

Enabling robust biological discovery through consortium-scale plasma proteomics

Caitlin A. Finney¹, Lijun An², Laura M. Winchester³, Jacob Vogel², Heather M. Wilkins⁴, Jeffrey M. Burns⁴, Russell H. Swerdlow⁴, Chad Slawson⁴, Jeffrey D. Rothstein⁵, Global Neurodegeneration Proteomics Consortium, Michael W. Lutz⁶, Rowan Saloner⁷, & **Artur Shvetcov**¹

1. Westmead Institute for Medical Research, Australia; 2. Lund University, Sweden; 3. University of Oxford, UK; 4. University of Kansas Alzheimer's Disease Research Center, USA; 5. Johns Hopkins University, USA; 6. Duke University, USA; 7. University of California San Francisco, USA

INTRODUCTION

Large-scale plasma proteomics offers unprecedented opportunities to identify biomarkers and biological signatures of neurodegenerative disease. However, technical heterogeneity, site-specific artifacts, and clinical imbalance remain major barriers to reproducible inference in multi-site studies. In real-world large consortium datasets, technical variance can rival or exceed disease-associated signal, while aggressive harmonization risks attenuating meaningful biology. Critically, batch correction is often evaluated in isolation, without explicit assessment of biological signal preservation or cross-site robustness under realistic confounding.

METHOD

We leveraged the new Global Neurodegeneration Proteomics Consortium (GNPC) dataset that captures 6,430 proteins for the plasma of 13,733 individuals with Alzheimer's disease (AD), Parkinson's disease (PD), frontotemporal dementia (FTD), Parkinson's disease dementia (PDD), amyotrophic lateral sclerosis (ALS), and non-impaired controls across 16 cohorts (Fig. 1).

We developed a multi-component benchmarking framework to evaluate batch correction performance. Five common approaches (z-score normalization, linear model residualization, ComBat, ComBat with site rebalancing, and limma) were assessed using global variance structure, multivariate site mixing, and protein-level variance decomposition

