The American Journal of Human Genetics, Volume *108*

Supplemental information

Anatomy of DNA methylation signatures:

Emerging insights and applications

Eric Chater-Diehl, Sarah J. Goodman, Cheryl Cytrynbaum, Andrei L. Turinsky, Sanaa Choufani, and Rosanna Weksberg

Supplemental Methods

We used 10 cases each clinically diagnosed with Kleefstra syndrome (KS) and had pathogenic variants, i.e. either point mutation in *EHMT1* or microdeletions overlapping with *EHMT*1. We generated signatures for all possible combinations of cases ranging from 2-10 KS samples, compared to the 40 controls. The downsampling was performed using "combn" function in the CRAN package "utils". Groups of cases were limited to those that 1) contain both males and females, and 2) had both array batches represented[6]. This study design has a class imbalance between cases and controls, reflecting the current reality of sample acquisition and study design for signature derivation in rare disorders[14]. To generate the *EMHT1* signatures, we used R statistical software and the same methods as previously published[6]. Briefly, linear regression was performed on each CpGs, using all combinations of cases vs 40 controls, while covarying for sex and technical batch. Note, that this differs from the published DNAm signature, as blood cell composition and age were not included here as covariates to simplify the data analysis. This analysis was performed using the Bioconductor package "limma" and effect size was calculated as a difference between group means using base R functions. For each iteration of downsampling, an FDR-corrected p-value<0.05 and an absolute mean sample group difference of >10% was applied to generate the DNAm signature.

We then filtered the 837 signatures for those generated using less than three discovery cases and those which were composed of less than 2 CpGs. The remaining 609 unique signatures were used to train predictive models for KS using a support vector machine (SVM) algorithm. Typically, prior to SVM a feature selection step is performed which removes highly correlated CpGs; due to often small sizes of the signatures and the variable effect this step would have, this was not performed. For each signature, only discovery samples (KS cases and controls) from which the signature had been derived were used to train the corresponding predictive model. The analysis was performed using the Bioconductor package "caret". Each of the 609 models was then tested on a separate validation cohort of 10 KS cases to measure signature sensitivity and 175 controls to measure specificity. As all signatures were generated from the same number of control sample numbers, the resulting specificity of the DNAm signatures did not vary strongly.

We further tested specificity by running SVM on an additional 31 samples from individuals with NDDs. These samples were generated on the EPIC array and previously published as "discovery cases" used to derive signatures for pathogenic variants in the following epigenetic genes (disorder and sample sizes), *DYRK1A* (MIM: 600855; *DYRK1A*-related ID; n=14), *SMARCA2* (MIM: 600014; Nicolaides-Baraitser; n=8), *SRCAP* (MIM: 611421; Floating Harbour syndrome [FLHS]; n=4, non-FLHS *SRCAP*-related NDD; n=5). Of 837 signatures, 610 signatures contained enough CpGs (2 or more) and discovery cases (3 or more) to generate SVM classifications.

We then regenerated the 837 signatures using an FDR-corrected p-value<0.1 and an absolute mean sample group difference of 10%, using the same regression model. SVM was then applied to all signatures generated using three or more discovery cases and those which were composed of two or more CpGs (827 signatures). We imposed a maximum limit of 2000 CpGs on our signatures before passing them into the machine learning models, in order to make the computations more efficient. For signatures that exceeded this threshold, we ranked the CpGs by

their DNAm variance and then selected only the top 2000 most-varying CpGs, which should be quite sufficient to represent the most salient differences in DNAm in the data for the purpose of prediction. Similarly, to the previous set of signatures, each model was tested on a validation cohort of 10 KS cases and 175 controls. Datasets are not publicly available due to institutional ethics restrictions.

