Exploration of Feature Selection Strategies in DNA Methylome-based Cancer Classification

Jenny Li (COL ’26) University of Pennsylvania College of Arts and Sciences
PI: Wanding Zhou, Assistant Professor of Pathology and Laboratory Medicine, Perelman School of Medicine

Introduction

- DNA cytosine modification at CpG dinucleotides are rich encoders of a cancer cell’s mitotic history and cell-of-origin information, establishing the DNA methylome as a powerful molecular analyte for cancer diagnosis
- Previous studies affirm the effectiveness of machine learning algorithms, notably Random Forest classifiers, in the accurate prediction and categorization of tumors
- Our research is centered on predicting 66 Central Nervous System (CNS) brain tumor categories (82 subcategories) from the Capper cohort and 33 cancer types from The Cancer Genome Atlas (TCGA) cohort
- Methylation data is generated from various iterations of Infinium BeadChip assays, including HM450, EPICv1, and EPICv2 platforms
- Classifiers trained on one methylation assay platform may not translate effectively to other assay platforms due to probe selection changes and platform-specific technical artifacts such as signal background and amplification bias
- Objective: Explore various feature selection methods to optimize the performance of random forest classifiers in predicting tumor classes, with the aim of expanding their utility in clinical diagnostic contexts

Methods

- Employed the randomForest and ggplot2 packages in R for classifier training and data visualization
- Utilized the High-Performance Computing (HPC) cluster at CHOP (Children's Hospital of Philadelphia) for data retrieval and job script execution
- Created t-SNE plots with various datasets, including 210 samples from CHOP’s Division of Genomic Diagnostics (DGD), TCGA, CBTN (Children's Brain Tumor Network), and Capper datasets
- Feature Selection Strategies: Conducted investigation into two primary avenues for feature selection:
    1. Biologically-informed CpG aggregation guided by tissue signature knowledgesets
    2. Nonparametric rank transformation of beta values
- Biological Knowledgebase-Aided Feature Aggregation: Aggregated CpG methylation levels by calculating mean across all CpG sites within each feature set, guided by individual tissue signature databases for model training
- Tissue signature databases included transcription factor binding sites (TFBS), chromatin states (chromHMM), histone modifications (HM), and chromosomal loci, among others
- Nonparametric rank transformation: Executed nonparametric rank transformation on raw CpG probe values after feature selection, focusing on 6,636 features of highest significance

