ems.hrbmu.edu.cn; Guiyou Liu, School of Life Science and Technology, Harbin Institute of Technology, Harbin, China. E-mail: 16b328001@hit.edu.cn
*These authors contributed equally to this work.

Abstract

At present, understanding of DNA methylation at the population level is still limited. Here, we first extended the classical framework of population genetics, such as single nucleotide polymorphism allele frequency, linkage disequilibrium (LD), LD block and haplotype, to epigenetics. Then, as an example, we compared the DNA methylation disequilibrium (MD) maps between HapMap CEU (Caucasian residents of European ancestry from Utah) population and YRI (Yoruba people from Ibadan) population (lymphoblastoid cell lines). We analyzed the differences and similarities between CEU and YRI from the following aspects: SMP (single methylation polymorphism) allele frequency, SMP allele association, MD, MD block and methylation haplotype (meplotype) frequency. The results showed that CEU and YRI had similar distribution of SMP allele frequency, and shared many MD block region. We believe that the framework of population genetics can be used in the population epigenetics. The population epigenetic framework also has potential prospects in the study of complex diseases, such as epigenome-wide association study.

Key words: DNA methylation; population epigenetics; SMP; single methylation polymorphism; population genetics; epigenome-wide association study

Linna Zhao is a master at College of Bioinformatics Science and Technology, Harbin Medical University. She is also a member of genetic testing group at Training Center for Students Innovation and Entrepreneurship Education of Harbin Medical University. Her research interest focuses on statistical genetics.
Di Liu is a master at College of Bioinformatics Science and Technology, Harbin Medical University. She is also a member of genetic testing group at Training Center for Students Innovation and Entrepreneurship Education of Harbin Medical University. Her research interest focuses on statistical genetics.
Jing Xu is a bachelor at College of Bioinformatics Science and Technology, Harbin Medical University. She is also a member of genetic testing group at Training Center for Students Innovation and Entrepreneurship Education of Harbin Medical University. Her research interest focuses on statistical genetics.
Zhaoyang Wang is a bachelor at College of Bioinformatics Science and Technology, Harbin Medical University. She is also a member of genetic testing group at Training Center for Students Innovation and Entrepreneurship Education of Harbin Medical University. Her research interest focuses on statistical genetics.
Yang Chen is a bachelor at College of Bioinformatics Science and Technology, Harbin Medical University. He is also a member of genetic testing group at Training Center for Students Innovation and Entrepreneurship Education of Harbin Medical University. His research interest focuses on statistical genetics.
Changgui Lei is a bachelor at College of Bioinformatics Science and Technology, Harbin Medical University. He is also a member of genetic testing group at Training Center for Students Innovation and Entrepreneurship Education of Harbin Medical University. His research interest focuses on statistical genetics.
Ying Li is a bachelor training at College of Bioinformatics Science and Technology, Harbin Medical University. She is also a member of genetic testing group at Training Center for Students Innovation and Entrepreneurship Education of Harbin Medical University. Her research interest focuses on statistical genetics.
Guiyou Liu is an assistant professor and PhD candidate at School of Life Science and Technology, Harbin Institute of Technology. His research interest focuses on bioinformatics, population genetics and statistical genetics.
Yongshuai Jiang is an associate professor at College of Bioinformatics Science and Technology, Harbin Medical University. He is the leader of genetic testing group at Training Center for Students Innovation and Entrepreneurship Education of Harbin Medical University. His research interest focuses on bioinformatics, population genetics and population epigenetics.
Submitted: 14 July 2016; **Received (in revised form):** 11 September 2016

Introduction

DNA methylation is an important epigenetic mechanism that occurs by adding a methyl group to DNA [1, 2]. It often plays important roles in the regulation of gene expression [3, 4], aging [5, 6], genomic imprinting [7, 8], X-chromosome inactivation [9] and development [10–12].

Although the role of DNA methylation in the modification of gene function is widely studied, the understanding of population characteristics of DNA methylation is still limited. In this study, we do not consider the dynamic changes of DNA methylation, only discuss the population characteristics. Some previous studies had indicated that DNA methylation site can form polymorphisms (also known as single methylation polymorphisms, SMPs) over an evolutionary timescale, and researchers have successfully carried out some population epigenetic studies for plant and animal genomes [13–18].

For human, Shoemaker et al. applied the principle of linkage disequilibrium (LD) to analyze the DNA methylation data in 16 cell lines, and observed the correlation of methylation between adjacent CpG loci [19]. Moen et al. conducted a population study using DNA methylation data [20]. They detected the cytosine modification levels (using the Illumina HumanMethylation450 BeadChip) of 133 lymphoblastoid cell lines (LCLs) derived from individuals of Yoruba people from Ibadan, Nigeria (YRI) and Caucasian residents of European ancestry from Utah (CEU). All the samples were from the International HapMap Project [21, 22]. Using the population data, they identified some differentially methylation loci between the CEU and YRI, assessed the genetic contribution to epigenetic differences and analyzed the role of modification quantitative trait loci (mQTL) on gene expression [20]. As an important application, they also build a SCAN database to facilitate the search of expression quantitative trait loci and mQTL [23]. Though they have achieved great success in explaining the relationship between genetic, epigenetic and expression, they only provide limited information on population epigenetic structure itself.

In this study, we did not consider how the genetic and epigenetic influence expression. Our aim is to elaborate and illustrate the population characteristics of DNA methylation. We extended the classical methods of population genetics to epigenetics, and constructed a population epigenetic analysis system. We then compared the differences and similarities of epiallele frequency, epiallele association, methylation disequilibrium (MD), MD block and meplotype (methylation haplotype) frequency between CEU population and YRI population.

Extended the classical framework of population genetics to epigenetics

Convert DNA methylation β-value into methylation genotype data

To compare the DNA MD maps with LD maps, we put the DNA methylation level data and single nucleotide polymorphism (SNP) genotype data into a unified framework, that is, we convert the DNA methylation β-value into methylation genotype data.

For a CpG locus, the methylation status β-value can be described as β=methylated signals/(methylated signals+ unmethylated signals + α). For Illumina chip, the α is set to 100 [24]. For an individual, if he/she has two modified cytosine at the same CpG locus on both members of a pair of homologous chromosomes, he/she would be detected with a high level of cytosine modifications by methylation chip, and will have a higher β-value. If he/she has one modified cytosine and one unmodified cytosine at the same locus, he/she would be detected with a moderate level of cytosine modifications, and will have a moderate β-value. If he/she has two unmodified cytosine at the same locus on both homologous chromosomes, he/she would be detected with a low level of cytosine modifications by methylation chip, and will have a lower β-value. Then we put the DNA methylation level data and SNP genotype data into a unified framework, and define some related concepts as follows:

1. **SMP**: is defined as single methylation polymorphism, or single cytosine modification polymorphism. A SMP locus is a specific chromosome location where cytosine base can be methylated.
2. **SMP allele**: is defined as the DNA methylation modification status of one member of homologous chromosomes at a specific chromosome cytosine location. There are two alleles for a SMP locus: methylation (M allele) and unmethylation (U allele) (Figure 1A).
3. **SMP allele frequency**: is defined as the frequency of methylation allele M and unmethylation allele U. For n samples, the frequency of methylation allele M p_M can be calculated as:

$$p_M = \frac{\text{number of methylation allele M}}{\text{number of samples } (n) \times 2}$$

