SNPs on DNA methylation microarrays: Precautions Against Confounding

Timothy Barrow[1, 2, 3] and Hyang-Min Byun[4,*]

[1]Institute for Prevention and Tumor Epidemiology, Freiburg Medical Center, University of Freiburg, 79106, Germany

[2]German Consortium for Translational Cancer Research (DKTK), Heidelberg, Germany

[3]German Cancer Research Center (DKFZ), Heidelberg, Germany

[4]Laboratory of Environmental Epigenetics, Exposure Epidemiology and Risk Program, Harvard School of Public Health, Boston, MA, USA

*Corresponding author

Hyang-Min Byun, hmbyun@hsph.harvard.edu

Key words: DNA methylation array, Infinium 450k, SNP

Financial disclosure

The authors have no financial or competing interests.

The purpose of this letter is solely to raise and emphasize the issue of possible artifacts arising from single nucleotide polymorphisms (SNPs) for users of Illumina methylation microarrays [1-4]. The authors would like to stress that there is no intention of discrediting or criticizing published works by other researchers or Illumina.

The Illumina Infinium microarray platform (Infinium Methylation 450K; Illumina, Inc. CA, USA) is the most widely-used approach for epigenome-wide analysis of DNA methylation. The HumanMethylation450 BeadCheap microarray interrogates methylation at >485,000 methylation sites that correspond to approximately 99 % of Refseq human genes and 96% of CpG islands [5]. The methylation detection probes on the array are 50 nucleotides in length (50-mer) and hybridize bisulfite-converted human genomic DNA sequences. These probes can be of two forms: the first, the Infinium I assay, uses separate “unmethylated” and “methylated” query probes; while the second, the Infinium II assay, represents the “all-or-none” approach and utilizes a single probe. Further details will not be provided in this letter, but can be found elsewhere [6]. The 50-mer probes are designed to interrogate a single CpG site that can potentially be methylated, at the 3’ end of probe. However, there may also be multiple other CpGs within the 50-mer probe. DNA methylation at the 3’ CpG site is measured using quantitative “genotyping” of bisulfite-converted genomic DNA. The bisulfite-conversion of DNA results in the deamination of unmethylated cytosines to uracil, leading to the incorporation of thymines during subsequent PCR-based amplification, while methylated cytosines are protected [7]. Therefore, at known CpG sites a ‘CG’ genotype will correspond to methylated sites within bisulfite-converted DNA and ‘TG’ to unmethylated sites. The level of DNA methylation at a given CpG site is then calculated by the ratio of fluorescent intensity (beta value) over the total, M/(M+U); M and U denote the average fluorescent signals from the methylated and unmethylated bead types of the probes, respectively.

The Infinium methylation microarrays offer the researcher a powerful tool to assess DNA methylation across the genome, but the possibility of technical artifacts needs to be taken into account, some of which can also be seen with other bisulfite-conversion-based techniques. In particular, the Infinium probes overlap with positions of known DNA variants. Based on the HumanMethylation450 BeadChip manifest file (a.k.a. Infinium 450k annotation file), 56% of the probes on the array (273,660 of 485,512 probes) contain at least one SNP within their 50-mer length [10]. . In this letter, we will discuss the issue of artifacts arising from SNP-associated probes, and how SNPs can potentially interfere with the assessment of DNA methylation to different degrees within the Infinium array.

Considerations for potential confounding by SNPs

The potential for confounding of results is dependent upon a number of factors associated with the SNP(s).

1) Type of Alleles: DNA variants can exist in different forms. Cytosine methylation has a spontaneous risk of deamination to thymine, thus eliminating the potential for the site to be methylated, and this event is more common than would be expected by random chance. If the SNP allele is either C to T or G to A (*vice versa*) at either position within a CpG site, the site may no longer be a candidate for DNA methylation. However, it is not possible to distinguish between changes in DNA sequence and changes in DNA methylation within the bisulfite-converted DNA, as a CpG to TpG genomic variant cannot be distinguished from an unmethylated cytosine following bisulfite conversion. Indeed, if the SNP exists at the C of the target CpG, therefore existing as either CG or TG, the DNA methylation array will serve only to perform as if a SNP array [4]. In contrast, if the SNP exists at the G of the target CpG, it will potentially inhibit the efficiency of hybridization of the bisulfite-converted genomic DNA sequences to the probes.