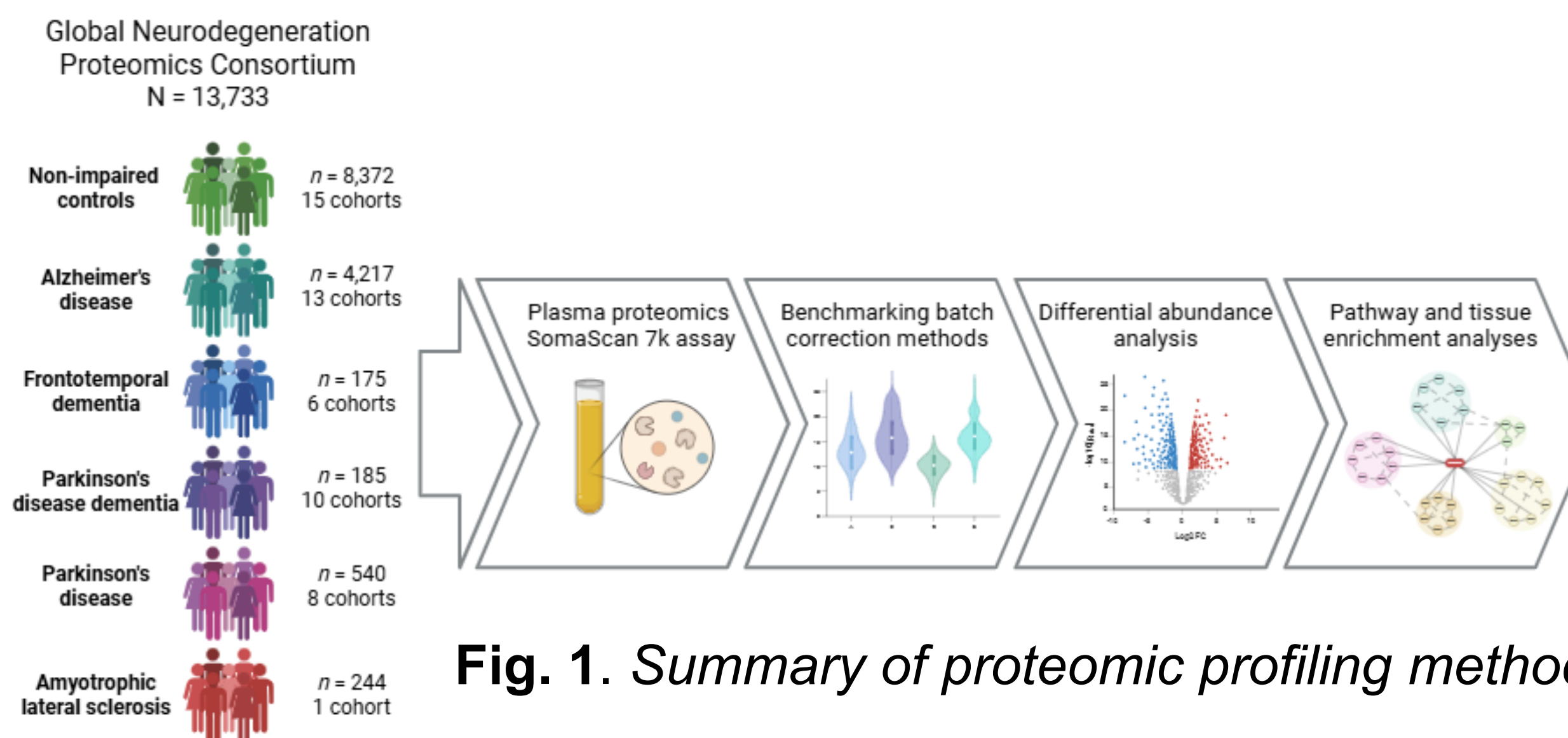


Fig. 1. Summary of proteomic profiling method

via mixed-effects modelling. Biological robustness was evaluated using site-stratified differential modelling, meta-analytic consistency diagnostics, and a high-dimensional perturbation framework that injects controlled outcome signals while varying batch-outcome confounding, heteroskedasticity, nonlinearity, and signal sparsity across 144 analytically realistic scenarios.

RESULTS

In the uncorrected data, site effects (Fig. 2a) accounted for substantially more variance than diagnostic status (Fig. 2b), underscoring the need for explicit harmonization.

