

Robust Integrative Genomic Association (RIGA): Uncovering Complex Genetic Architecture from Diverse Multi-Cohort Studies

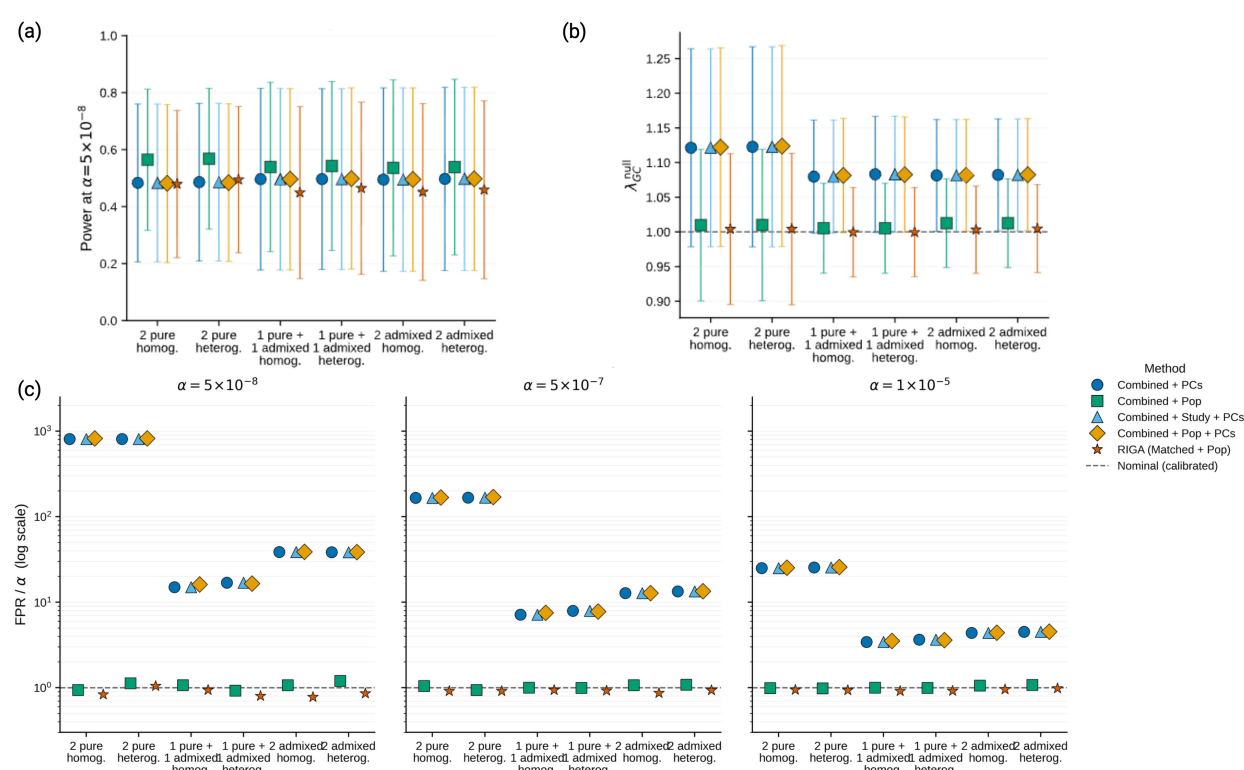
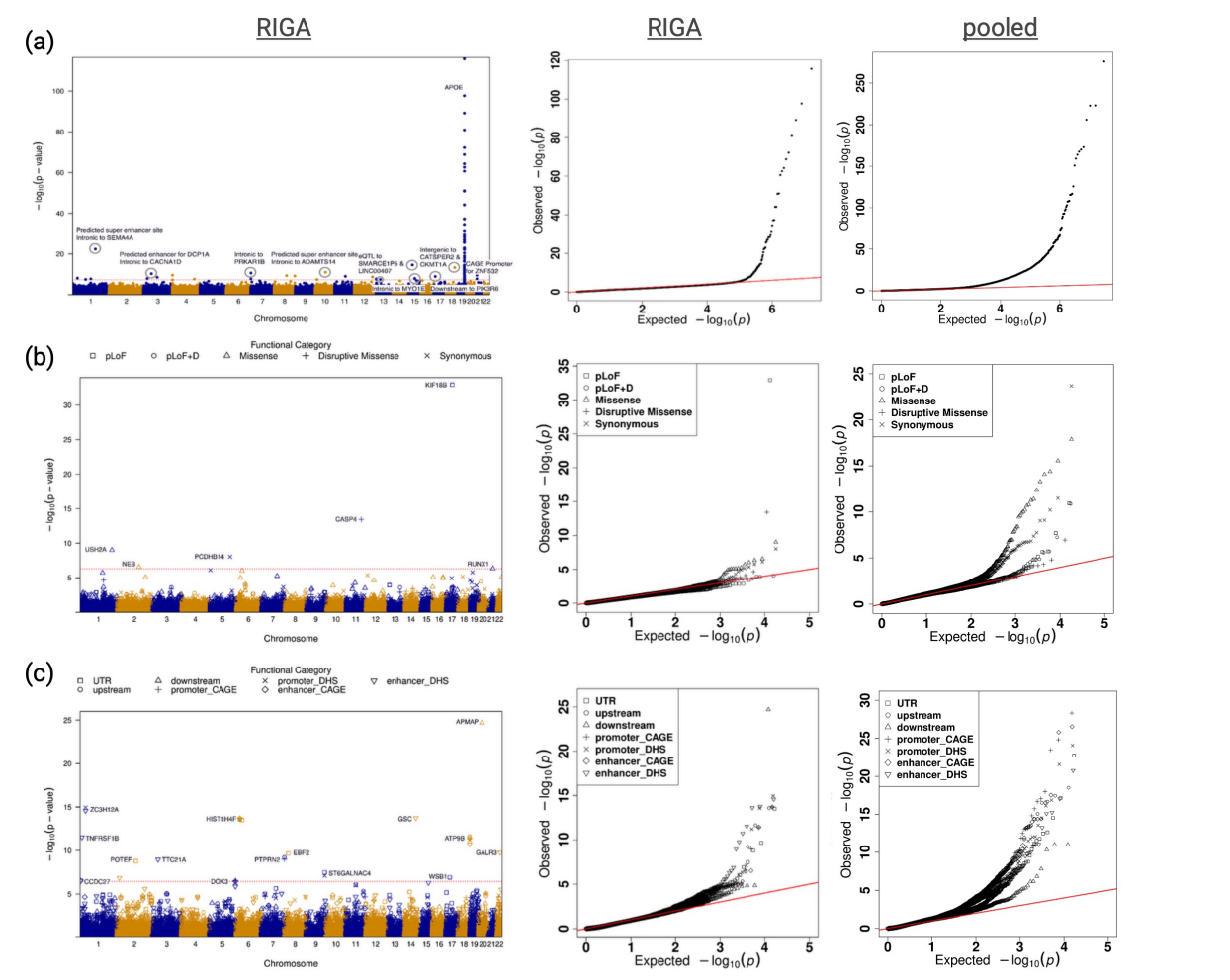
P2-03