Feature Aggregation Results

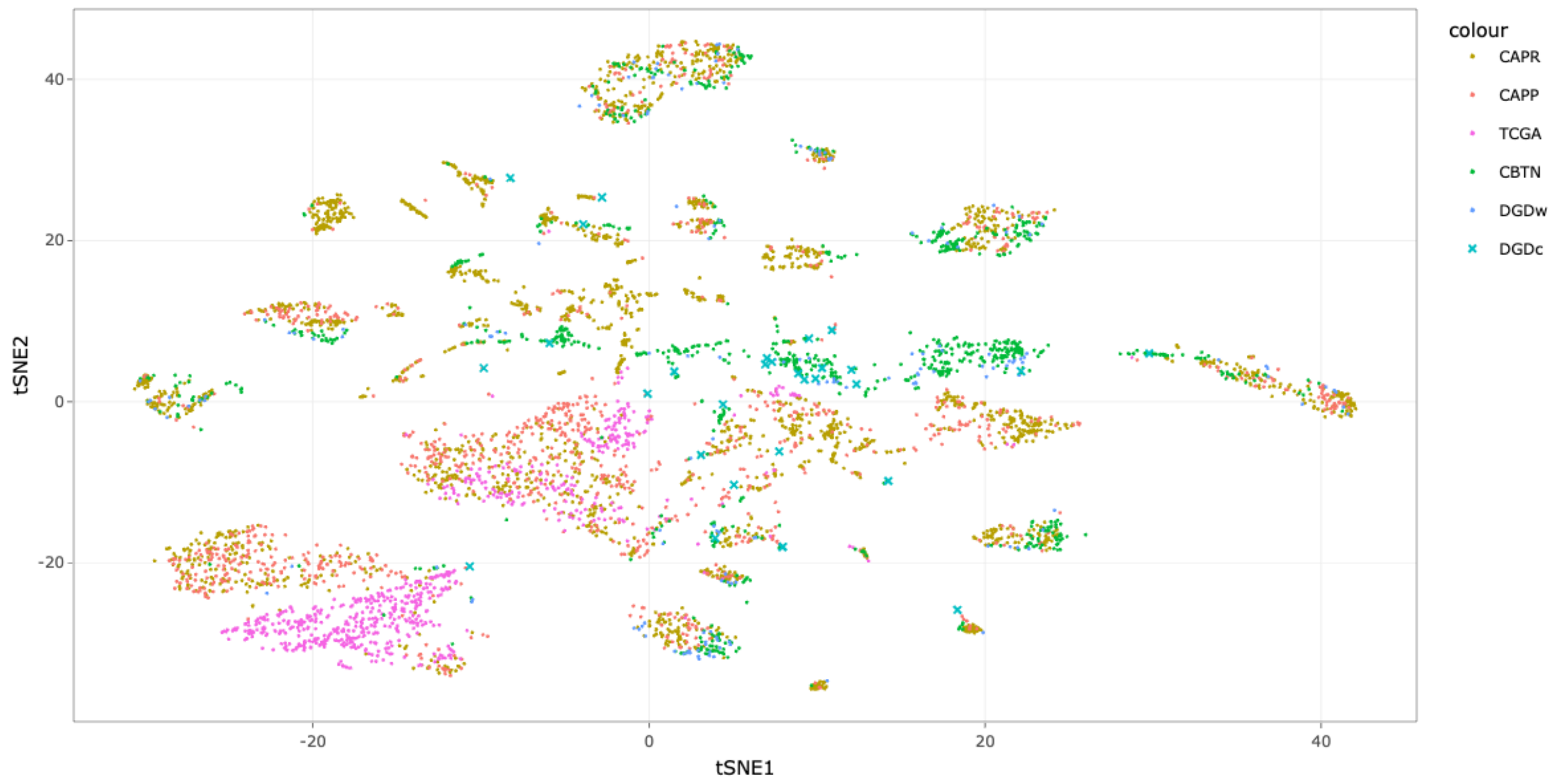

Fig 1. Capper, TCGA, CBTN, and DGD cohorts plotted in t-SNE plot

- During clustering analysis, we identified pronounced batch effects that manifested as distinct separations among CBTN, TCGA, and Capper cohorts
- Batch effects compromise our ability to definitively categorize certain samples within a given class, making it difficult to add to the re-curated training set
- Limits generalizability of machine learning models trained on one dataset but tested on another, thereby affecting the reliability of predictive outcomes
- In light of these challenges, our research pivoted towards investigating various feature selection and aggregation strategies

- TiSig (Tissue Signature) databases showed highest accuracies, with TiSigLoyfer (Tissue-specific methylation in human sorted cell types) up to 91.8% estimated accuracy
- Our analysis revealed direct correlation between number of features within knowledgebase set and classification accuracy, with larger feature sets contributing to increased accuracy
- Upon conducting feature selection processes of equal size across various databases, we observed that biologically-informed feature sets generally yielded superior performance compared to randomly selected feature sets This trend was especially pronounced in the ChromHMM, HM, Chromosome, and ABCCompartment knowledge sets
- Contrarily, models based on Centromere and nFlankCG features exhibited suboptimal performance when compared to their randomly selected counterparts