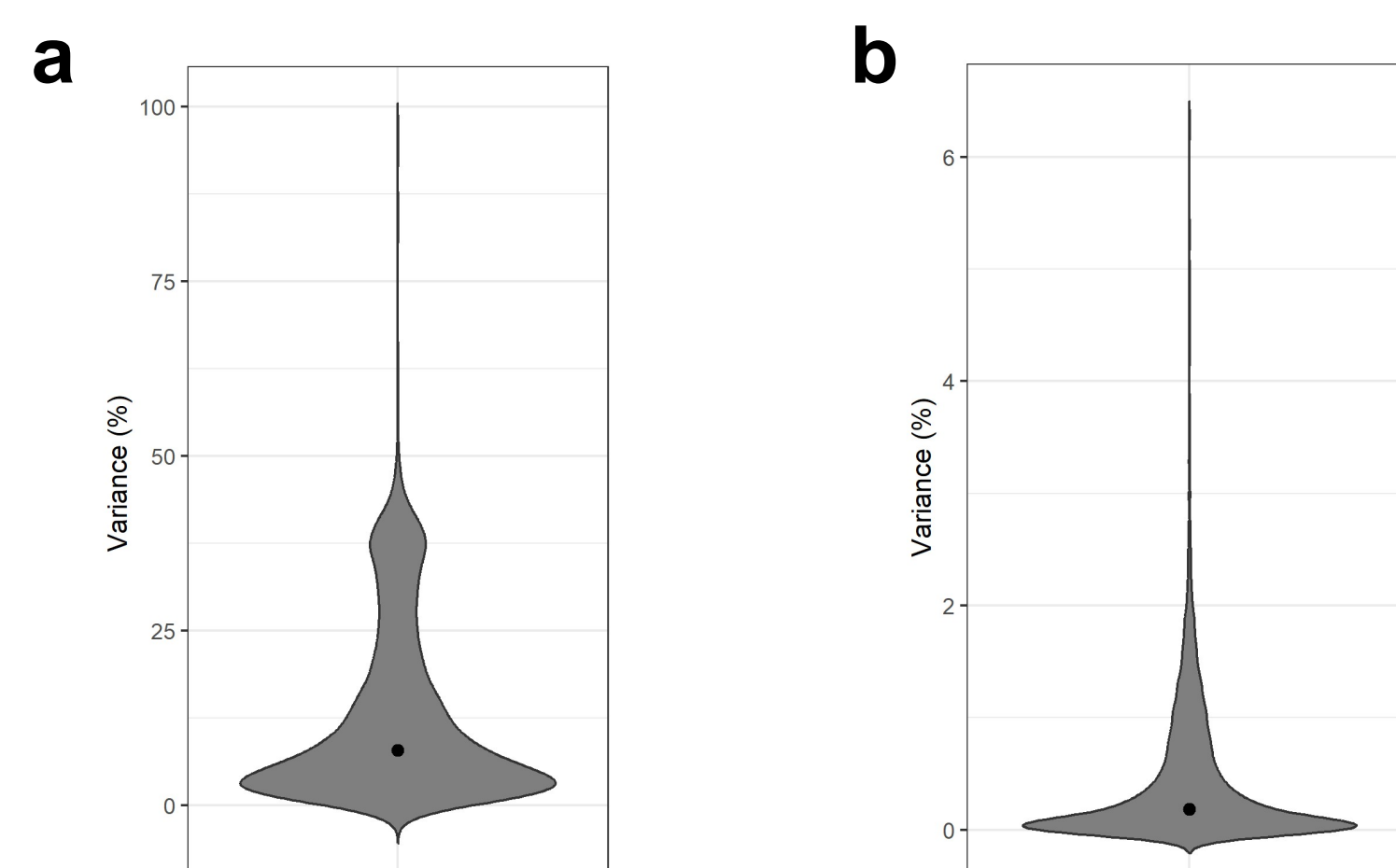


Fig. 2. Variance in uncorrected raw plasma proteomic data for (a) site and (b) outcome.

Although all methods reduced site-associated variance (Fig. 3a), their effects on biological signal preservation differed markedly. Z-score normalization and linear residualization consistently attenuated diagnosis-associated variance, with only 33-40% retained (Fig. 3b,c). In contrast, empirical Bayes approaches preserved substantially more biological signal (Fig. 3b,c). ComBat approaches retained 52-54% of signal and limma preserved the greatest fraction of diagnostic variance (66%) (Fig. 3b,c).

