Epigenetics Communications

CORRESPONDENCE

Open Access

Characterization of the Illumina EPIC array for optimal applications in epigenetic research targeting diverse human populations

Zhou Zhang[1], Chang Zeng[1] and Wei Zhang[1,2*]

Abstract

The Illumina EPIC array is widely used for high-throughput profiling of DNA cytosine modifications in human samples, covering more than 850,000 modification sites across various genomic features. The application of this platform is expected to provide novel insights into the epigenetic contribution to human complex traits and diseases. Considering the diverse inter-population genetic and epigenetic variation, it will benefit the research community with a comprehensive characterization of this platform for its applicability to major global populations. Specifically, we mapped 866,836 CpG probes from the EPIC array to the human genome reference. We detected 91,034 CpG probes that did not align reliably to the human genome reference. In addition, 21,256 CpG probes were found to ambiguously map to multiple loci in the human genome, and 448 probes showing inaccurate genomic information from the original Illumina annotations. We further characterized those uniquely mapped CpG probes in terms of whether they contained common genetic variants, i.e., single nucleotide polymorphisms (SNPs), in major global populations, by utilizing the 1000 Genomes Project data. A list of optimal CpG probes on the EPIC array was generated for major global populations, with the aim of providing a resource to facilitate future studies of diverse human populations. In conclusion, our analysis indicated that studies of diverse human populations using the EPIC array would be benefited by taking into account of the technical features of this platform.

Keywords: Epigenetics, DNA methylation, Microarray, Single nucleotide polymorphism, Population epigenetics

Introduction

DNA methylation (i.e., cytosine modification) has been implicated in various human complex traits and diseases [1]. The Illumina Infinium HumanMethylation assay, such as the Infinium 27K and 450K arrays, offers a cost-efficient, high-throughput platform of genome-wide CpG profiling for investigating complex traits and diseases [2–4], including for example detecting epigenetic variation between human populations, dissecting the genetic architecture of cytosine modifications, and detecting epigenetic contributors to human diseases [5]. However, it has become known that a set of the Infinium CpG probes did not perform as designed due to technical biases, such as probes ambiguously mapped to the human genome, probes containing common genetic variants in the form of single nucleotide polymorphisms (SNPs), and probes containing mismatched nucleotides [6–12].

Of particular interest to us is the potential technical bias caused by inter-population genetic variation, given that probes containing common genetic variants for a particular population may not generate reliable profiling data in that population [10]. Therefore, a thorough analysis of the CpG probes for these technical features will help determine which CpG probes may perform

*Correspondence: wei.zhang1@northwestern.edu

[1] Department of Preventive Medicine, Northwestern University Feinberg School of Medicine, 680 N. Lake Shore Dr., Suite 1400, Chicago, IL 60611, USA
Full list of author information is available at the end of the article

more reliably in a given population. We herein described a comprehensive analysis of the Illumina EPIC array, which is the current and widely used Infinium platform designed to interrogate more than 850,000 modification sites across the human genome, including for example sites within and outside of CpG islands, ENCODE (The Encyclopedia of DNA Elements) open chromatin, DNase hypersensitive sites, and > 90% of CpGs from the 450K array [10, 13]. Specifically, we sought to provide a technical characterization of the EPIC array with the aim of facilitating epigenomic research of complex traits and diseases targeting different populations.

Materials and methods

Retrieval of CpG probe sequences and annotations

Probe sequences and genomic annotations of the EPIC array were obtained at the support website maintained by Illumina, Inc. (http://www.illumina.com/). In total, sequences of 866,836 probes (50 bp/probe) were retrieved. Illumina provides basic information such as chromosomal location, genomic position (hg19), and strand. The original genic assignment was checked for accuracy by aligning the probe sequences to the human genome reference (hg19) as described below. The EPIC array employs two types of probes: Type I (2 probes per CpG locus, representing 1 "unmethylated" and 1 "methylated" query sequence) and Type II (one probe per CpG locus).