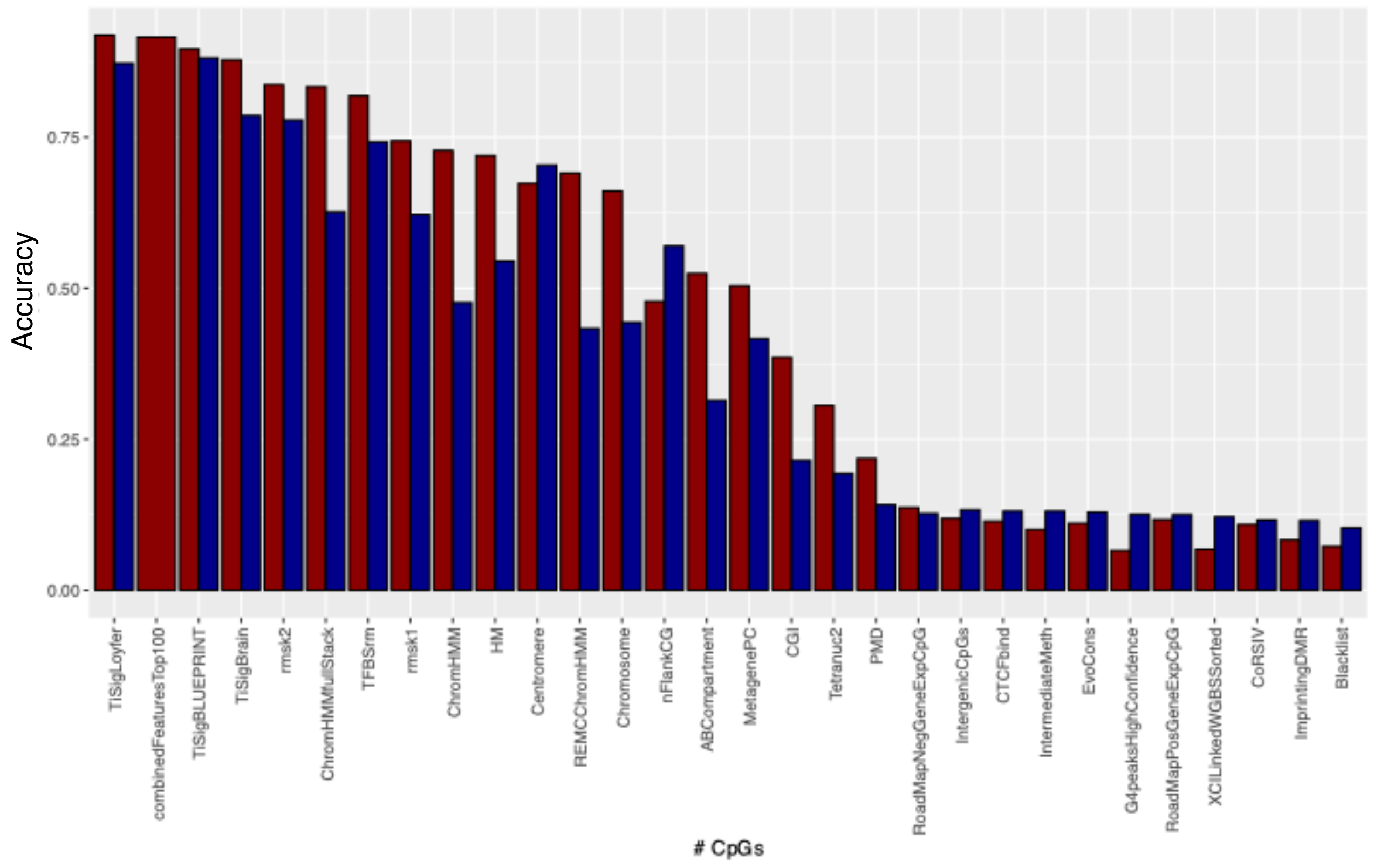

Fig. 2 Aggregated Methylation over Tissue Signatures vs Equal-Sized Random Samples