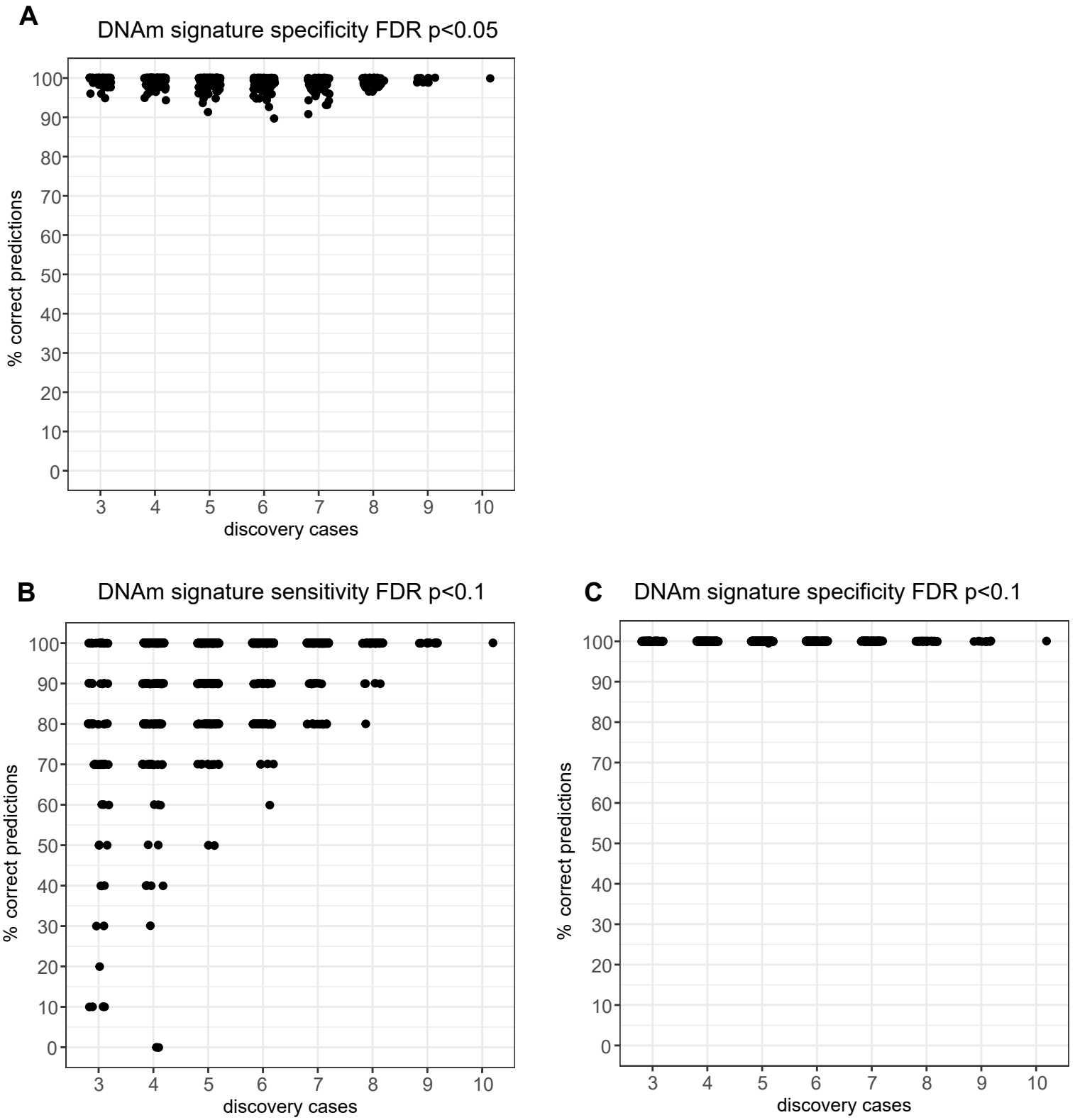

Figure S1. DNAm signature specificity and sensitivity. Predictive models from signatures generated for Kleefstra syndrome by iteratively resampling 3-9 discovery cases (n=10 total) vs. n=40 age- and sex-matched controls. A) Specificity of each signature meeting an FDR-corrected p-value<0.05, tested on controls (n=175; 609 signatures). Corresponding sensitivity plotted in Fig. 1C. B) Sensitivity of each signature meeting an FDR-corrected p-value<0.1, tested on the validation set of KS cases (n=10; 827 signature). C) Specificity of each signature meeting an FDR-corrected p-value<0.1, tested on controls (n=175; 827 signature).