Obtaining genetic variant information from global populations

We downloaded variant calls in VCF format for 26 global populations from the 1000 Genomes Project [14–17]. Given that SNPs represent the majority of common genetic variants, we focused on SNPs in this work [14]. We calculated allele frequency for each bi-allelic SNP in each population. In addition to individual populations, we also obtained common SNPs for each of the five major global groups: African (7 populations), European (5 populations), East Asian (5 populations), South Asian (5 populations), and ad Admixed American (4 populations).

Detection of cross-hybridized probes, i.e., ambiguously mapped to the human genome

We aligned 866,836 probe sequences to hg19 following the method presented in a previous study [18]. Briefly, to represent all possible post-bisulfite conversion sequences, four single-stranded genomes (i.e., forward methylated, forward unmethylated, reverse methylated, and reverse unmethylated) were generated *in silico*. Both forward and reverse strands of hg19 were bisulfite converted *in silico*, in which all cytosines (C) of non-CpG dinucleotides were converted to thymines (T). For the unmethylated genome, all Cs of CpG sites were also converted to T's, whereas in the methylated genome, all Cs of CpG sites remain as Cs (Fig. 1a). Type I probes were extracted from the annotated file provided by the

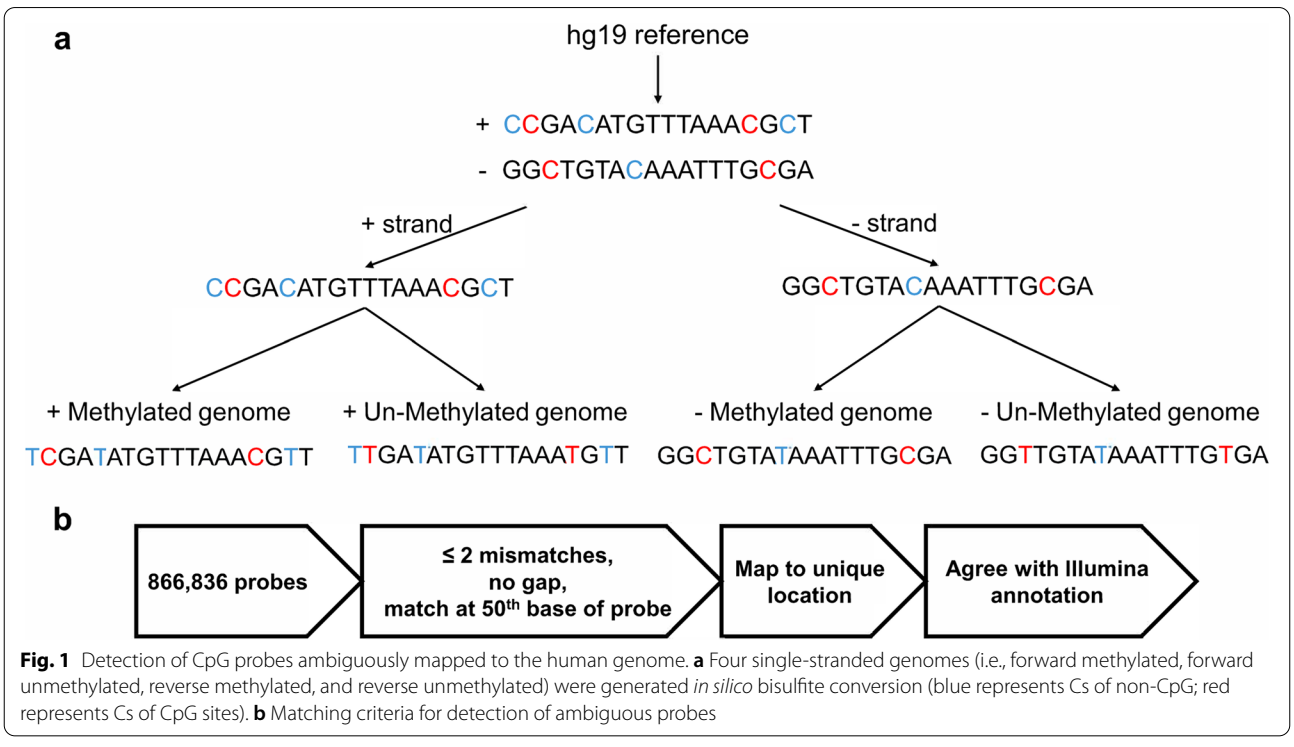

Fig. 1 Detection of CpG probes ambiguously mapped to the human genome. **a** Four single-stranded genomes (i.e., forward methylated, forward unmethylated, reverse methylated, and reverse unmethylated) were generated *in silico* bisulfite conversion (blue represents Cs of non-CpG; red represents Cs of CpG sites). **b** Matching criteria for detection of ambiguous probes

