KilnBio connects assay processing, multiomics integration and biological validation

Functional & Integrative Genomics
Varun B. Dwaraka and Michael J. Corley
Corresponding author: Varun B. Dwaraka, Independent researcher, Brooklyn, New York, USA
varun.dwaraka@gmail.com

Online Resource 6

GTEx methods and results for sensitivity to promoter construction and donor partitioning.

GTEx promoter sensitivity analysis

Online Resource 6 for KilnBio connects assay processing, multiomics integration and biological validation

Changing the promoter methylation summary changed the combined target shortlist on the same donors and gene family. This supplement describes how the analysis separated promoter representation, donor allocation and statistical confirmation, enabling researchers to trace the resulting changes in target priorities.

Data and sample matching

The published GSE213478 matrix contains 754,119 CpGs in 987 GTEx EPIC profiles, with the deposited ssNoob/BMIQ preprocessing. GTEx v8 gene read counts supplied RNA measurements. Matching donor and tissue identifiers yielded 555 pairs, including 516 with a shared specimen prefix. The primary analysis used the specimen-prefix matches in lung, ovary and transverse colon; pairing was defined at this deposited identifier level. Exact age and extended donor covariates came from supplied GTEx v9 phenotype metadata, provided with permission for this study. The source studies are Oliva et al. (2023) and Jain et al. (2024).

A deterministic donor hash assigned approximately 60% of donors to discovery and 40% to validation, with a donor assigned consistently across tissues. The original assignment used the prefix gtex-20261104: and five hash folds; folds 0–1 were validation and folds 2–4 were discovery. Singleton methylation plates were excluded within each tissue/partition to keep the fixed-plate HC3 model estimable. The original lung groups contained 89 discovery and 62 validation profiles; ovary contained 76 and 47; transverse colon contained 47 and 36.

Promoter features and age models

Autosomal cg probes annotated to a gene's TSS200 or TSS1500 regions were combined, requiring at least two probes per gene. Beta values were bounded at 10^{-6} and $1 - 10^{-6}$ before the $\log_2[\text{beta}/(1 - \text{beta})]$ transform. The original feature was the arithmetic mean of probe M-values. Alternative features were the median M-value, the M-transform of the mean beta value, and separate TSS200 and TSS1500 means, each requiring at least two probes. Gene summaries can share probes annotated to more than one gene.

Genes required a unique RNA symbol, nonzero molecular variation and RNA abundance of at least one count per million in 20% of discovery profiles. Separate linear models estimated RNA or promoter association with chronological age per decade. Covariates were sex where variable, RNA integrity, ischemic time and methylation plate. Standard errors used HC3 with residual-degree-of-freedom t inference. The estimand was age-associated variation in bulk tissue.

Ranking and confirmation

Each tissue supplied three 25-gene lists. RNA-only and methylation-only lists ranked the absolute age t statistic in the respective assay; the combined list averaged their percentile ranks. Gene symbols broke exact ties. The original lists and both assay directions were frozen before validation, yielding 208 distinct nominees across tissues and strategies. The frozen-nomination file has SHA-256 2dfafca7f07bb43a728a0d5a1b450428c95916f6e9c3071acb1dccbff32caf4f.

For each source/validation-tissue pair, the testing family was the union of that source tissue's three lists. One-sided validation P values tested the discovery direction. Two-assay confirmation used the maximum of the RNA and methylation P values, followed by BH adjustment across the union at $q \leq 0.05$. Individual-assay BH results and a Benjamini–Yekutieli conjunction sensitivity were also retained as separate outputs. Missing regional summaries and untestable RNA received $P = 1$ and retained their follow-up slots. Conjunction support was assessed using its own adjusted test across the declared family.

Sensitivity specifications

RNA sensitivities used edgeR TMM factors relative to a discovery-selected reference, retaining the same 0.5-count logCPM transformation and HC3 inference to isolate normalization. A separate standard TMM/voom/limma analysis assessed inference sensitivity. Mean-versus-median ranking comparisons used identical normalized RNA, discovery donors and genes available under both summaries. Region-specific comparisons likewise used the intersection of available gene families.

Additional covariates were BMI, recorded race and smoking status. BMI was complete in the primary tissues. Recorded race code 3 was contrasted with other known codes; two unknown values were replaced by the discovery modal category. Smoking was coded Yes/No; two Unknown records and one missing value were replaced by the discovery modal category. Recorded race and smoking entered as categorical donor metadata. The model omitted singleton missingness indicators to preserve HC3 estimability.

EpiSCORE estimates used the published lung and colon references with wRPC weight threshold 0.4 (Zhu et al. 2022). Reference genes sharing measured promoter probes with original nominees were excluded. Lung retained 237/369 reference genes and 139/190 informative genes; colon retained 80/104 and 60/75. Cell-composition adjustment covered these two tissues. A separate sensitivity added three principal components learned in discovery from non-nominated methylation genes, excluding genes sharing target probes. These components represented unprotected nuisance directions that could overlap age-related variation.