The frequency of unmethylation allele U p_U can be calculated as:

$$p_U = \frac{\text{number of unmethylation allele U}}{\text{number of samples } (n) \times 2} = 1 - p_M$$

where $p_M + p_U = 1$.
4. **rmSMP**: is defined as a SMP locus with rare M allele, that is, the frequency of M allele is <1% ($p_M < 0.01$).
5. **ruSMP**: is defined as a SMP locus with rare U allele, that is, the frequency of U allele is <1% ($p_U < 0.01$). $p_U < 0.01$ is equivalent to $p_M > 0.99$. In other words, for a ruSMP locus, the frequency of M allele is >99%.
6. **Common SMP**: is defined as a SMP locus with common M allele, that is, the frequency of M allele is from 1% to 99%. For a common SMP, the frequency of U allele is also from 1% to 99%.
7. **Menotype (methylation genotype):** At a specific chromosome cytosine location, the menotype is defined as the combination of SMP alleles located on homologous chromosomes. In other words, for an individual, the menotype at a cytosine locus is the DNA methylation genotype at the locus. There are three possible menotypes at a cytosine locus: methylation homozygote, methylation heterozygote and unmethylation homozygote (Figure 1B).
8. **Methylation homozygote:** For an individual, if he/she has two modified cytosine at the same locus on both members of homologous chromosomes, he/she has a homozygote menotype, denoted MM.
9. **Methylation heterozygote:** For an individual, if he/she has one modified cytosine and one unmodified cytosine at the

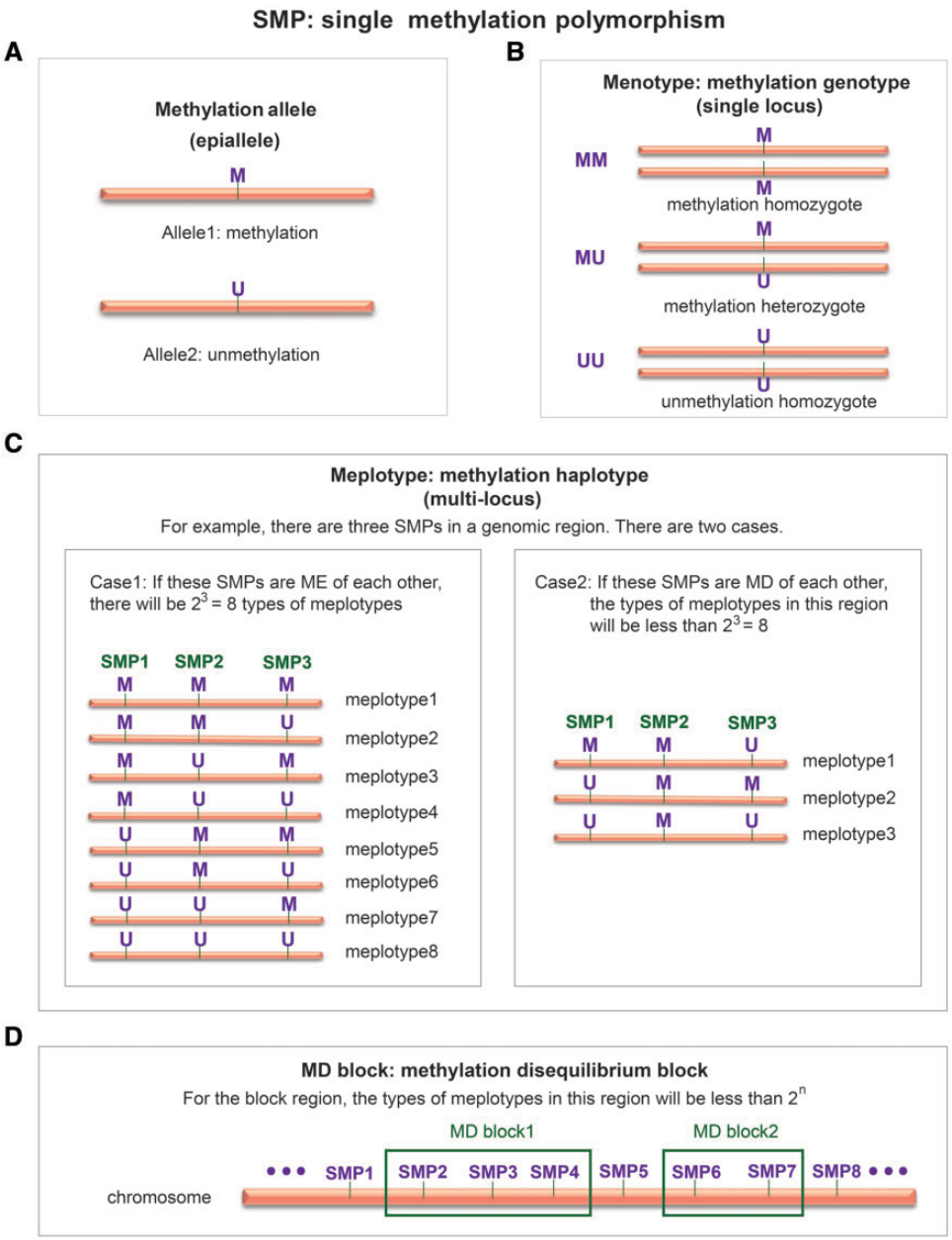

Figure 1. A schematic diagram of (**A**) SMP allele; (**B**) menotype; (**C**) meplotype; and (D) MD block.

same locus, he/she has a heterozygote menotype, denoted MU.

10. **Unmethylation homozygote:** For an individual, if he/she has two unmodified cytosine at the same locus on both members of homologous chromosomes, he/she has a homozygote menotype, denoted UU.
11. **Menotype frequency**: is defined as the frequency of three menotypes: MM, MU and UU. For n samples, the frequency of menotype MM p_{MM} can be calculated as:

$$p_{MM} = \frac{\text{number of menotype MM}}{\text{number of samples } (n)}$$

The frequency of menotype MU f_{MU} can be calculated as:

$$p_{MU} = \frac{\text{number of menotype MU}}{\text{number of samples } (n)}$$

The frequency of menotype UU f_{UU} can be calculated as:

$$p_{UU} = \frac{\text{number of menotype UU}}{\text{number of samples } (n)} = 1 - p_{MM} - p_{MU}$$

where $p_{MM} + p_{MU} + p_{UU} = 1$.

12. **Convert methylation β-value into menotype:**We convert the Methylation β-value into menotype as follows:

$$\text{individual menotype} = \begin{cases} \text{MM}, & \text{if } 0.7 < \beta \leq 1 \\ \text{MU}, & \text{if } 0.3 < \beta \leq 0.7 \\ \text{UU}, & \text{if } 0 \leq \beta \leq 0.3 \end{cases}$$

13. **SMP allele association**: We suppose that a SMP locus has two alleles M and U. There may be some association between the two alleles M and U. We used the SMP Hardy–Weinberg equilibrium (SMP-HWE) to analyze the association between two SMP alleles in the same locus. SMP-HWE was defined as: M and U are independent, that is, no association exists between two alleles of the SMP. Based on the principle of independence, SMP-HWE can be described using formula:

$$\begin{cases} p_{MM} = p_M p_M \\ p_{MU} = 2p_M p_U \\ p_{UU} = p_U p_U \end{cases}$$

where p_{MM} is the frequency of menotype MM, p_{MU} is the frequency of menotype MU, p_{UU} is the frequency of menotype UU, p_M is the frequency of SMP allele M, p_U is the frequency of SMP allele U. In this study, we use Wigginton *et al.*'s [25] method to test the SMP-HWE. The significance level is 0.001. For a SMP locus, if the *P*-value of SMP-HWE <0.001, we believe that the SMP deviated from SMP-HWE, that is, one allele of the SMP is associated with the other allele. The SMP deviated from equilibrium is also called Hardy–Weinberg disequilibrium (SMP-HWD). Based on the relationship between two alleles, the SMP loci with allele association (or SMP-HWD) can be classified into two groups: synSMP and excSMP.