While interference of hybridization occurs with a variety of types of SNP, it is not known whether pyrimidine-pyrimidine (cytosine and thymine) mismatches are as problematic as purine-purine (adenine and guanine), as is the case with PCR primer annealing [11].

2) Distance: a SNP can be present anywhere within the 50-mer probe. The Infinium 450k annotation file classifies SNP-associated probes based upon the distance between the SNP and the target CpG (only those within 10 bases), on the assumption that SNPs closer to the target CpG site present a greater risk of impacting upon accurate measurement of methylation level. This issue is especially problematic when the risk of beta value variation from SNP artifacts is higher than actual biological variation, as is commonly the case with environmental epigenetics studies. Thus, technical data is required to establish the validity of this assumption.

3) Population diversity: a SNP can present with different frequencies across subsections of the human population, with some more prevalent in certain ethnicities than in others. Therefore, it is important to consider the population being studied and to utilize data regarding the frequency of SNPs within it, where available [12].

4) Number of SNPs: a probe may contain multiple SNPs within its sequence. Indeed, we have observed some probes to contain more than 16 SNPs (e.g. cg24170212) [10]. The number of SNPs within a probe will further impact upon the hybridization efficiency, and there is a need for data describing possible synergistic effects of multiple SNPs within the probe sequence.

The efficiency of hybridization of the probe is therefore dependent upon a range of factors. The impacts of SNPs upon beta values from associated probes are not expected to be uniform, and therefore can be difficult to identify. Some SNP loci clearly show binary beta values as would be expected from a SNP array, with beta values of 0.0 and 1.0 corresponding to the two homozygote genotypes and a value of 0.5 corresponding to heterozygotes [4]. However, the beta values from most SNP-associated probes are less predictable due to the multitude of factors affecting probe hybridization. Thus, caution must be taken in validating the ‘top hits’, such as the most significantly hyper- or hypomethylated loci between experimental groups by ranking, in order to determine whether beta values are truly reflecting DNA methylation patterns.

The impact of potential SNP-induced artifacts from DNA methylation microarrays is expected to be much greater within epigenetic epidemiology studies, where DNA methylation displays less variation than with cancer epigenetics, for example. Where technical variation from SNP loci is greater than biological variation, as would be the case with epigenetic epidemiology studies, it is reasonable to expect that ‘top hits’ may contain a higher prevalence of SNP-associated probes than would be expected by chance.

The potential for SNP artifacts on the methylation array can also be misinterpreted with the study of mQTLs (methylation quantitative trait loci), regarding SNPs whose genotype correlates with DNA methylation. If a SNP occurs in either the C or G of the target CG site, the SNP will directly interfere with DNA methylation, rather than simply correlate with it. For example, the rs8133082 (G/T) SNP is present within a CG site, thereby resulting in a CG or CT genotype. As DNA methylation can only occur at CG sites, the T genotype removes the potential for the cytosine to be methylated.

Possible Solutions

It is therefore important to be stringent when performing the data analysis, in order to determine whether beta values truly reflect variation in DNA methylation or could be the product of genotype or technical variation resulting from SNPs. Subsequently, it is now common to exclude SNP-associated loci in order to limit one source of confounding of results. However, as 56% of the probes on the Infinium array contain SNPs [10], it would not be appropriate to remove all SNP-associated probes from the analysis process, and therefore consideration must be paid to what parameters should be set. Accounting for the potential confounding by SNPs can be performed *a priori*, or *post hoc* of identifying the ‘top hits’.

1) a priori exclusion of SNP-associated loci: the exclusion criteria need to be carefully considered and based upon the aforementioned factors. This is sometimes performed, for example, by excluding probes where SNPs are present within 10 bases and show a minor allele frequency of more than 5% [3] (or 1% [13]) within the ethnicity of the study population. The identification of significantly different loci is then performed having excluded the probes which met these criteria.