Xiaoqi Li¹, Xihao Li^{1,2}, Borui Tang¹, Tengfei Li^{1,3}, Weili Lin³, Hongtu Zhu^{1*}

¹Department of Biostatistics, University of North Carolina at Chapel Hill ²Department of Biostatistics, University of North Carolina at Chapel Hill

³Biomedical Research Imaging Center, Department of Radiology, UNC–Chapel Hill *Corresponding: htzhu@email.unc.edu

Conventional pooled and PC-adjusted GWAS models inflate false positives up to ~800× in heterogeneous multi-cohort data. **RIGA** — latent-profile population modeling + retrospective case–control matching — restores genome-wide calibration ($\lambda \approx 1.00$), recovering canonical Alzheimer’s genes (*APOE*) and novel risk loci across 58,507 ADSP whole genomes.

INTRODUCTION	METHODS	RESULTS	CONCLUSIONS	REFERENCES
The premise. Pooling multi-ancestry cohorts boosts GWAS power, but between-cohort heterogeneity in design, demographics, and population structure could inflate false positives and obscure true signals. The challenge (ADSP). 50+ studies spanning case–control, family, and prospective designs; 10 sequencing centers; 4 ancestry groups, forming multi-layer heterogeneity of center, cohort, and ancestry. Four nested confounders. Study design, sampling design, covariate imbalance (age, sex, <i>APOE</i> ϵ4), and fine-scale population structure. Standard fixes fail. Principal components, study fixed-effects, and discrete ancestry labels cannot absorb this structure → genomic inflation and unstable models.	Data — ADSP R5 WGS. 58,507 individuals, 50 studies, 10 centers; 361M biallelic variants (GRCh38). Ancestry: 29.3K European, 15.1K Admixed American, 8.6K African, 2.9K East Asian, 2.7K South Asian. Quality control. bcftools filtering; ancestry-aware Hardy–Weinberg (RUTH); PCs + sparse GRM (fastSparseGRM) [1]; FAVOR functional annotation [2]. RIGA – two stages. (1) Gaussian mixture model on genetic PCs → BIC-optimal latent subpopulations [3]; (2) retrospective case–control matching on age, sex, and latent population. Association testing. STAARpipeline single-variant and gene-centric STAAR-O [4]; multicategorical SKATMC [5]. Validated on 10,000 simulation replicates × 27 architectures × 6 confounding scenarios.	 Figure 1. Simulation results benchmarking five association strategies across six simulated multi-cohort scenarios (10,000 replicates).  Figure 2. RIGA-matched Manhattan (left) and QQ (center) vs. pooled-analysis QQ (right) for (a) single-variant, (b) rare coding STAAR-O, and (c) non-coding STAAR-O analyses. The pooled column (right) shows the genome-wide inflation that RIGA's matching design removes.	Standard frameworks could fail on large, diverse consortia. RIGA restores calibration via matching + latent-profile correction, validated in simulation and ADSP. Biology. Positive controls (<i>APOE</i>, <i>PSEN1</i>, <i>TOMM40</i>) plus novel loci in calcium signaling, brain energy metabolism, cytoskeletal dynamics, and the non-canonical inflammasome (<i>CASP4</i>, <i>KIF18B</i>). Disease modifiers. Multicategorical SKATMC highlights <i>SNCAIP</i> ($p = 4.8 \times 10^{-28}$)—α-synuclein co-pathology as a modifier of AD severity. Impact. A scalable framework for valid, equitable multi-ancestry genomic discovery.	- Lin et al. Scalable computation of principal components and a sparse GRM (fastSparseGRM). 2024. - Li X, et al. Dynamic incorporation of in silico functional annotations for rare-variant association testing (STAAR). Nat Genet. 2020. - Scrucca et al. mclust 5: clustering, classification and density estimation. R Journal. 2016. - Li X, et al. A framework for detecting noncoding rare-variant associations of large-scale WGS studies. Nat Methods. 2022. - Jiang et al. A sequence kernel association test for multicategorical outcomes (SKATMC). Genetic Epidemiology. 2023.
CONTACT				
 Xiaoqi Li Dept. of Biostatistics, UNC–Chapel Hill Corresponding: Hongtu Zhu htzhu@email.unc.edu Acknowledge grant (TODO): 				