Twenty additional deterministic donor partitions repeated discovery selection and validation, preserving donor separation across tissues within each partition. Their overlapping donors make the results dependent. CpG checks fitted probe-level effects and repeated the promoter model after each single-probe removal; the conservative summary used the worst directional P value across removals before BH correction over the original nomination union. Fifty equal-size validation resamples per tissue retained 36 profiles and at least two profiles per plate. The sensitivity plan was recorded after the original validation outcomes were known.

Changes in the combined shortlist

Replacing the mean by the median promoter M-value changed 44–60% of the 25-gene list in the original discovery groups. The result recurred across the specified donor partitions, identifying promoter representation as a source of variation in the experiments prioritized by the combined ranking.

Tissue	Original changed / 25	Median changed / 25	Range in 20 splits
Lung	11	12	10–18
Ovary	15	11.5	6–18
Transverse colon	14	12	7–16

Confirmation and CpG coherence

None of the original frozen nominees achieved corrected two-assay confirmation in any of the 31 same-tissue or 89 source/validation-tissue specifications using the full original validation groups. In ovary, the original RNA confirmation yield was 15/25 for combined nominations and 13/25 for RNA-only nominations; TMM/HC3 gave 13/25 versus 14/25, and voom gave 17/25 in both lists.

Tissue	Direction stable after every probe removal	Methylation $q \leq 0.05$ after every removal
Lung	21/25	0/25
Ovary	24/25	0/25
Transverse colon	23/25	0/25

Across the 20 additional partitions, newly selected ovarian combined lists yielded 0–14 joint confirmations (median 3). RNA-only lists yielded 0–5 (median 1), and methylation-only lists yielded 0–7 (median 1.5). Combined selection exceeded RNA-only selection in 12 partitions, tied in six and trailed in two. Lung and colon lists had no joint confirmations in these partitions. Equal-size subsamples of the original ovarian validation group yielded 0–1 joint confirmations for the original combined list, compared with zero in lung and colon.

Target stability and tissue differences

AQP1 was selected in all 20 ovarian combined lists and jointly confirmed in ten partitions. Its original conjunction q was 0.111, and no original-split sensitivity specification reduced it below 0.05. AQP1 therefore combined stable nomination with partition-dependent confirmation. Donor-clustered tissue-by-age comparisons produced 15 globally BH-significant individual-assay interactions and zero robust two-assay tissue-specific targets. No nominated gene passed global BH correction for age- and covariate-adjusted RNA–promoter coupling. Tissue specificity was assessed through direct interaction tests.

Interpretation and reproducibility

The empirical contribution is that a choice within the methylation assay can change which experiments a combined ranking prioritizes. GTEx methylation–expression relationships and chronological-age associations were reported by the source studies (Oliva et al. 2023; Jain et al. 2024). The present analysis quantifies the sensitivity of target selection to promoter representation while recording the donors, feature family and confirmation rule used for each comparison.

The outputs distinguish probe-direction coherence, promoter-level association, shortlist membership and corrected two-assay confirmation. Reading these together shows whether a gene retains its nomination across feature definitions and donor partitions, and how that stability relates to statistical support. This makes the consequences of a preprocessing choice inspectable before a laboratory selects experiments for follow-up.

Reproducing the GTEx summary

Online Resource 7 contains aggregate tables and the analysis scripts used for the GTEx extension. Running reproduce_summary.py with Python's standard library reconstructs the principal counts and ranges from the supplied tables. Full molecular rerunning uses separately obtained donor-level matrices, exact ages, extended phenotypes and the scientific Python/R environment documented in the package. Online Resource 7 distributes the aggregate results and code.

The aggregate tables cover all original-list specifications, CpG coherence, donor-partition nominations and confirmations, controlled promoter-summary comparisons, equal-size validation and tissue/coupling checks. An inventory records SHA-256 hashes. Source scripts retain their original analysis paths relative to the project root, with personal filesystem identifiers removed if present. The original frozen-nomination file is included unchanged.

Computational checks

The preparation read all 754,119 published beta rows, retained 155,227 relevant CpGs and reproduced the original promoter means. The original confirmation counts were reproduced. HC3 coefficients, standard errors and P values were checked against statsmodels, as were BH calculations. All 126 pairwise TMM checks agreed with edgeR to numerical precision. Discovery and validation donors were disjoint within each partition. Together these checks establish numerical consistency of preparation and inference.

Data access and computational environment

Public inputs are GSE213478 methylation beta values and GTEx Analysis v8 gene read counts. Exact age and extended donor metadata must be obtained through the applicable GTEx access route or an authorized source; the supplement distributes aggregates to protect donor-level information. The aggregate tables report association estimates. Residual-pair permutation diagnostics assume exchangeability. The supplemental analysis used Python 3.12.14, NumPy 2.5.4, pandas 3.1.0, SciPy 1.18.1 and statsmodels 0.15.0; the R session is recorded in Online Resource 7.