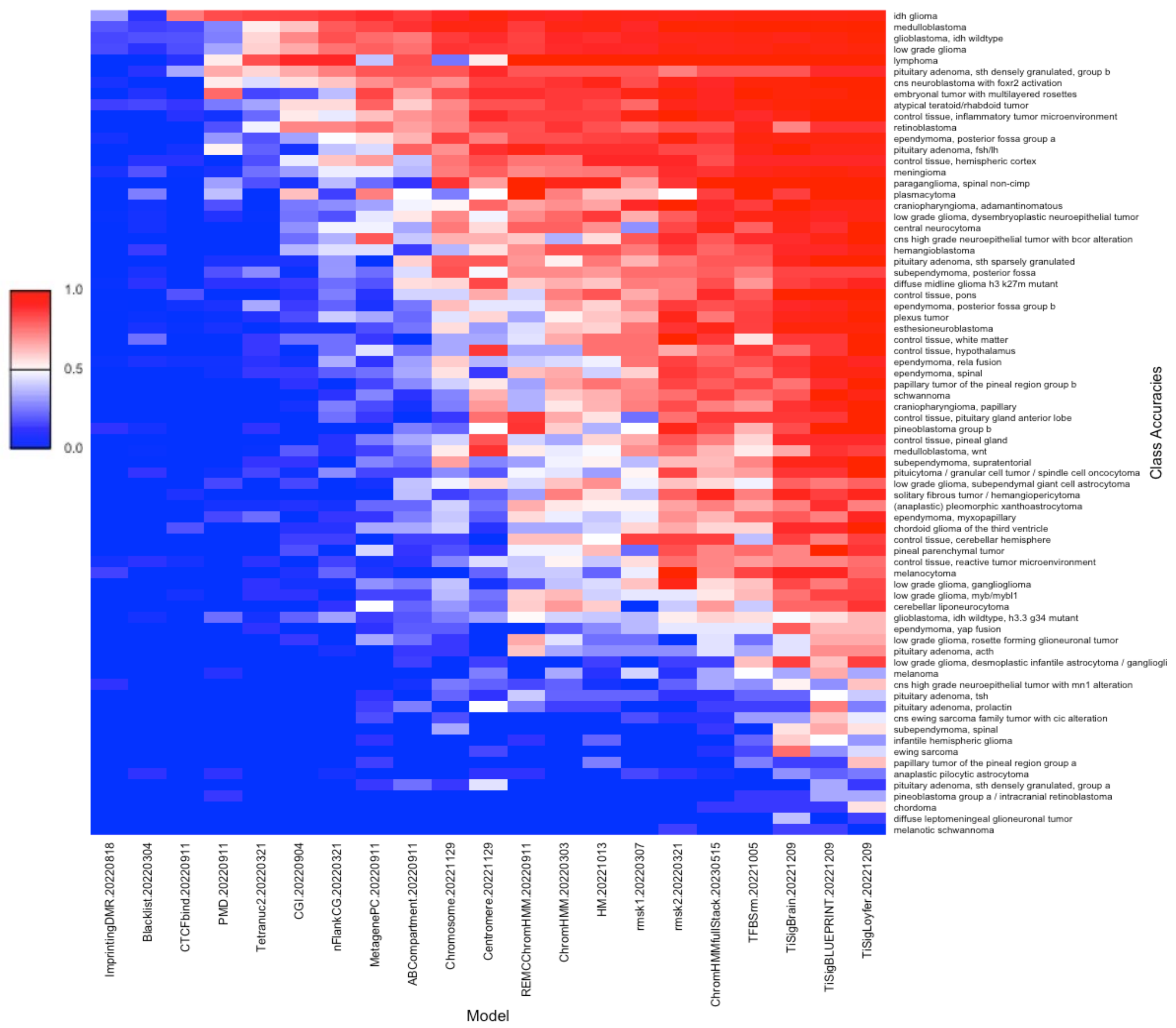

Fig. 3 Class accuracies of aggregated CpG signatures

- Employed heatmap visualization techniques that elucidated the accuracies across various cancer classes to assess the performance of models
- Upon averaging metrics across the aggregated models, we discovered certain neoplastic categories such as Isocitrate dehydrogenase (IDH) gliomas, medulloblastomas, glioblastoma (IDH wildtype), low-grade glioma, lymphoma, pituitary adenoma sth densely granulated group B, and CNS neuroblastoma with FOXR2 activation, demonstrated accuracies exceeding 97%
- Cancer types that exhibited the lowest predictive accuracies were consistently those for which the training sets within the Capper datasets were most limited in size
- Among ChromHMM full stack model, Gap_Artf2 (repeat element) and EnhA6 (brain enhancer) were two of the most important features
- Among the high-performing tissue signature models, the features that proved to be most significant were those that uniquely characterize glial cells