2) post hoc exclusion of SNP-associated loci: as an alternative approach, the researcher may opt to carefully review all the top hits using a resource such as dbSNP or BLAT search [14], in order account for potential confounding. In the Illumina annotation file, the genomic sequence of the probe can be found under the header 'SourceSeq', along with the bisulfite sequence of the probe under the headers 'AlleleA_ProbeSeq' and 'AlleleB_ProbeSeq' (depending upon the Infinium probe type). The genomic sequences can then be used to perform a BLAT search in order to identify SNPs within the probe sequence.

In addition to taking SNP-associated loci into account during the analysis of the microarray data, the researcher may wish to consider further measures to ensure against confounding of results. The verification of DNA methylation at the 'top hits' by another method, such as pyrosequencing, would serve to inform upon the veracity of the results by utilizing approaches that are not as influenced by the presence of SNPs. It should be noted, however, that C/T SNPs directly affecting the cytosine residue of the target CpG sites would not be detectable by such methods. In such cases where a C/T SNP may be present, genotyping of the loci, such as by pyrosequencing or Sanger sequencing of unconverted genomic DNA would enable clarification.

SNPs represent a substantial challenge to researchers using microarray platforms, but with careful consideration it is possible to determine *bona fide* differences in DNA methylation and overcome confounding induced by SNPs, thereby enabling the researcher to utilize this tremendously powerful platform.

Reference:

1. Byun HM, Siegmund KD, Pan F *et al.* Epigenetic profiling of somatic tissues from human autopsy specimens identifies tissue- and individual-specific DNA methylation patterns. *Hum Mol Genet.* 18(24), 4808-4817 (2009).
2. Hinoue T, Weisenberger DJ, Lange CP *et al.* Genome-scale analysis of aberrant DNA methylation in colorectal cancer. *Genome Res.* 22(2), 271-282 (2012).
3. Touleimat N, Tost J Complete pipeline for Infinium((R)) Human Methylation 450K BeadChip data processing using subset quantile normalization for accurate DNA methylation estimation. *Epigenomics.* 4(3), 325-341 (2012).
4. Price ME, Cotton AM, Lam LL *et al.* Additional annotation enhances potential for biologically-relevant analysis of the Illumina Infinium HumanMethylation450 BeadChip array. *Epigenetics Chromatin.* 6(1), 4 (2013).
5. Infinium HumanMethylation450 BeadChip Kit.
6. Dedeurwaerder S, Defrance M, Calonne E, Denis H, Sotiriou C, Fuks F Evaluation of the Infinium Methylation 450K technology. *Epigenomics.* 3(6), 771-784 (2011).
7. Xi Y, Li W BSMAP: whole genome bisulfite sequence MAPping program. *BMC Bioinformatics.* 10, 232 (2009).
8. Triche TJ, Jr., Weisenberger DJ, Van Den Berg D, Laird PW, Siegmund KD Low-level processing of Illumina Infinium DNA Methylation BeadArrays. *Nucleic Acids Res.* 41(7), e90 (2013).
9. dbSNP Short Genetic Variations.
10. HumanMethylation450 BeadChip manifest file

11. Stadhouders R, Pas SD, Anber J, Voermans J, Mes TH, Schutten M The effect of primer-template mismatches on the detection and quantification of nucleic acids using the 5' nuclease assay. *J Mol Diagn.* 12(1), 109-117 (2010).

12. Internatinal HapMap Project.

13. Grundberg E, Meduri E, Sandling JK *et al.* Global analysis of DNA methylation variation in adipose tissue from twins reveals links to disease-associated variants in distal regulatory elements. *Am J Hum Genet.* 93(5), 876-890 (2013).

14. UCSC BLAT Search Genome.

15. Shenker NS, Polidoro S, Van Veldhoven K *et al.* Epigenome-wide association study in the European Prospective Investigation into Cancer and Nutrition (EPIC-Turin) identifies novel genetic loci associated with smoking. *Hum Mol Genet.* 22(5), 843-851 (2013).