Published online 22 April 2021

NAR Genomics and Bioinformatics, 2021, Vol. 3, No. 2 **1**
doi: 10.1093/nargab/lqab025

Human methylome variation across Infinium 450K data on the Gene Expression Omnibus

Sean K. Maden [1,2], Reid F. Thompson [1,2,3,4,5], Kasper D. Hansen [6,7,*] and Abhinav Nellore [1,2,8,*]

[1]Computational Biology Program, Oregon Health & Science University, Portland, OR 97239, USA, [2]Department of Biomedical Engineering, Oregon Health & Science University, Portland, OR 97239, USA, [3]VA Portland Healthcare System, Portland, OR 97239, USA, [4]Department of Medical Informatics & Clinical Epidemiology, Oregon Health & Science University, Portland, OR 97239, USA, [5]Department of Radiation Medicine, Oregon Health & Science University, Portland, OR 97239, USA, [6]Department of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, MD 21287, USA, [7]Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD 21205, USA and [8]Department of Surgery, Oregon Health & Science University, Portland, OR 97239, USA

Received December 01, 2020; Revised February 11, 2021; Editorial Decision March 09, 2021; Accepted April 19, 2021

ABSTRACT

While DNA methylation (DNAm) is the most-studied epigenetic mark, few recent studies probe the breadth of publicly available DNAm array samples. We collectively analyzed 35 360 Illumina Infinium HumanMethylation450K DNAm array samples published on the Gene Expression Omnibus. We learned a controlled vocabulary of sample labels by applying regular expressions to metadata and used existing models to predict various sample properties including epigenetic age. We found approximately two-thirds of samples were from blood, one-quarter were from brain and one-third were from cancer patients. About 19% of samples failed at least one of Illumina's 17 prescribed quality assessments; signal distributions across samples suggest modifying manufacturer-recommended thresholds for failure would make these assessments more informative. We further analyzed DNAm variances in seven tissues (adipose, nasal, blood, brain, buccal, sperm and liver) and characterized specific probes distinguishing them. Finally, we compiled DNAm array data and metadata, including our learned and predicted sample labels, into database files accessible via the `recountmethylation` R/Bioconductor companion package. Its vignettes walk the user through some analyses contained in this paper.

INTRODUCTION

DNA methylation (DNAm, Table 1) has been widely studied for its roles in normal tissue development (1–4), biological aging (5–7) and disease (8–12). DNAm regulates gene expression, either in *cis* if it occurs in a gene's promoter, or in *trans* if it overlaps an enhancer or insulator (4,9,13). Whole-genome DNAm (or 'methylome') analysis, especially in epigenome-wide association studies (EWAS), is a common strategy to identify epigenetic biomarkers with potential for clinical applications such as in prognostic or diagnostic panels (14–16).

Most investigations probe DNAm with array-based platforms. Published DNAm array data and sample metadata are commonly available through several public resources. These include cross-study databases like the Gene Expression Omnibus (GEO) (17,18) and ArrayExpress (19), as well as landmark consortium studies like the Cancer Genome Atlas (TCGA) (20) and the Encyclopedia of DNA Elements (ENCODE) (21,22). Recently published databases and interfaces provide access to samples from these sources (23–27).

While over 1604 DNAm array studies and over 104 000 samples have been submitted to GEO since 2009 (Supplementary Figure S1), there have been few attempts to rigorously characterize technical and biological variation across these studies. In 2013, two studies independently compiled DNAm array samples from GEO and elsewhere, analyzing epigenetic age across tissues and diseases (5), and investigating cross-study normalization (28). More recent cross-study analyses include (29) from 2018, which evaluated metadata

*To whom correspondence should be addressed. Tel: +1 443 910 1925; Email: nellore@ohsu.edu
Correspondence may also be addressed to Kasper D. Hansen. Tel: +1 510 333 5322; Fax: +1 410 955 0958; Email: khansen@jhsph.edu