Illumina. For type II probes, according to Illumina, some of them contain R nucleotides, representing either A or G due to the presence of CpG sites within the probe sequence. In order to represent all possible probe sequences, we replaced all R nucleotides in these probes with all possible combinations of A and G. In the end, we generated 3,505,864 probe sequences for all array probes. All these probe sequences were then aligned to the *in silico* converted reference using BLAT [19] with default parameters.

The matching criteria were shown in Fig. 1b. Briefly, matches with $\leq$ 2 mismatches were retained. Only matches with a match at the 50th nucleotide of probe sequences were retained. Duplicate matches of the same probe that map to the same chromosomal location were also removed. Matches with gaps were also removed. Probes that satisfied the above criteria, while having multiple matches, were defined as ambiguous probes. The chromosomal locations of unambiguous probes were compared with Illumina's information.

Enrichment analysis with cis-regulatory elements

We used DNase I hypersensitivity sites (DNase), transcription factor binding sites (TFBS), and annotations of histone modification peaks pooled across cell lines, downloaded at the ENCODE Analysis Hub at the European Bioinformatics Institute. For each regulatory element, we calculated the number of overlapping regions with the ambiguous probes (observed). To generate the null (expected) distribution of the number of overlaps between the regulatory elements and array probes, a random set of probes (same number as the ambiguous probes) from the array was randomly selected 10,000 times and overlapped with each regulatory element. The expected mean number of overlaps was derived from the distribution of 10,000 random sets. We then calculated the ratio of observed to mean expected as the enrichment fold and obtained an empirical p-value from the distribution of expected.

Detection of SNP-containing CpG probes by population

We examined the 754,546 unambiguous EPIC array probes for SNPs in each of the 26 populations from the 1000 Genomes Project [15]. We searched for common SNPs, i.e., MAF (minor allele frequency) >0.05 within 20 bp of the interrogated CpG sites in each population. The common SNP-containing probes were summarized by the location of the SNP within a probe. To provide an

