On the analysis of the Illumina 450K array data: probes ambiguously mapped to the human genome

Xu Zhang[1], Wenbo Mu[2] and Wei Zhang[3,4,5]*

[1] Department of Ecology and Evolution, University of Chicago, Chicago, IL, USA
[2] Department of Bioengineering, University of Illinois at Chicago, Chicago, IL, USA
[3] Department of Pediatrics, University of Illinois at Chicago, Chicago, IL, USA
[4] Institute of Human Genetics, University of Illinois at Chicago, Chicago, IL, USA
[5] University of Illinois Cancer Center, Chicago, IL, USA
*Correspondence: weizhan1@uic.edu

A commentary on

The impact of recent alcohol use on genome wide DNA methylation signatures
by Philibert, R. A., Plume, J. M., Gibbons, F. X., Brody, G. H., and Beach, S. R. (2012). Front. Genet. 3:54. doi: 10.3389/fgene.2012.00054

The newly developed Illumina HumanMethylation450 BeadChip (450K array; Illumina, Inc., San Diego, CA, USA) allows unprecedented genome-wide profiling of DNA methylation at >450,000 CpG and non-CpG methylation sites (Sandoval et al., 2011). Utilizing the 450K array, Philibert et al. (2012) examined the relationship of recent alcohol intake to genome-wide methylation patterns in lymphoblast DNA samples derived from 165 female subjects participating in the Iowa Adoption Studies. The authors' interesting paper demonstrated that the 450K array could be a useful tool for ongoing and newly designed epigenome projects. However, given the unique design of the platform (for detailed annotations for the 450K array including probe sequences: http://www.illumina.com/), some cautions might need to be exerted when analyzing the 450K array data, in addition to some general challenges for analyzing the whole-genome DNA methylation data (Laird, 2010). Particularly, we found that a substantial proportion of the >450,000 DNA methylation probes on the 450K array are not aligned to unique, unambiguous loci in the human genome (Moen et al., 2012). In total, we found ~140,000 methylation probes ambiguously mapped to multiple locations in the human genome (hg19) with up to two mismatches in the probe sequences using Bowtie (v2.0.0 beta2; Langmead et al., 2009; Langmead and Salzberg, 2012). Briefly, Bowtie is an ultrafast, memory-efficient short read aligner by indexing the genome with an extended Burrows–Wheeler technique, which implements a novel quality-aware backtracking algorithm that permits mismatches (Langmead et al., 2009; Langmead and Salzberg, 2012). Different alignment algorithms, e.g., BLAT (Kent, 2002) and MAQ (Li et al., 2008), would provide similar estimates (unpublished data). In comparison, ~1,000 methylation probes were found to be ambiguously mapped to the human genome hg18 in the earlier 27K Illumina Human Methylation array (27K array; Bell et al., 2011). Because the much more comprehensive 450K array covers not only promoters, but also gene bodies, untranslated regions (UTRs) and "open sea" methylation sites, the problem of ambiguous alignment may particularly need to be taken into account when analyzing the data from this new platform. Notably, 20 CpG methylation probes (e.g., cg24023553 in Table 2; cg00004209 in Table 3; cg24675557 in Table 5) out of the 90 top-ranking probes reported by Philibert et al. (2012) were mapped to ambiguous loci in the current human reference (hg19) using Bowtie (Langmead et al., 2009; Langmead and Salzberg, 2012). Since the problem of ambiguous alignment to the human genome may cause unreliable measurement of DNA methylation level at a particular methylation site, considering this unique problem for this platform may not only facilitate the data analysis (e.g., by improving the multiple-testing problem by removing those affected probes), but also help interpret the results by focusing on more reliable biological signals. In addition, other factors (e.g., polymorphisms in the target sequences, potential batch effects) that may affect other platforms (e.g., the 27K array; Bell et al., 2011; Fraser et al., 2012) as well may also need to be considered in the analysis of these data.

ACKNOWLEDGMENT

REFERENCES

Bell, J. T., Pai, A. A., Pickrell, J. K., Gaffney, D. J., Pique-Regi, R., Degner, J. F., Gilad, Y., and Pritchard, J. K. (2011). DNA methylation patterns associate with genetic and gene expression variation in HapMap cell lines. *Genome Biol.* 12, R10.

Fraser, H. B., Lam, L. L., Neumann, S. M., and Kobor, M. S. (2012). Population-specificity of human DNA methylation. *Genome Biol.* 13, R8.

Kent, W. J. (2002). BLAT – the BLAST-like alignment tool. *Genome Res.* 12, 656–664.

Laird, P. W. (2010). Principles and challenges of genomewide DNA methylation analysis. *Nat. Rev. Genet.* 11, 191–203.

Langmead, B., and Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nat. Methods* 9, 357–359.

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* 10, R25.

Li, H., Ruan, J., and Durbin, R. (2008). Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res.* 18, 1851–1858.

Moen, L. E., Mu, W., Delaney, S., Wing, C., McQuade, J., Godley, L. A., Dolan, M. E., and Zhang, W. (2012). Differences in DNA methylation between the African and European HapMap populations. *Proc. Am. Assoc. Cancer Res.* 5010. [Abstract]

Philibert, R. A., Plume, J. M., Gibbons, F. X., Brody, G. H., and Beach, S. R. (2012). The impact of recent alcohol use on genome wide DNA methylation signatures. *Front. Genet.* 3:54. doi: 10.3389/fgene.2012.00054

Sandoval, J., Heyn, H., Moran, S., Serra-Musach, J., Pujana, M. A., Bibikova, M., and Esteller, M. (2011). Validation of a DNA methylation microarray for 450,000 CpG sites in the human genome. *Epigenetics* 6, 692–702.