14. **synSMP**: For a SMP locus, if the methylation status of two members of homologous chromosomes has synergic relationship, we define the SMP as synSMP. In other words, the two members of homologous chromosomes tend to be methylated simultaneously ($p_{MM} > p_M p_M$).
15. **excSMP**: For a SMP locus, if the methylation status of two members of homologous chromosomes has exclusion relationship, we define the SMP as excSMP. In other words, if one member of homologous chromosomes is methylated, the other member of homologous chromosomes tends to be unmethylated ($p_{MM} < p_M p_M$).

The DNA MD

For two loci, there may be some correlations. For the SNP data, if there is non-random association of alleles between two SNP loci in a population, we call LD. LD has been widely used in complex disease and population analysis [26–29]. Here, we extend this concept to DNA methylation analysis, and define some related concepts as follows:

1. **Meplotype (methylation haplotype):** is defined as a collection of specific SMP alleles (M or U) on a chromosome (or a haploid) (Figure 1C).
2. **Meplotype frequency:** is defined as the frequency of meplotypes. In this study, we used the Maximum Likelihood Estimate method, which was described by Excoffier *et al.* [30], to estimate the meplotype frequencies in CEU population and YRI population.
3. **Methylation equilibrium:** For two SMP loci SMP1 and SMP2, we suppose that SMP1 has two alleles M1 and U1, and SMP2 has two alleles M2 and U2. The four SMP allele frequencies are denoted as p_{M1}, p_{U1}, p_{M2}, p_{U2}. Methylation equilibrium (ME) is defined as: SMP1 and SMP2 are independent, that is, no association exists between SMP alleles at the two SMP loci. Based on the principle of independence, ME can be described using formula:

$$p_{M1M2} = p_{M1} p_{M2},$$

where p_{M1M2} is the frequency of meplotype M1M2, p_{M1} is the frequency of SMP1 allele M1, p_{M2} is the frequency of SMP2 allele M2.

4. **Methylation disequilibrium:** For two SMP loci SMP1 and SMP2, the MD is defined as non-random association between SMP alleles at the two SMP loci. Therefore, we define the MD coefficient *md* as follows:

$$md = p_{M1M2} - p_{M1} p_{M2}$$

We also provided other four equivalent definitions [31, 32]:

$$\begin{aligned} md &= p_{U1U2} - p_{U1} p_{U2} \\ &= -(p_{M1U2} - p_{M1} p_{U2}) \\ &= -(p_{U1M2} - p_{U1} p_{M2}) \\ &= p_{M1M2} p_{U1U2} - p_{M1U2} p_{U1M2} \end{aligned}$$

5. **MD coefficient mD' and mr^2:** We standardized the *md* using the following two methods.
mD' is defined as [33] follows:

$$mD' = \frac{md}{md_{\max}}$$

where

$$md_{\max} = \begin{cases} \min(p_{M1} p_{U2}, p_{U1} p_{M2}) & md > 0 \\ \max(-p_{M1} p_{M2}, -p_{U1} p_{U2}) & md < 0 \end{cases}$$

The range of mD' is between 0 and 1 [31].
mr^2 was defined as [34] follows:

$$mr^2 = \frac{(md)^2}{p_{M1} p_{U1} p_{M2} p_{U2}}$$

The range of mr^2 is also between 0 and 1.
In this study, the MD coefficient mD' and mr^2 are used to measure the degree of association between two SMP loci.

6. **MD block:** is defined as a chromosome region where SMPs inside were high MD (Figure 1D). In this study, we used the Gabriel *et al.*'s algorithm [35] to identify the MD blocks.
7. **A note about meplotype:** We suppose that there are *n* SMP loci in a genomic region, and each locus has two epialleles M and U. If these SMPs are ME (they are independent of each other), there will be 2^n types of meplotypes. If these SMPs are high MD, the types of meplotypes in this region will be $< 2^n$ (Figure 1C and D).

A case study: comparing the DNA MD maps between HapMap CEU population and YRI population

In this study, we compared the similarities and differences of population epigenetic structure between HapMap CEU population and YRI population from the following aspects: SMP allele frequency, SMP-HWE, MD, MD block and meplotype.

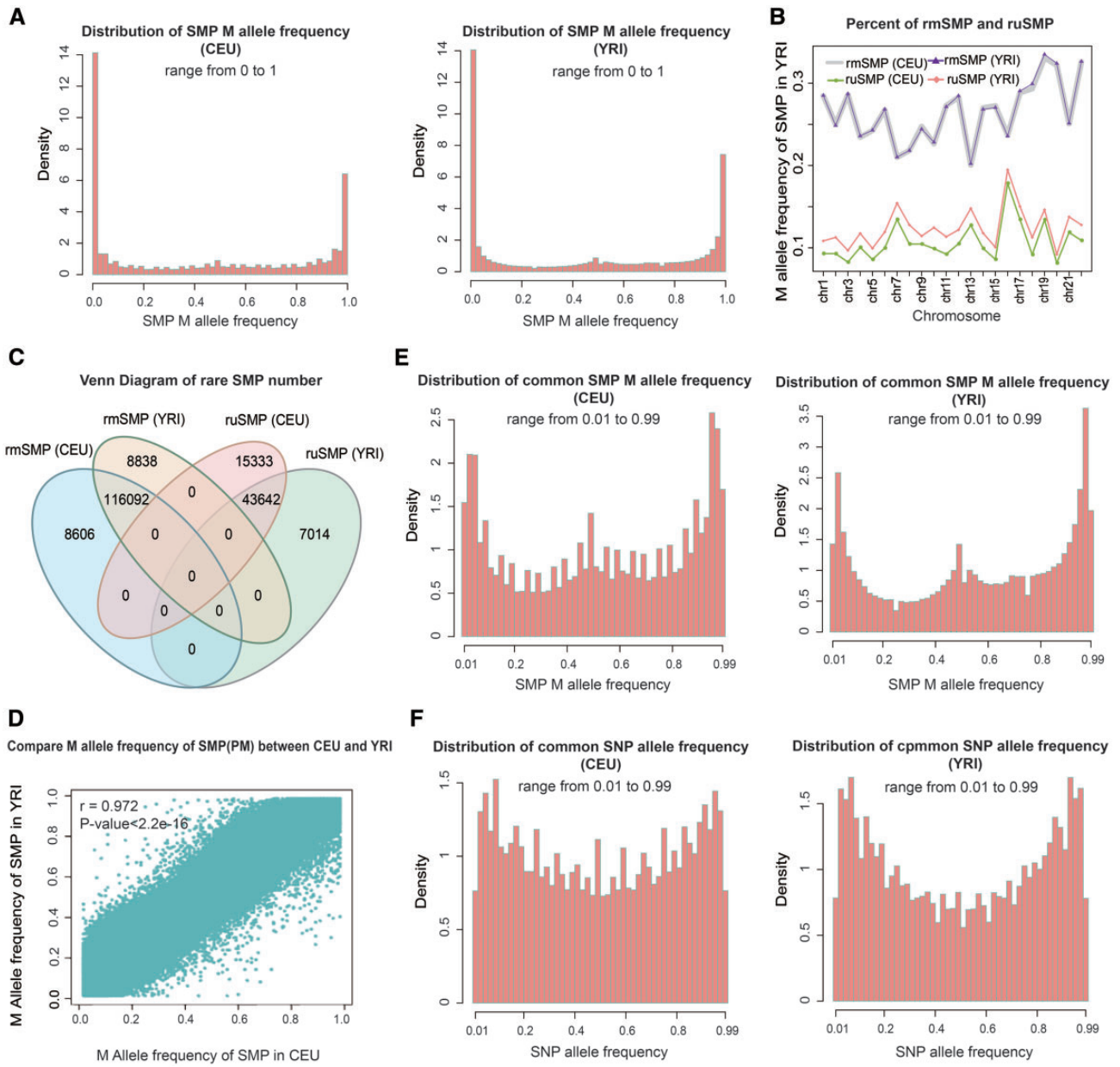

Figure 2. (A) The distribution of SMP M allele frequency for CEU and YRI; (B) compare the percentage of rmSMP/ruSMP in each chromosome; (C) the Venn diagram of shared rmSMP/ruSMP between CEU and YRI; (D) a scatter plot of common SMP frequency in CEU against YRI. (E) the distribution of SMP M allele frequency (removed the rare SMPs) for CEU and YRI; (F) the distribution of common SNP allele frequency for CEU and YRI.