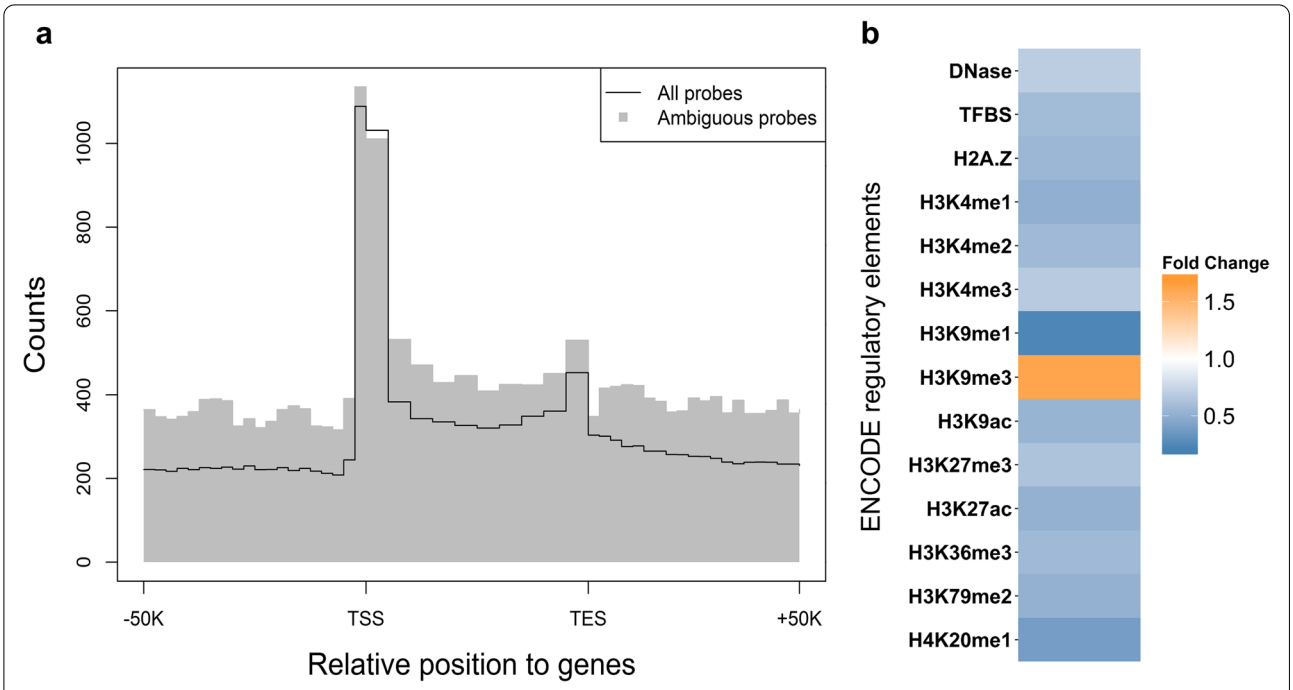

Fig. 2 Summary of the EPIC probes with Illumina annotations. **a** Ambiguous probes generally follow the distribution of all EPIC probes in terms of relative position to the gene. The black line represents all EPIC probes. The grey bar represents ambiguous probes. **b** Enrichment analysis indicates that ambiguous probes are more likely to co-localize with H3K9me3. DNase, DNase I hypersensitivity sites; TFBS, transcription factor binding sites; H2A.Z, histone H2A variant; H3, histone H3; H4, histone H4; K, lysine; me1, monomethylation; me2, demethylation; me3, trimethylation; ac, acetylation

overview of major global populations, we also examined common SNP-containing probes in the five major global groups.

Results

Evaluation of the cross-hybridized probes in the EPIC array

Out of the total number of 866,836 probes on the EPIC array, in total 91,034 probes did not meet the criteria for ≤ 4 ambiguous bases (i.e., N) and ≤ 2 mismatches. We detected 21,256 probes that were not aligned to unique loci in the human genome reference. Figure 2 shows the summary of the ambiguous EPIC probes in terms of genomic distribution and enrichment of the ENCODE *cis*-regulatory elements according to Illumina's annotated chromosomal locations. The ambiguous probes generally followed the distribution of all EPIC probes, while we observed a trend of enrichment with H3K9me3 (empirical $p < 0.001$). We further compared the mapping results with the original genomic annotations provided by Illumina. In total, 448 probes were mapped to a different gene from the original Illumina annotation file (Supplemental Table 1).

Common SNP-containing probes by population

Table 1 shows the summary of common SNPs (MAF > 0.05) detected by population for those unambiguous probes on the EPIC array. Notably, for the East Asian populations there existed the least common SNP-containing probes, 54,810 at MAF >0.05, followed by European (68,155), South Asian (70,068), Admixed American (70,576), and African populations (96,674). SNP-containing probes by population are provided in Supplemental Table 2.

Conclusion

Epigenetic research of complex traits and diseases including epigenome-wide association studies (EWAS) rely on the robustness of profiling measurements from high through-put platforms. The EPIC array provides a powerful tool that covers almost 1 million modification sites across the human genome, as well as cost efficiency for population-based studies. However, there have been many instances of technical artifacts that can lead to erroneous results [10, 13, 18, 20, 21].