Nonparametric Feature Results

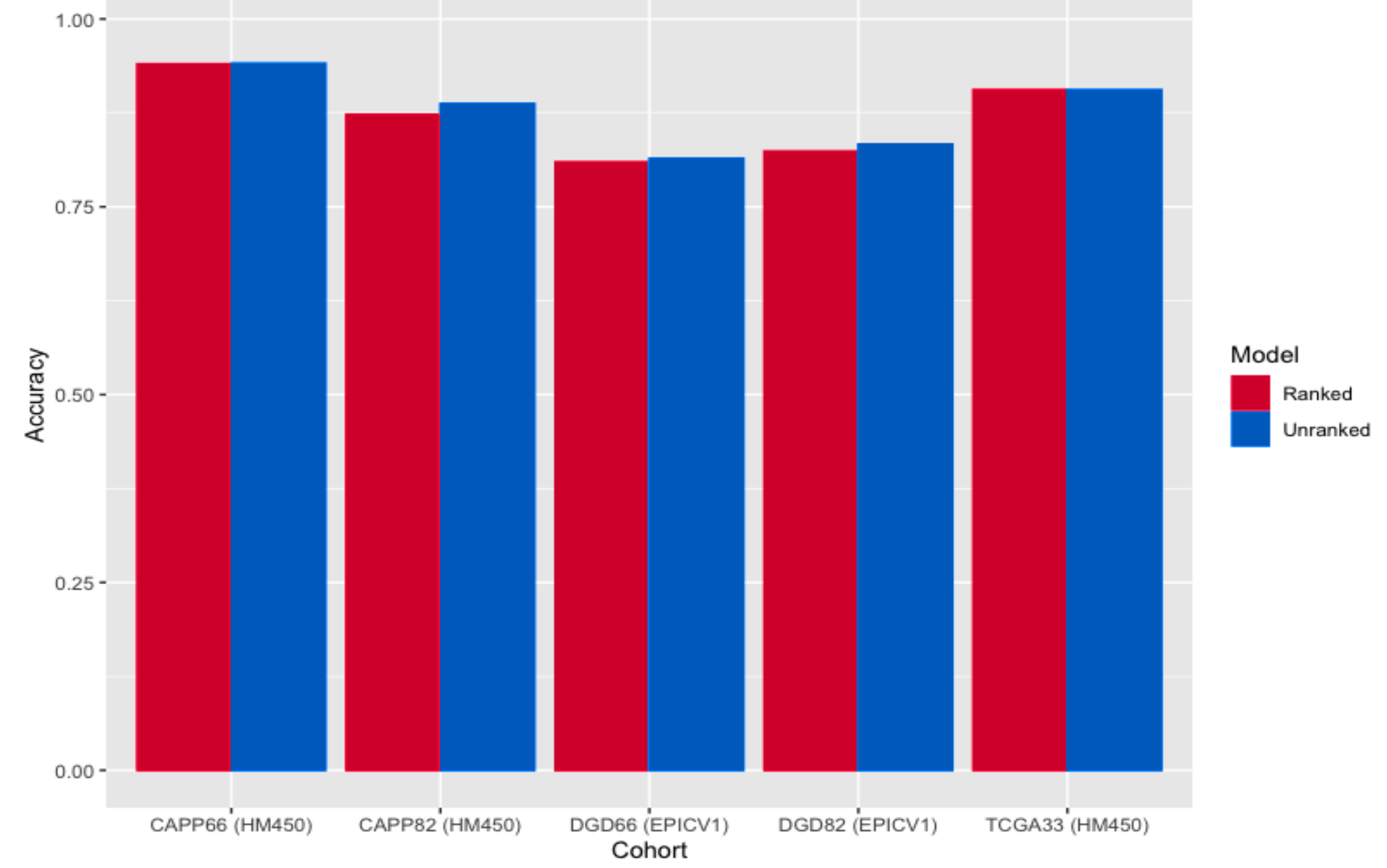

Fig. 4 Comparison of ranked vs. raw methylation data in random forest classifiers

- Trained models on Capper Reference and TCGA data in same manner as unranked, with only difference being ranking
- Tested on Capper Prospective, DGD samples, and TCGA samples
- Upon evaluation across three cohorts, we observed that these models exhibited comparable to marginally inferior performance relative to classifiers trained on unprocessed methylation data
- The classification schema comprising 66 labels yielded more consistent and higher accuracy compared to the 82 labels system

Conclusions

- Feature aggregation based on biological knowledgebases generally improves classifier performances compared to random samples
- Cell signatures & repeat elements are significant factors in predicting cancer types
- Certain cancer subtypes such as IDH tumors are distinctly easier to classify than others across models
- Cancer types with smaller training sets had decreased predictive accuracy, emphasizing the need for larger and more diverse training sets to improve model performance
- Ranking of CpG probes may serve as an effective marker for methylation status, offering potential avenues for further investigation into how rank-order might be incorporated into feature selection or model interpretability processes
- More cohorts with different assay platforms are necessary to further research these feature selection strategies

References & Acknowledgements

- This study was supported by the Penn Undergraduate Research Mentoring Program. Special thanks to Dr. Wanding Zhou for his invaluable support and guidance throughout this research.

- Capper D, Jones DTW, Sill M, et. al. Nature. 2018 Mar 22;555(7697):469-474. doi: 10.1038/nature26000. Epub 2018 Mar 14. PMID: 29539639; PMCID: PMC6093218.
- Liaw A, Wiener M (2002). “Classification and Regression by randomForest.” R News, 2(3), 18-22. https://CRAN.R-project.org/doc/Rnews/.