DNA methylation data

In this study, the DNA methylation data of LCLs in two HapMap populations, CEU and YRI, were obtained from NCBI GEO database. We downloaded the 60 CEU and 73 YRI samples from GEO series GSE39672 [20]. The methylation levels of samples were profiled by Illumina 450K array (including 485577 probes). In this study, we only chose to analyze the CpG loci on the autosomal chromosomes. At last, 473 844 loci were included in our study (call ratio > 75% for both CEU and YRI). For each sample, the β-value was used to present the methylation status at a locus.

The SNP genotype data

For each of the 60 CEU and 73 YRI samples, we also extracted their SNP genotype data from the HapMap project [21, 22]. The 3 472 118 autosomal SNPs in both CEU and YRI were selected (call ratio >75%). In other words, each sample in this study has both SNP profile and DNA methylation profile (The coordinate system is GRCh37).

Compare the SMP allele frequency

We first convert the DNA methylation β-value into menotype data. For each CpG locus, we then calculated the frequency of allele M (p_M). The distribution of p_M can be found in Figure 2A. For CEU and YRI populations, we found that p_M had similar distribution. In both of the two populations, there was high proportion of extreme values (rmSMPs with $p_M < 0.01$ or ruSMPs with $p_M > 0.99$). In CEU population, 26.3% (124 698) SMP loci were rmSMPs, and 10.7% (50 656) SMP loci were ruSMPs. In YRI population, 26.4% (124 930) SMP loci were rmSMPs, and 12.4% (58 975)

SMP loci were ruSMPs. We also calculated the percentage of rmSMP/ruSMP on each chromosome (Figure 2B). From Figure 2B, we can see that, in both CEU and YRI, the percentage of rmSMP was higher than the percentage of ruSMP for each chromosome. For rmSMPs, the CEU and YRI population have almost the same percentage for each chromosome. While for ruSMPs, the percentage of ruSMP in YRI population was higher than the percentage of ruSMP in CEU population for all the chromosomes.

We also counted out the number of shared rmSMP/ruSMP between CEU and YRI. The Venn diagram (Figure 2C) showed that, for each of the two types of SMP (rmSMP and ruSMP), CEU and YRI shared a high proportion of SMP number (116 092 rmSMPs and 43 642 ruSMPs). We also noted that the two rmSMP groups (CEU and YRI) did not have any intersection with the two ruSMP groups (CEU and YRI). These suggest that the SMP loci with higher/lower frequency of M allele in one population (CEU) trend to have higher/lower frequency of M allele in another population (YRI). Then we think whether this population epigenetic phenomenon can still be observed in common SMPs. To illustrate this point, we drew a scatter plot of common SMP frequency (p_M) in CEU against YRI. From Figure 2D, we observed a high degree of positive correlation between CEU and YRI. The Pearson's correlation coefficient is 0.972 ($P < 2.2E\text{-}16$). This indicated that CEU and YRI shared similar population epigenetic characteristics.

Nevertheless, there are still some SMP loci that showed significant differences between CEU and YRI. We then used a 4-fold table Chi-square test to test the difference in proportions of M allele and U allele. The result showed that 601 common SMPs were significantly different ($P < 10^{-7}$) between YRI and CEU population. Among these loci, the most significant locus is cg12074150 in chromosome 2 ($P = 5.94E-31$). The p_M is 0.25 in CEU population, while p_M is 0.945 in YRI population. The 601 significant SMPs loci and more detailed gene annotation information can be found at http://www.ewas.org.cn/CEU-YRI.

To better understand the common SMP M allele frequency, we removed the rare SMPs loci and redrew the distribution of p_M for both CEU and YRI (Figure 2E). In addition, to compare with SNPs, we also drew the distribution of common SNP (the minor allele frequency is >1%) for both CEU and YRI (Figure 2F). We observed that the distribution of SMP alleles was different from the distribution of SNP alleles. For SMP allele frequency, CEU and YRI had similar W-type distribution (that is, the shape is similar to the letter 'W'), while for the SNP allele frequency, CEU and YRI had similar U-type distribution (that is, the shape is similar to the letter 'U'). This indicated that the epigenetic marker has different population characteristics with the genetic marker. In subsequent analyses, we use common SMPs to identify MD, MD block and meplotype.

Compare the SMP allele association between CEU and YRI

For each of the two populations (CEU and YRI), we used Wigginton *et al.*'s [25] method to test the SMP allele association for common SMPs. The results showed that most of the SMPs were SMP-HWE for both populations (80.3% SMPs for CEU population and 73.1% SMPs for YRI). Only 19.7% SMPs in CEU population (Figure 3A) and 26.9% SMPs in YRI population (Figure 3B) were significantly deviated from SMP-HWE. In other words, for these SMPs, the DNA methylation status of one member of homologous chromosomes may affect the methylation status of the other members of homologous chromosomes (that is, SMP-HWD). Moreover, we analyzed the percentage of two types of SMPs with allele association: synSMPs and excSMPs (Figure 3A and B). We observed an interesting phenomenon that almost all the SMPs with allele association were excSMPs (99.5% for CEU and 99.8% for YRI). Only 0.5% SMPs with allele association in CEU and 0.2% SMPs with allele association in YRI were synSMPs. This indicated that, for almost all the SMPs with allele association, if one member of homologous chromosomes was methylated, the other member of homologous chromosomes tended to be unmethylated. This population epigenetic phenomenon is consistent in the two HapMap populations CEU and YRI.

We then investigate the sharing of SMPs with allele association between CEU and YRI. Totally, 58 692 SMPs in CEU and 77 867 SMPs in YRI were SMP-HWD (SMP allele association). CEU and YRI shared most of the SMPs with allele association. The Venn diagram (Figure 3C) showed that 49 804 SMPs with allele association were shared between CEU (84.9% = 49 804/58 692) and YRI (64.0% = 49 804/77 867). This indicated that if two SMP alleles were associated in one population (CEU), they also trended to be associated in the other population (YRI). Nevertheless, there are still 8888 SMPs that are CEU specific, and 28 063 SMPs are YRI specific. The more detailed gene annotation information for these loci can be found at http://www.ewas.org.cn/CEU-YRI. For the three subsets of SMPs (interaction, CEU specific and YRI specific), we also investigated the percentage of synSMPs and excSMPs. We found the percentage of synSMPs were different between CEU specific SMP subset and YRI specific SMP subset. The CEU specific SMP subset had highest percentage of synSMPs (2.9%), and was about nine times that of YRI specific SMP subset (0.3%) (Figure 3C).