Table 1 Summary of common SNPs within +/− 20 bp of CpG probes by population

Population	Description	Number of individuals	MAF > 0.05	Super population	MAF > 0.05
CHB	Han Chinese in Beijing, China	103	64,521	EAS (East Asian)	64,810
JPT	Japanese in Tokyo, Japan	104	64,785		
CHS	Southern Han Chinese	105	64,509		
CDX	Chinese Dai in Xishuangbanna, China	93	64,267		
KHV	Kinh in Ho Chi Minh City, Vietnam	99	65,534		
CEU	Utah Residents (CEPH) with Northern and Western European Ancestry	99	68,818	EUR (European)	68,155
TSI	Toscani in Italia	107	68,808		
FIN	Finnish in Finland	99	69,113		
GBR	British in England and Scotland	91	67,296		
IBS	Iberian Population in Spain	107	68,676		
YRI	Yoruba in Ibadan, Nigeria	108	100,424	AFR (African)	96,674
LWK	Luhya in Webuye, Kenya	99	100,029		
GWD	Gambian in Western Divisions in the Gambia	113	97,879		
MSL	Mende in Sierra Leone	85	101,091		
ESN	Esan in Nigeria	99	101,502		
ASW	Americans of African Ancestry in SW USA	61	91,790		
ACB	African Caribbeans in Barbados	96	97,684		
MXL	Mexican Ancestry from Los Angeles, USA	64	68,247	AMR (Admixed American)	70,576
PUR	Puerto Ricans from Puerto Rico	104	72,095		
CLM	Colombians from Medellin, Colombia	94	70,452		
PEL	Peruvians from Lima, Peru	85	63,122		
GIH	Gujarati Indian from Houston, Texas	103	69,747	SAS (South Asian)	70,068
PJL	Punjabi from Lahore, Pakistan	96	70,290		
BEB	Bengali from Bangladesh	86	69,960		
STU	Sri Lankan Tamil from the UK	102	69,496		
ITU	Indian Telugu from the UK	102	69,499		

Cross-hybridization due to imperfect matches (e.g., mismatches/INDELs) and genetic variation (e.g., SNPs) is one of the most significant technical artifacts, where probes can map to multiple places in the genome and consequently assess a mixture of genuine and spurious signals [22]. Multiple studies have identified cross-hybridized probes in both the 450k and EPIC arrays, which affects a significant number of probes (6 to 11% of all probes). These findings contributed to the establishment of quality control practices for problematic probes on the EPIC array, resulting in the exclusion of a number of probes [10, 13, 18, 21]. Here, our technical characterization of the EPIC array showed that special care would be necessary when using this platform in epigenetic studies targeting diverse populations, because a substantial proportion of the interrogated CpG probes on the array contained common SNPs for major global populations. We therefore would like to provide those CpG probes on the array with potential technical biases (Supplemental Tables 1 and 2) as a useful resource for population-specific studies. For example, we can envision that investigators who identify any differential CpGs using the EPIC array will be benefited by referring this resource to make sure their findings are less biased to potential issues caused by cross-hybridization or common SNPs in a particular population, which will be critical for enhancing our knowledge of epigenetic contribution to complex traits and diseases.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s43682-022-00015-9.

Additional file 1.

Additional file 2.

Acknowledgements
The authors would like to thank colleagues from Northwestern University and University of Illinois for discussion on an early version of the manuscript.

Authors' contributions
Z.Z. and Z.W. made substantial contributions to conception. Z.Z. analyzed the data. Z.Z., Z.C., and Z.W. drafted the manuscript. All authors reviewed the manuscript. The author(s) read and approved the final manuscript.

Funding
This study was supported partially by a grant from the National Institutes of Health: R21MD011439. The funding body had no role in the experiment design, collection, analysis and interpretation of data, and writing of the manuscript.

Availability of data and materials
Not applicable.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Author details
[1]Department of Preventive Medicine, Northwestern University Feinberg School of Medicine, 680 N. Lake Shore Dr., Suite 1400, Chicago, IL 60611, USA. [2]The Robert H. Lurie Comprehensive Cancer Center, Northwestern University Feinberg School of Medicine, Chicago, IL 60611, USA.