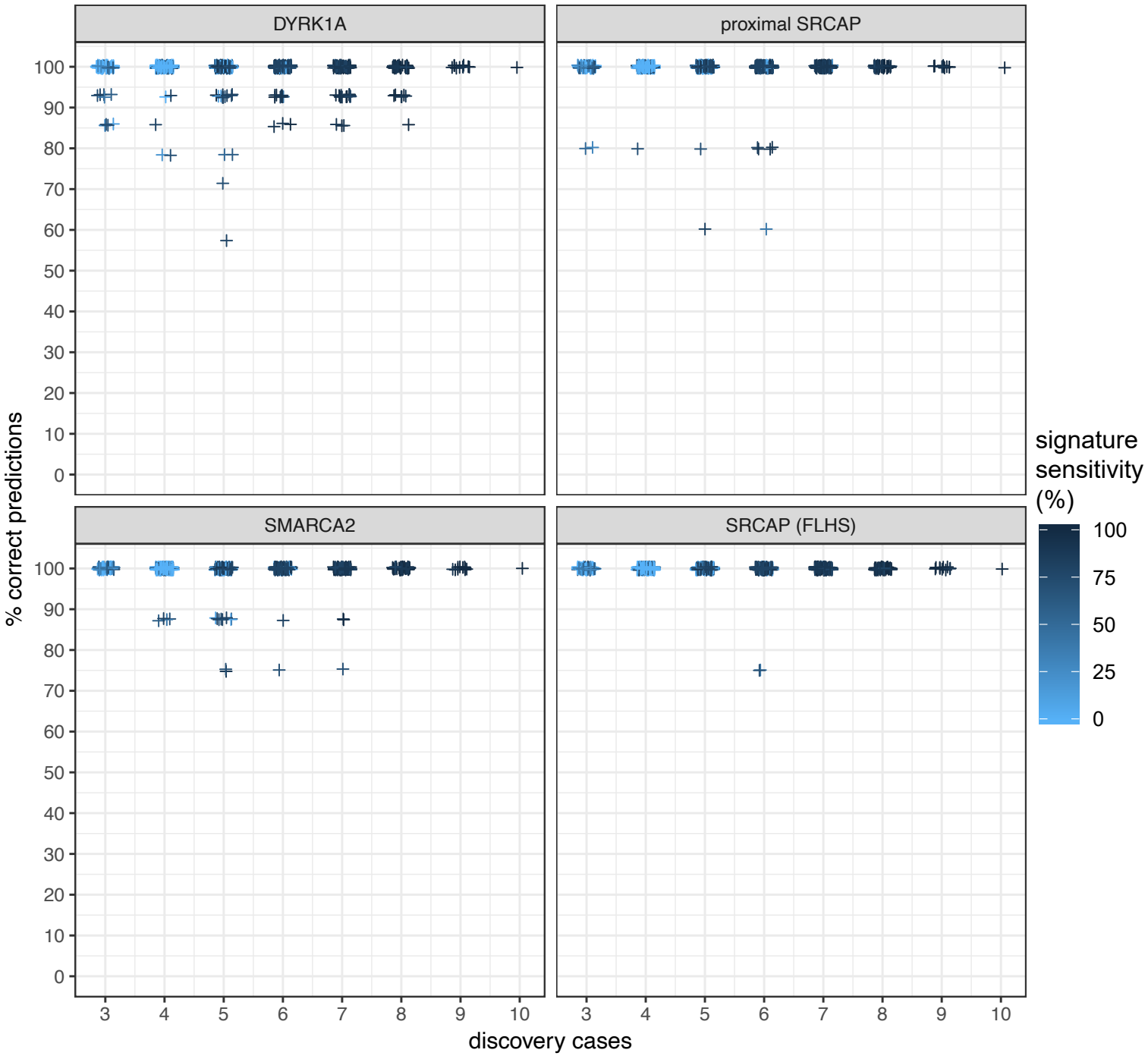

Figure S2. Specificity of DNAm signatures, as tested on 31 samples from individuals with NDDs. Signatures were generated for Kleefstra syndrome by iteratively resampling 3-9 discovery cases (n=10 total) vs. n=40 age- and sex-matched controls. Test samples include individuals with the pathogenic variants in the following genes: DYRK1A (DYRK1A-related ID; n=14), SMARCA2 (Nicolaides-Baraitser; n=8), SRCAP (Floating Harbour syndrome [FLHS]; n=4, non-FLHS SRCAP-related NDD; n=5). A total of 610 of 837 signatures were tested and plotted in each facet, represented by crosses.