We also analyzed the number of shared synSMP/excSMP between CEU and YRI. (1) We observed that synSMP/excSMP in CEU did not have any intersection with excSMP/synSMP in YRI (Figure 3D). This suggests that the two types of SMP (synSMP and excSMP) are not compatible even between different populations. (2) We also noted that, between CEU and YRI, the synSMPs had different intersection percentage from excSMP. For synSMPs, the intersection percentage was lower for both CEU and YRI populations. There were 289 synSMPs in CEU and 131 synSMPs in YRI. The number of intersection was only 33, accounting for 11.4% (33/289) of CEU synSMPs, and 25.2% (33/131) of YRI synSMPs (Figure 3D). However, for excSMP, the intersection percentage was higher in both CEU and YRI populations. The number of intersection accounted for 85.2% (49 771/58 403) of CEU excSMPs, and 64% (49 771/77 736) of YRI excSMPs (Figure 3D). This suggested that synSMP has high population specificity; however, excSMP trends to be common between populations.

Although there were some SMP loci with allele association, as we described above, most of the SMPs were SMP-HWE (M and U are independent). In the following analysis, we will use the SMPs with SMP-HWE to compare the MD pattern between CEU and YRI.

Compare the MD between CEU and YRI

To analyze the MD, we first filtered the SMP loci based on the following criteria: (1) SMP-HWE $P \geq 0.001$ in an individual population (CEU or YRI); (2) the minor SMP allele frequency is >1% (common SMPs). After filtering, there were 239 798 SMPs in CEU population and 212 072 SMPs in YRI population.

We then analyzed the decay of MD with distance. For all the 22 chromosomes, we split the SMPs into a number of bins. Each bin includes 1000 SMPs. For all the pair-wise SMPs in a bin, we calculated the MD coefficient mr^2 and the physical distance.

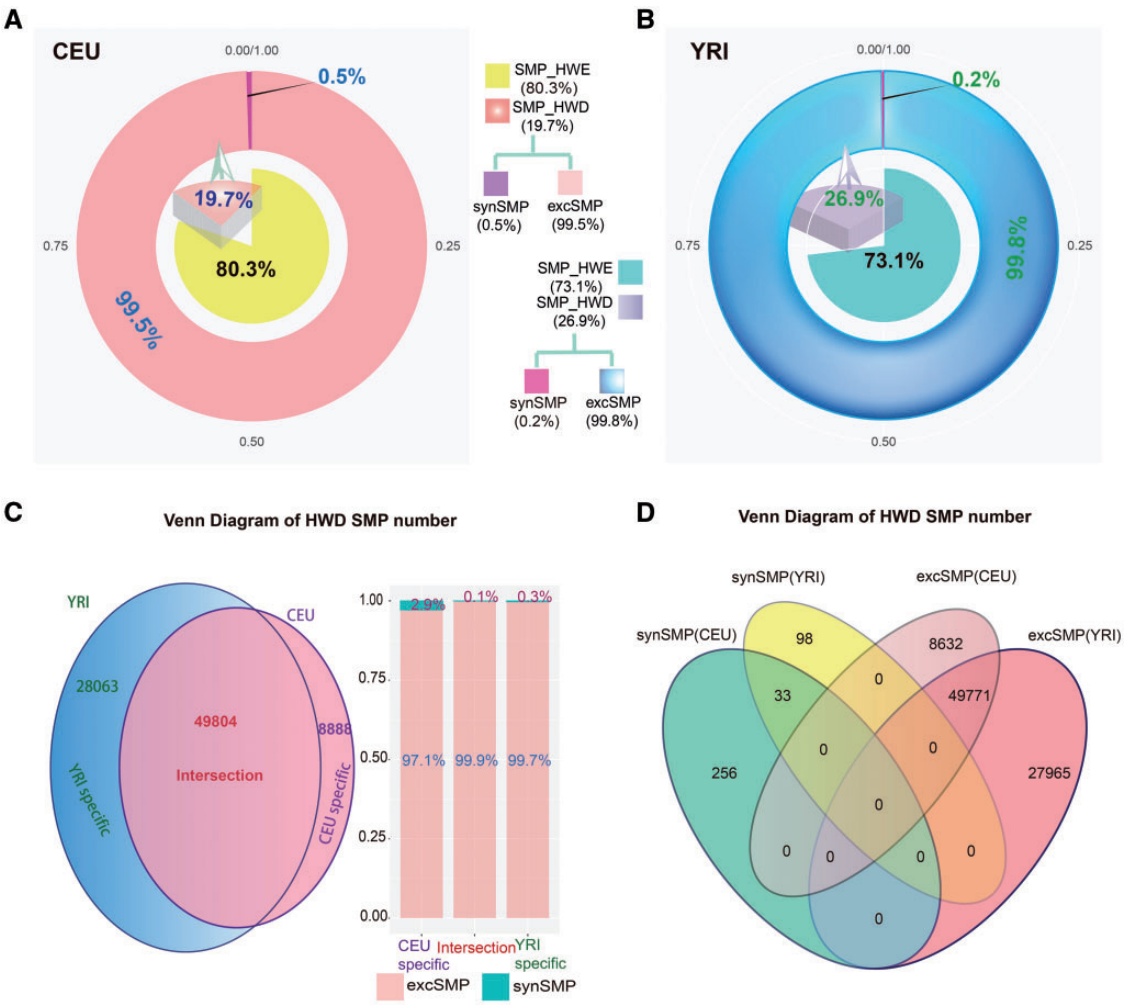

Figure 3. (**A**) The percentage of SMP with SMP-HWE and SMP-HWD in CEU population; (**B**) the percentage of SMP with SMP-HWE and SMP-HWD in YRI population; (**C**) the Venn diagram of SMP with SMP-HWD between CEU and YRI; (**D**) the Venn diagram of synSMP/excSMP between CEU and YRI.

Figure 4A showed the relationship between pair-wise mr^2 and the physical distance of the pair-wise SMPs. The points in a box plot at the *nth* coordinate represent the average mr^2 values of pair-wise SMPs with distance between the $(n-1)th$ coordinate and the *nth* coordinate (each box contains 22 points). From Figure 4A, we can see that, for both CEU and YRI population, the MD coefficient mr^2 decayed with distance and had similar tendency. To compare the SMP with SNP, we also split the SNP data (meet HWE $P \geq 0.001$ and the minor allele frequency$\geq$1%) into a number of bins (length 1000 SNPs), and plot the relationship between pair-wise LD coefficient r^2 and the physical distance (Figure 4A). We observed that the decay rate of SMP is faster than that of SNP (the two curves of SNP were higher than that of SMP). For the SNP, the decay distance of LD coefficient r^2 is about 500 kb, while for the SMP, the decay distance of MD coefficient mr^2 is only about 1 kb. In Figure 4A, we also noted that the two curves of CEU (CEU_SNP and CEU_SMP) were above the two curves of YRI (YRI_SNP and YRI_SMP). This indicated that the degree of LD and MD in CEU were higher than that of YRI. In addition, we observed that the distance between two curves of SMP was less than the distance between two SNP curves. This suggested that the difference of MD degree between CEU and YRI was less than the difference of LD degree.

Further, we investigated the correlation between SMP MD and SNP LD. For each of the 22 chromosomes, to accurately analyze the correlation, we first split the CEU SMPs into a number of little bins (each bin include 3 SMPs). Then we mapped the CEU SNP loci to bins based on the physical position. For each bin, we calculated the average mr^2 of all pair-wise SMPs and the average r^2 of all pair-wise SNPs. At last, for each chromosome, we calculated the Pearson's correlation coefficient between average mr^2 and average r^2. We observed the low correlation between SMP MD and SNP LD. Most of the Pearson's correlation coefficient of 22 chromosomes were <0.01. For YRI population, we also repeated the same analysis, and observed a similar phenomenon (Figure 4B). This suggested, whether in CEU population or in YRI population, the SMP MD was less affected by SNP LD.

