Bioinformatics Advances, 2023, vbad020
https://doi.org/10.1093/bioadv/vbad020
Advance Access Publication Date: 20 February 2023
Original Paper

Databases and ontologies

recountmethylation enables flexible analysis of public blood DNA methylation array data

Sean K. Maden [1,2], Brian Walsh [1,2], Kyle Ellrott [1,2], Kasper D. Hansen [3,4,5], Reid F. Thompson [1,2,6,7,8] and Abhinav Nellore [1,2,9,*]

[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]Department of Genetic Medicine, Johns Hopkins School of Medicine, Baltimore, MD 21205, USA, [4]Department of Biostatistics, Johns Hopkins School of Public Health, Baltimore, MD 21205, USA, [5]Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD 21218, USA, [6]VA Portland Healthcare System, Portland, OR 97239, USA, [7]Department of Medical Informatics & Clinical Epidemiology, Oregon Health & Science University, Portland, OR 97239, USA, [8]Department of Radiation Medicine, Oregon Health & Science University, Portland, OR 97239, USA and [9]Department of Surgery, Oregon Health & Science University, Portland, OR 97239, USA

*To whom correspondence should be addressed.

Associate Editor: Aida Ouangraoua

Received on July 18, 2022; revised on December 29, 2022; editorial decision on January 26, 2023; accepted on February 17, 2023

Abstract

Summary: Thousands of DNA methylation (DNAm) array samples from human blood are publicly available on the Gene Expression Omnibus (GEO), but they remain underutilized for experiment planning, replication and cross-study and cross-platform analyses. To facilitate these tasks, we augmented our recountmethylation R/Bioconductor package with 12 537 uniformly processed EPIC and HM450K blood samples on GEO as well as several new features. We subsequently used our updated package in several illustrative analyses, finding (i) study ID bias adjustment increased variation explained by biological and demographic variables, (ii) most variation in autosomal DNAm was explained by genetic ancestry and CD4+ T-cell fractions and (iii) the dependence of power to detect differential methylation on sample size was similar for each of peripheral blood mononuclear cells (PBMC), whole blood and umbilical cord blood. Finally, we used PBMC and whole blood to perform independent validations, and we recovered 38–46% of differentially methylated probes between sexes from two previously published epigenome-wide association studies.

Availability and implementation: Source code to reproduce the main results are available on GitHub (repo: recountmethylation_flexible-blood-analysis_manuscript; url: https://github.com/metamaden/recountmethylation_flexible-blood-analysis_manuscript). All data was publicly available and downloaded from the Gene Expression Omnibus (https://www.ncbi.nlm.nih.gov/geo/). Compilations of the analyzed public data can be accessed from the website recount.bio/data (preprocessed HM450K array data: https://recount.bio/data/remethdb_h5se-gm_epic_0-0-2_1589820348/; preprocessed EPIC array data: https://recount.bio/data/remethdb_h5se-gm_epic_0-0-2_1589820348/).

Contact: anellore@gmail.com

Supplementary information: Supplementary data are available at *Bioinformatics Advances* online.

1 Introduction

DNA methylation (DNAm) is the most commonly studied epigenetic mark, and most public DNAm array samples are generated from blood (Maden *et al.*, 2021c). In prior work (Maden *et al.*, 2021c), we conducted comprehensive cross-study analyses of human DNAm array studies with raw data deposited on the Gene Expression Omnibus (GEO) (Barrett *et al.*, 2012; Edgar *et al.*, 2002), the largest archive of publicly available array data. We confined attention to the HumanMethylation450K (HM450K) platform introduced by Illumina in 2012. HM450K arrays profile 485 577 CpG loci concentrated in protein-coding genes and CpG island regions (Bibikova *et al.*, 2011; Sandoval *et al.*, 2011). We found that: (i) a subset of Illumina's prescribed BeadArray quality metrics explained most quality variances; (ii) samples clustered by tissue and cancer status in a principal component analysis (PCA) of autosomal DNAm; and (iii) subsets of CpG probes showed high tissue-specific DNAm