Received: 20 September 2022 Accepted: 28 November 2022
Published online: 01 December 2022

References

1. Egger G, Liang G, Aparicio A, Jones PA. Epigenetics in human disease and prospects for epigenetic therapy. Nature. 2004;429(6990):457–63.
2. Bibikova M, Fan JB. Genome-wide DNA methylation profiling. Wiley Interdiscip Rev Syst Biol Med. 2010;2(2):210–23.
3. Ma X, Wang YW, Zhang MQ, Gazdar AF. DNA methylation data analysis and its application to cancer research. Epigenomics. 2013;5(3):301–16.
4. Tang J, Fang F, Miller DF, Pilrose JM, Matei D, Huang TH, et al. Global DNA methylation profiling technologies and the ovarian cancer methylome. Methods Mol Biol. 2015;1238:653–75.
5. Flanagan JM. Epigenome-wide association studies (EWAS): past, present, and future. Methods Mol Biol. 2015;1238:51–63.
6. Liu J, Siegmund KD. An evaluation of processing methods for HumanMethylation450 BeadChip data. BMC Genomics. 2016;17:469.
7. Marabita F, Almgren M, Lindholm ME, Ruhrmann S, Fagerstrom-Billai F, Jagodic M, et al. An evaluation of analysis pipelines for DNA methylation profiling using the Illumina HumanMethylation450 BeadChip platform. Epigenetics. 2013;8(3):333–46.
8. Niu L, Xu Z, Taylor JA. RCP: a novel probe design bias correction method for Illumina Methylation BeadChip. Bioinformatics. 2016;32(17):2659–63.
9. Andrews SV, Ladd-Acosta C, Feinberg AP, Hansen KD, Fallin MD. "Gap hunting" to characterize clustered probe signals in Illumina methylation array data. Epigenetics Chromatin. 2016;9:56.
10. Pidsley R, Zotenko E, Peters TJ, Lawrence MG, Risbridger GP, Molloy P, et al. Critical evaluation of the Illumina MethylationEPIC BeadChip microarray for whole-genome DNA methylation profiling. Genome Biol. 2016;17(1):208.
11. Moen EL, Zhang X, Mu W, Delaney SM, Wing C, McQuade J, et al. Genome-wide variation of cytosine modifications between European and African populations and the implications for complex traits. Genetics. 2013;194(4):987–96.
12. Zhang X, Mu W, Zhang W. On the analysis of the illumina 450k array data: probes ambiguously mapped to the human genome. Front Genet. 2012;3:73.
13. Zhou W, Laird PW, Shen H. Comprehensive characterization, annotation and innovative use of Infinium DNA methylation BeadChip probes. Nucleic Acids Res. 2017;45(4):e22.
14. Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, et al. A global reference for human genetic variation. Nature. 2015;526(7571):68–74.
15. Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, Huddleston J, et al. An integrated map of structural variation in 2,504 human genomes. Nature. 2015;526(7571):75–81.
16. Abecasis GR, Altshuler D, Auton A, Brooks LD, Durbin RM, Gibbs RA, et al. A map of human genome variation from population-scale sequencing. Nature. 2010;467(7319):1061–73.
17. Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, et al. An integrated map of genetic variation from 1,092 human genomes. Nature. 2012;491(7422):56–65.
18. Chen YA, Lemire M, Choufani S, Butcher DT, Grafodatskaya D, Zanke BW, et al. Discovery of cross-reactive probes and polymorphic CpGs in the Illumina Infinium HumanMethylation450 microarray. Epigenetics. 2013;8(2):203–9.

19. Kent WJ. BLAT - The BLAST-like alignment tool. Genome Res. 2002;12(4):656–64.
20. Dedeurwaerder S, Defrance M, Calonne E, Denis H, Sotiriou C, Fuks F. Evaluation of the Infinium Methylation 450K technology. Epigenomics. 2011;3(6):771–84.
21. Naeem H, Wong NC, Chatterton Z, Hong MK, Pedersen JS, Corcoran NM, et al. Reducing the risk of false discovery enabling identification of biologically significant genome-wide methylation status using the HumanMethylation450 array. BMC Genomics. 2014;15:51.
22. Hop PJ, Zwamborn RAJ, Hannon EJ, Dekker AM, van Eijk KR, Walker EM, et al. Cross-reactive probes on Illumina DNA methylation arrays: a large study on ALS shows that a cautionary approach is warranted in interpreting epigenome-wide association studies. NAR Genom Bioinform. 2020;2(4):lqaa105.

Publisher's Note