For SMP MD and SNP LD, we also separately calculated the Pearson's correlation coefficient between CEU and YRI. We found that, for both SMP and SNP, CEU and YRI have a high correlation.

All of the above analyses were repeated by using bin sizes 5 and 7. From Figure 4B, we can see that the results of bin sizes 5 and 7 are consistent with bin size 3. These results indicated that our results are highly stable and reliable.

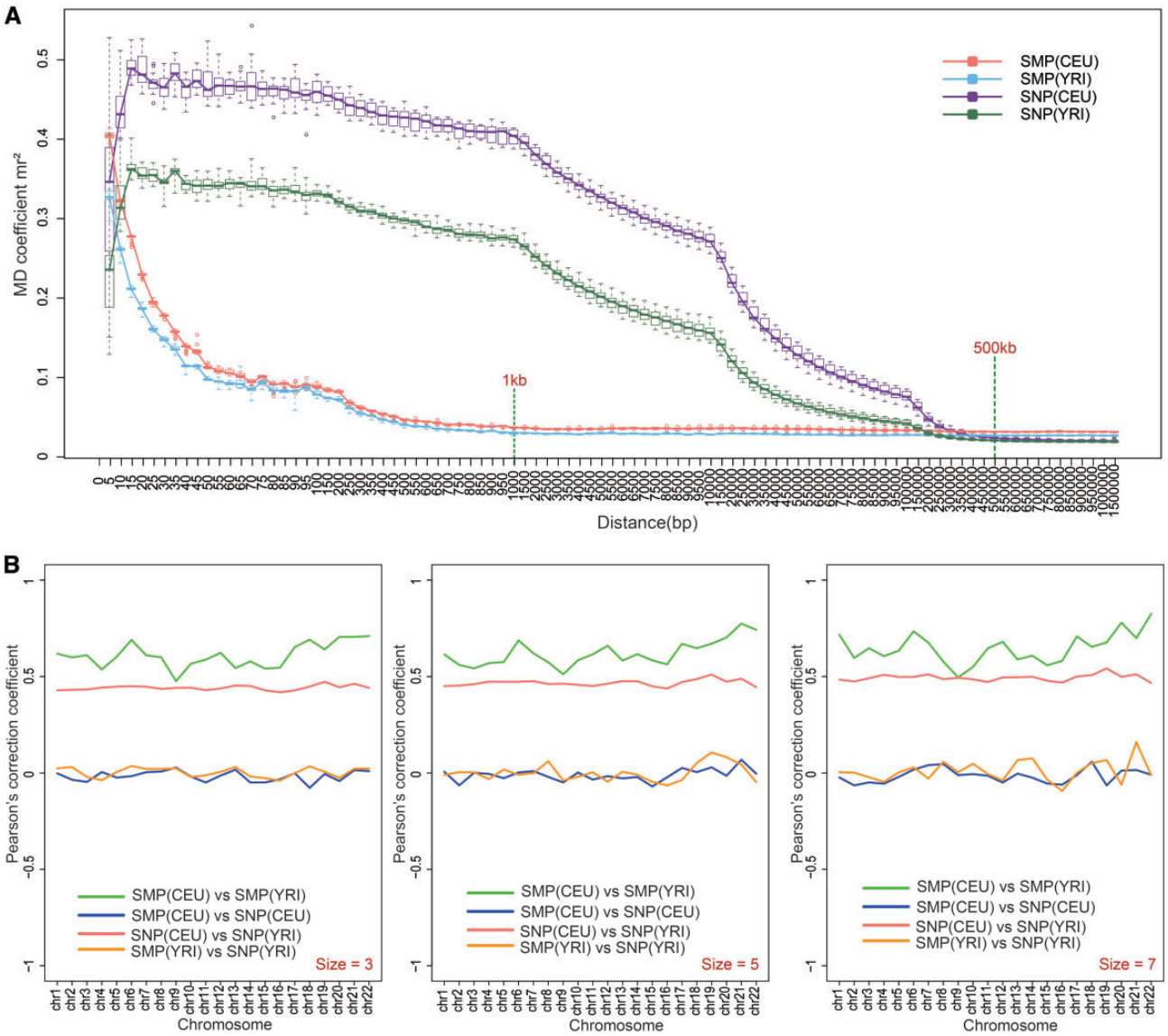

Figure 4. (**A**) The decay of MD and LD with distance; (**B**) the correlation between MD and LD.

Compare the MD blocks between CEU and YRI

In this study, we used the Gabriel *et al.*'s algorithm [35] to identify the MD blocks. Table 1 showed the number of MD blocks. For SMP, we identified 571 MD blocks in CEU population and 489 MD blocks in YRI population. Chromosome 6 has the most number of MD blocks (89 MD blocks in CEU and 64 MD blocks in YRI), and chromosome 9 has the fewest number of blocks (only 1 MD blocks in CEU and 2 MD blocks in YRI). To compare with LD blocks, we also identified the LD blocks using Gabriel *et al.*'s algorithm. We see that, the number of LD blocks is far more than that of MD blocks (Table 1). There were 119 719 LD blocks in CEU and 182 481 LD blocks in YRI. We also calculated the average length of MD blocks and LD blocks. We found that the average length of LD blocks (16 931 bp in CEU and 8152 bp in YRI) was also far more than that of MD blocks (298 bp in CEU and 722 bp in YRI). In summary, SMP had lower block number and shorter block size than SNP. This phenomenon is partly because the decay rate of SMP MD is faster than that of SMP LD (see Figure 4A).

Shared MD blocks between CEU and YRI

We then analyzed the overlap of blocks between CEU and YRI. For SMP blocks, 199 CEU MD blocks (34.9% = 199/571) shared some locations with YRI MD blocks, and 197 YRI MD blocks (41.5% = 203/489) shared some locations with CEU MD blocks. To compare with random, we fixed the length of MD blocks and shuffled MD blocks across the chromosome. We found that there was almost no overlap between CEU blocks and YRI blocks in the case of random. This indicated that the distribution of MD block is not random, CEU and YRI shared some specific population epigenetic structure. We also counted out the number of completely overlap blocks. There were 88 MD blocks that shared the same starting and ending positions (Supplementary Table S1). The block that contains the largest number of SMPs located in HOXA5 (27 180 671–27 192 309 bp) gene body region on chromosome 7. It ranges from 27 182 637 to 27 183 861 bp (length 1.224 kb), and contains 17 SMP loci: cg05076221, cg23936031, cg02248486, cg25866143, cg09549073, cg01370449, cg04863892, cg19759481, cg12128839, cg02916332, cg17569124,