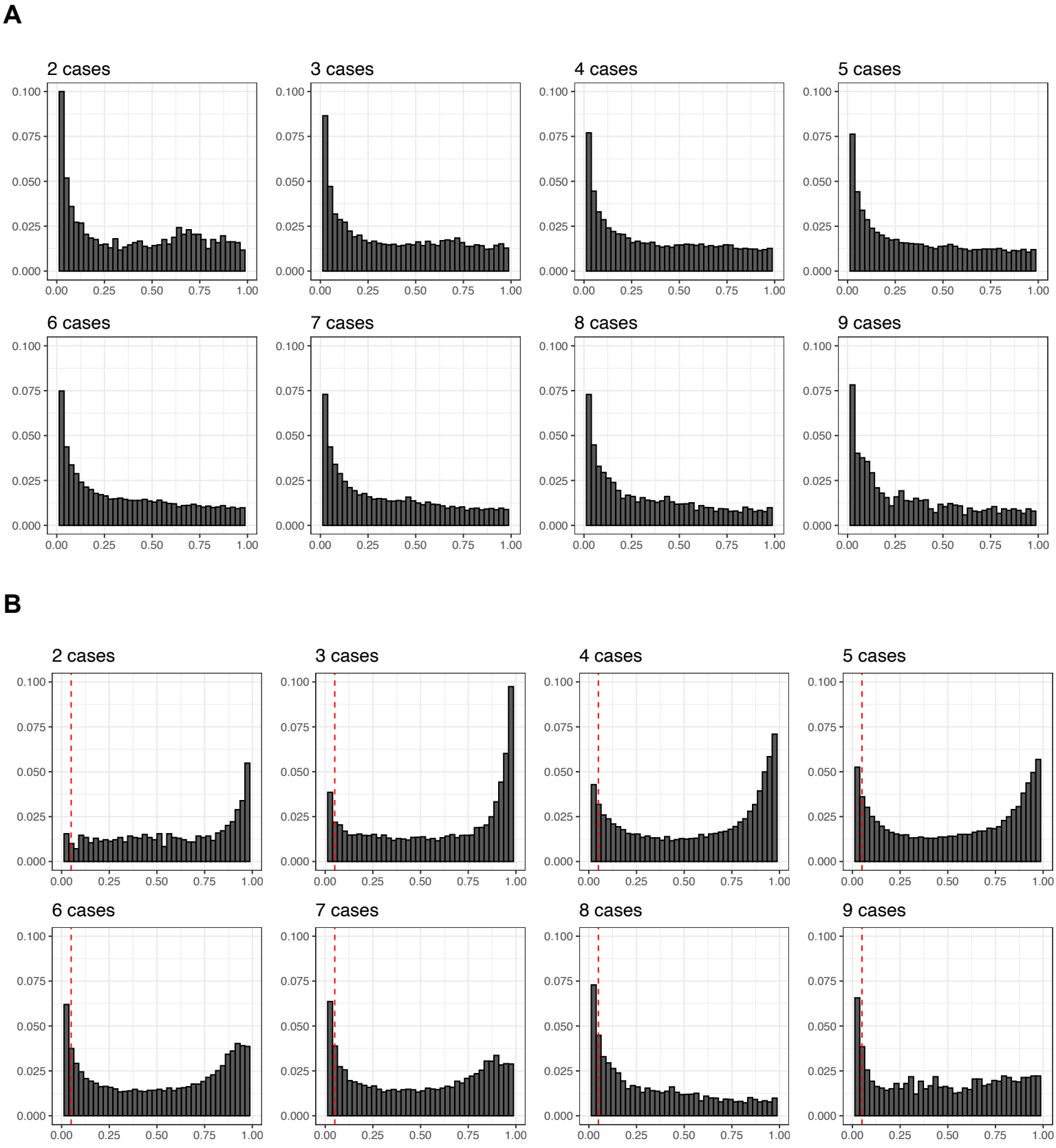

Figure S3. Uncorrected p-value (A) and FDR-corrected p-value (B) distributions of the 239 signature CpGs identified using 10 Kleefstra syndrome cases. Each line represents the results from all combined iterations of down-sampling, grouped based on the number of discovery cases used. Red dashed line represents FDR-corrected p-value=0.05.