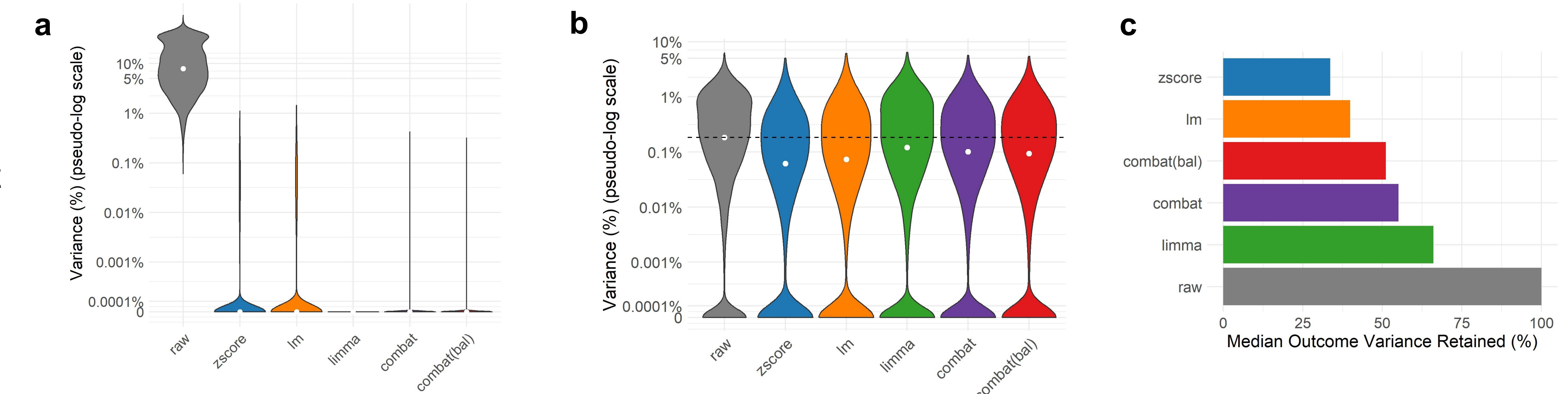


Fig. 3. Variance in retained across batch correction methods for (a) site and (b-c) outcome.

Across perturbation scenarios, limma showed the highest signal recovery and stable performance under severe confounding and heteroskedasticity, matching or outperforming alternative methods.

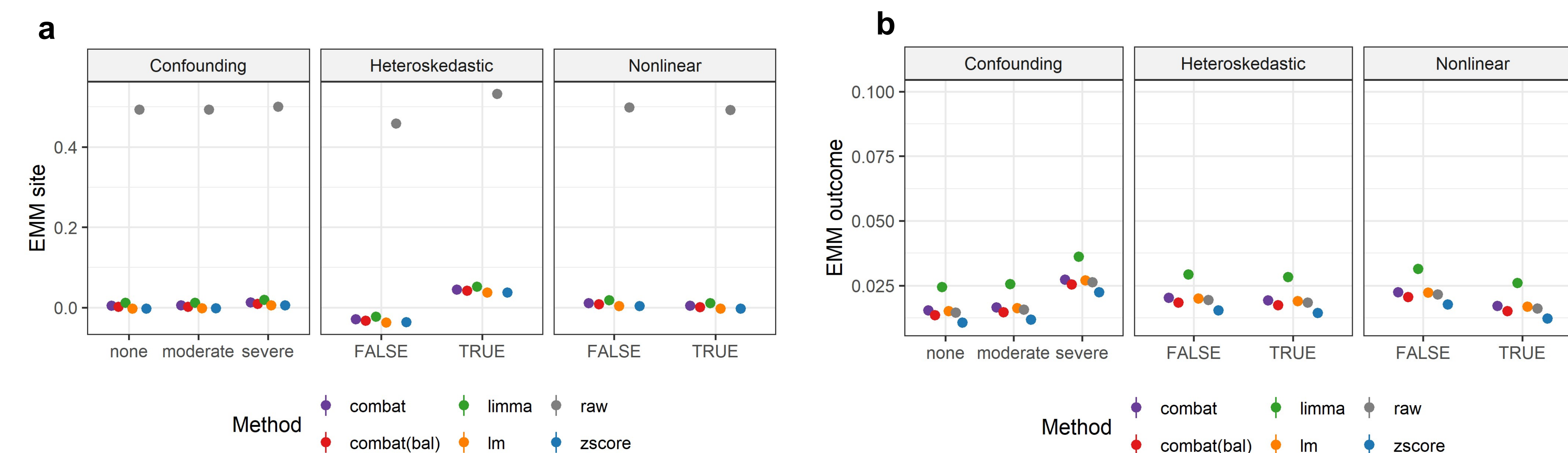


Fig. 4. Variance in retained across perturbation scenarios for (a) site and (b) outcome.

CONCLUSION

We introduce a generalizable framework for evaluating batch correction in consortium-scale plasma proteomics that integrates variance partitioning, meta-analytic robustness, and perturbation-based signal recovery. Model-based empirical Bayes approaches, particularly limma, provide superior robustness for harmonizing multi-site clinical proteomic data while preserving biological signal. This framework provides a rigorous foundation for reproducible biomarker discovery and biological inference in large neurodegenerative disease consortia.

CONTACT: artur.shvetcov@wimr.org.au