Table 1. The number of MD blocks and LD blocks in CEU and YRI

Chr	Methylation disequilibrium (MD)						Linkage disequilibrium (LD)					
	MD_blocks number (CEU)	MD_blocks average length (bp, CEU)	MD_blocks number (YRI)	MD_blocks average length (bp, YRI)	CEU overlap with YRI	YRI overlap with CEU	LD_blocks number (CEU)	LD_blocks average length (bp, CEU)	LD_blocks number (YRI)	LD_blocks average ength (bp, YRI)	CEU overlap with YRI	YRI overlap with CEU
chr1	40	1809	39	505	15	15	9277	18 648	14 074	8999	7876	13 204
chr2	31	124	39	918	14	14	9740	19 953	15 800	8932	8348	14 857
chr3	25	61	19	1463	8	8	8083	19 190	12 507	9254	6884	11 696
chr4	22	1154	14	988	4	4	7340	20 516	11 608	9446	6317	10 931
chr5	32	768	20	917	8	7	7366	19 819	11 676	9118	6291	11 037
chr6	89	73	64	392	32	34	7875	17 491	11 855	8463	6776	11 129
chr7	34	194	31	757	10	10	6658	18 536	10 140	8478	5683	9482
chr8	24	476	34	1106	12	12	6501	17 879	10 482	8366	5527	9805
chr9	1	12	2	74	0	0	5945	14 299	8824	6989	4964	8228
chr10	34	145	27	1281	9	9	6295	16 039	9556	7691	5361	8930
chr11	37	80	28	136	15	15	5854	17 914	9001	8707	4934	8435
chr12	27	165	30	1585	10	10	5904	17 774	8680	8780	4976	8068
chr13	14	69	15	86	5	5	4680	16 247	7242	7655	4031	7273
chr14	12	97	12	2696	4	5	4096	16 379	6102	7931	3446	6050
chr15	17	64	7	168	2	2	3737	15 884	5613	7256	3082	5559
chr16	13	171	9	215	4	4	3903	16 635	5696	8671	3121	5536
chr17	42	274	37	448	21	21	3336	17 927	4642	9047	2666	4551
chr18	8	149	5	78	3	3	3667	15 921	5702	7303	3059	5633
chr19	28	141	25	134	7	8	2503	14 972	3222	9276	2011	3074
chr20	18	72	13	141	8	9	3186	14 328	4753	6720	2620	4668
chr21	9	114	8	306	2	2	1842	13 765	2712	6338	1519	2712
chr22	14	347	11	1482	6	6	1931	12 362	2594	5926	1545	2603
Total	571	298	489	722	199	203	119 719	16 931	182 481	8152	101 037	173 461

cg02005600, cg25307665, cg14014955, cg20517050, cg23204968, cg05835726. There were 10 meplotypes in this block, and these meplotypes had similar frequencies for CEU and YRI (for more details, see Supplementary Table S1).

The population difference of meplotypes between CEU and YRI

For each meplotype in the 88 shared MD blocks (total 418 meplotypes), we carried out a chi-square test to analyze the statistical significance of population difference between CEU and YRI. The statistical significance level is 0.0001 (Bonferroni correction $\alpha = 0.05/418$). We found two meplotypes showing significant difference between CEU and YRI. The most significant meplotype is MMM (cg24891660, cg18000391 and cg04757492) located in PLEC gene body (144 989 321–145 050 913 bp) on chromosome 8 from 145 003 653 to 145 003 862 bp (length 209 bp). The frequency of meplotype MMM in YRI population is 0.419, while in CEU is only 0.124. The *P*-value is 1.23E-07. The other significant meplotype is UU located in MTERFD2 (MTERF4) gene body (242 026 509–242 041 747 bp) on chromosome 2. The UU meplotype consists of two SMP loci, cg24269863 and cg21773665, ranging from 242 027 434 to 242 027 464 bp. The frequency of UU meplotype in YRI (0.575) is higher than that of CEU (0.291). The *P*-value is 3.67E-06. In addition to this meplotype UU, there are still two meplotypes MM and MU in the block. For MM meplotype, although it is not significant, it still has a smaller *P*-value 0.0007. The frequency of MM in YRI (0.335) is lower than in CEU (0.541). This indicates that the chromosome region in this block tends to be methylated in CEU population, whereas it tends to not be methylated in YRI population. For more meplotype test results, see Supplementary Table S1.

Discussions

In this study, we analyzed the population epigenetic characteristics using the cytosine modification data. By comparing the SMP allele frequency, SMP allele association, MD, MD block and meplotype frequency between CEU and YRI, we observed some stable population epigenetic phenomena, such as the same distribution of SMP allele frequency, similar percentage of excSMP and shared MD block region. Although Schmitz *et al.* described that 'errors in the maintenance of methylation states may result in the accumulation of SMPs over an evolutionary timescale' [16], at present, it is difficult to explain the formation of more complex population epigenetic phenomenon, such as MD and MD block. In this study, we have only explored and described the population characteristics of DNA methylation. In the future, we will further analyze the reasons for the formation of stable population epigenetic characteristics. Here, we do not discuss more about this.

For both CEU and YRI populations, 19–27% common SMPs were SMP-HWD. This implies that the two SMP alleles of these SMPs were not independent. Thus, when we carry out frequency-based analysis (such as case-control design) and identify the SMPs related to some phenotypes, the association between the two SMP alleles may affect the results. Here, we believed that the SMPs with SMP-HWD should be filtered in these studies. In other words, it is better to use the SMPs with SMP-HWE as biomarker to identify the association with phenotypes or disease.

In this study, we also carried out an analysis of gene differential expression between CEU and YRI. We downloaded the gene expression data GSE9703 [36] from GEO database. There were 57 CEU samples and 54 YRI samples in both DNA methylation data set and SNP data set. For those genes that have significant differences in SMP allele frequency and meplotype frequency, we used *t*-test to analyze their difference of gene expression. We observed that the gene expression of the two genes PLEC and MTERFD2 (carrying the most significant meplotypes, mentioned above) did not show the significant differences between CEU and YRI ($P = 0.0497$ for PLEC and $P = 0.2131$ for MTERFD2). More detailed information about *t*-test can be found at http://www.ewas.org.cn/CEU-YRI.

Here, we did not discuss the regulatory relationship between SNP and methylation. Zhang Lab had carried out many excellent works to describe the interaction between SNP and methylation [20, 23, 37]. Although we do not consider these aspects, it does not mean that they are not important. In the future, we will discuss these factors when we explore the reasons for the formation of population epigenetic properties.

The samples in this study were from the LCLs, which is a pure population of B cells [20]. This avoids the effect of tissue-specific DNA methylation patterns on our results [38] because the primary samples from humans may include many cell types. Both CEU and YRI samples were from HapMap, and were simultaneously detected using Illumina chip. Therefore, there are few batch effects, and will not impact on our results. In addition, owing to technical limitations, we are still difficult to detect the menotype (SMP genotype) for a single cell. In this study, we adopt a discretization strategy to describe the menotype of LCLs. The advantage of this strategy is that we can analyze the combination of specific SMP alleles (M or U) on a single chromosome. The disadvantage is that the process of discretization will lose some information. With the development of technology, the detection of SMP alleles will be more accurate. By using thresholds 0.3 and 0.7, we convert the DNA methylation β-value into menotype (UU, UM and MM). We also tried some other thresholds, such as 0.25/0.75 and 0.35/0.65. We found that, the population epigenetic phenomena found by other thresholds were consistent with that found by threshold 0.3/0.7. This indicated that the population epigenetic properties are real, and the different menotype determination methods will not significantly affect our description of the population genetic phenomenon.

We split all the SMPs in 22 chromosomes into a number of bins (length 1000 SMPs). For pair-wise SMPs in each bin, we calculated pair-wise mr^2 and the physical distance. We found that the MD also exists for human population epigenetic data, and the global decay distance of MD is about 1 kb for both CEU and YRI population. For the decay of MD, Liu *et al.* observed that the methylation correlation was reduced by half in <500 bp, and the global decay distance was also about 1 kb [39]. Moen *et al.* analyzed the co-modification between CpG sites using Spearman's correlation coefficient, and found that 'beyond 1 kb, co-modification between CpG sites decreased to about the background level' [20]. The conclusions of these studies are consistent. This suggests that the markers or genes within 1 kb of a candidate DNA methylation locus should also be focused on. In addition, we should carry out some MD-based analysis using DNA methylation data.

Our research also has implications for the study of complex diseases. We advise that researchers should carry out the MD block identification and meplotype-based association study for their epigenome-wide case-control data. These analyses can help them find some complex disease-related meplotypes, regions, genes or pathways.

In conclusion, we believe that the framework of population genetics can be used to understand the DNA methylation from

population level. And we also hope that these concepts and principles can be widely used in population epigenetic studies, such as epigenome-wide association studies.

Future

In the future, we plan to develop some programs and software based on these population epigenetic concepts and principles. The progress of our software development and related data resources will be found on the epigenome-wide association studies web site: http://www.ewas.org.cn.

Key Points

- We extended the classical framework of population genetics to epigenetics.
- We compared the DNA methylation disequilibrium maps between HapMap CEU and YRI.
- We observed some stable population epigenetic phenomena.

Supplementary Data

Supplementary data are available online at http://bib.oxfordjournals.org/.

Acknowledgments

We thank Zhang Lab for providing such useful data.

Funding

This work was supported in part by grants from the National Natural Science Foundation of China (grant numbers 31200934 and 81300945) and the Natural Science Foundation of Heilongjiang Province (grant number C2016036). The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

References

1. Michelotti GA, Brinkley DM, Morris DP, *et al*. Epigenetic regulation of human alpha1d-adrenergic receptor gene expression: a role for DNA methylation in Sp1-dependent regulation. *Faseb J* 2007;**21**:1979–93.
2. Jones PA, Takai D. The role of DNA methylation in mammalian epigenetics. *Science* 2001;**293**:1068–70.
3. Linn F, Heidmann I, Saedler H, *et al*. Epigenetic changes in the expression of the maize A1 gene in Petunia hybrida: role of numbers of integrated gene copies and state of methylation. *Mol Gen Genet* 1990;**222**:329–36.
4. Shirodkar AV, St Bernard R, Gavryushova A, *et al*. A mechanistic role for DNA methylation in endothelial cell (EC)-enriched gene expression: relationship with DNA replication timing. *Blood* 2013;**121**:3531–40.
5. Horvath S. DNA methylation age of human tissues and cell types. *Genome Biol* 2013;**14**:R115.
6. Jones MJ, Goodman SJ, Kobor MS. DNA methylation and healthy human aging. *Aging Cell* 2015;**14**:924–32.
7. Li E, Beard C, Jaenisch R. Role for DNA methylation in genomic imprinting. *Nature* 1993;**366**:362–5.
8. Reik W. Stability and flexibility of epigenetic gene regulation in mammalian development. *Nature* 2007;**447**:425–32.
9. Heard E, Clerc P, Avner PX. Chromosome inactivation in mammals. *Annu Rev Genet* 1997;**31**:571–610.
10. Li E, Bestor TH, Jaenisch R. Targeted mutation of the DNA methyltransferase gene results in embryonic lethality. *Cell* 1992;**69**:915–26.
11. Okano M, Bell DW, Haber DA, *et al*. DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development. *Cell* 1999;**99**:247–57.
12. Illingworth R, Kerr A, Desousa D, *et al*. A novel CpG island set identifies tissue-specific methylation at developmental gene loci. *PLoS Biol* 2008;**6**:e22.
13. Becker C, Hagmann J, Muller J, *et al*. Spontaneous epigenetic variation in the *Arabidopsis thaliana* methylome. *Nature* 2011;**480**:245–9.
14. Schmitz RJ, Schultz MD, Lewsey MG, *et al*. Transgenerational epigenetic instability is a source of novel methylation variants. *Science* 2011;**334**:369–73.
15. Genereux DP, Miner BE, Bergstrom CT, *et al*. A population-epigenetic model to infer site-specific methylation rates from double-stranded DNA methylation patterns. *Proc Natl Acad Sci USA* 2005;**102**:5802–7.
16. Schmitz RJ, Schultz MD, Urich MA, *et al*. Patterns of population epigenomic diversity. *Nature* 2013;**495**:193–8.
17. Wenzel MA, Piertney SB. Fine-scale population epigenetic structure in relation to gastrointestinal parasite load in red grouse (Lagopus lagopus scotica). *Mol Ecol* 2014;**23**:4256–73.
18. Richards EJ. Population epigenetics. *Curr Opin Genet Dev* 2008;**18**:221–6.
19. Shoemaker R, Deng J, Wang W, *et al*. Allele-specific methylation is prevalent and is contributed by CpG-SNPs in the human genome. *Genome Res* 2010;**20**:883–9.
20. Moen EL, Zhang X, Mu W, *et al*. Genome-wide variation of cytosine modifications between European and African populations and the implications for complex traits. *Genetics* 2013;**194**:987–96.
21. International HapMap C, Frazer KA, Ballinger DG, *et al*. A second generation human haplotype map of over 3.1 million SNPs. *Nature* 2007;**449**:851–61.
22. International HapMap C, Altshuler DM, Gibbs RA, *et al*. Integrating common and rare genetic variation in diverse human populations. *Nature* 2010;**467**:52–8.
23. Zhang W, Gamazon ER, Zhang X, *et al*. SCAN database: facilitating integrative analyses of cytosine modification and expression QTL. *Database* 2015;**2015**:bav025.
24. Du P, Zhang X, Huang CC, *et al*. Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. *BMC Bioinformatics* 2010;**11**:587.
25. Wigginton JE, Cutler DJ, Abecasis GR. A note on exact tests of Hardy-Weinberg equilibrium. *Am J Hum Genet* 2005;**76**:887–93.
26. Balding DJ. A tutorial on statistical methods for population association studies. *Nat Rev Genet* 2006;**7**:781–91.
27. Rosenberg NA, Huang L, Jewett EM, *et al*. Genome-wide association studies in diverse populations. *Nat Rev Genet* 2010;**11**:356–66.
28. Hirschhorn JN, Gajdos ZK. Genome-wide association studies: results from the first few years and potential implications for clinical medicine. *Annu Rev Med* 2011;**62**:11–24.
29. Stranger BE, Stahl EA, Raj T. Progress and promise of genome-wide association studies for human complex trait genetics. *Genetics* 2011;**187**:367–83.

30. Excoffier L, Slatkin M. Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. *Mol Biol Evol* 1995;**12**:921–7.
31. Mueller JC. Linkage disequilibrium for different scales and applications. *Brief Bioinform* 2004;**5**:355–64.
32. Hill WG, Robertson A. Linkage disequilibrium in finite populations. *Theor Appl Genet* 1968;**38**:266–31.
33. Lewontin RC. The interaction of selection and linkage. I. General considerations; Heterotic models. *Genetics* 1964;**49**:49–67.
34. Hill WG, Robertson A. Linkage disequilibrium in finite populations. *Theor Appl Genet* 1968;**38**:226–31.
35. Gabriel SB, Schaffner SF, Nguyen H, *et al*. The structure of haplotype blocks in the human genome. *Science* 2002;**296**:2225–9.
36. Zhang W, Duan S, Bleibel WK, *et al*. Identification of common genetic variants that account for transcript ISOFORM variation between human populations. *Hum Genet* 2009;**125**:81–93.
37. Zhang X, Moen EL, Liu C, *et al*. Linking the genetic architecture of cytosine modifications with human complex traits. *Hum Mol Genet* 2014;**23**:5893–905.
38. Rakyan VK, Down TA, Thorne NP, *et al*. An integrated resource for genome-wide identification and analysis of human tissue-specific differentially methylated regions (tDMRs). *Genome Res* 2008;**18**:1518–29.
39. Liu Y, Li X, Aryee MJ, *et al*. GeMes, clusters of DNA methylation under genetic control, can inform genetic and epigenetic analysis of disease. *Am J Hum Genet* 2014;**94**:485–95.