

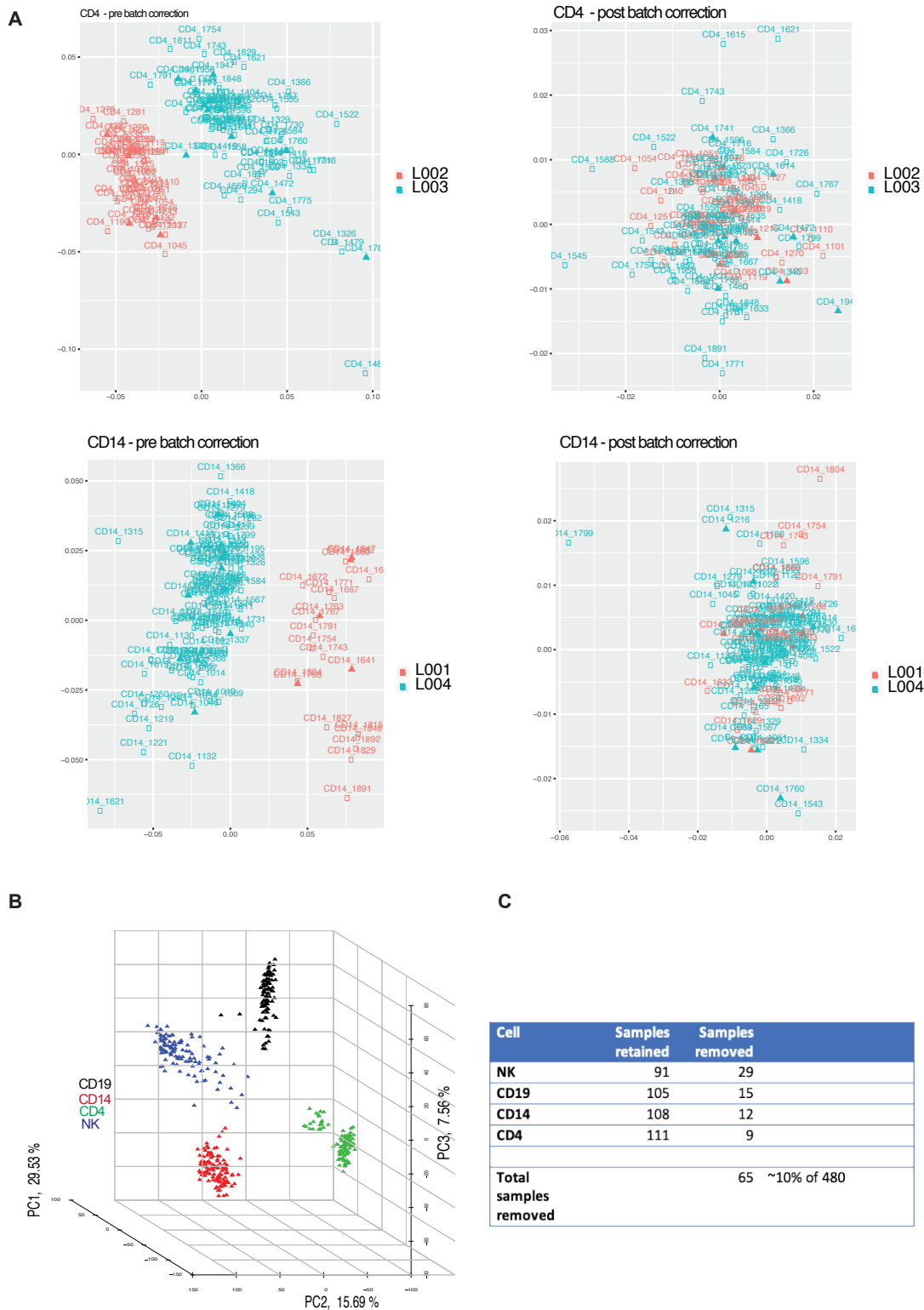
**Transcriptomic analysis of immune cells in a multi-ethnic cohort of systemic lupus erythematosus patients identifies ethnicity- and disease-specific expression signatures**

Gaia Andreoletti<sup>1</sup>, Cristina M. Lanata<sup>2</sup>, Laura Trupin<sup>2</sup>, Ishan Paranjpe<sup>1,3</sup>, Tia S. Jain<sup>1</sup>, Joanne Nititham<sup>2</sup>, Kimberly E. Taylor<sup>2</sup>, Alexis J Combes<sup>3</sup>, Lenka Maliskova<sup>2</sup>, Chun Jimmie Ye<sup>2-4</sup>, Patricia Katz<sup>2</sup>, Maria Dall'Era<sup>2</sup>, Jinoos Yazdany<sup>5</sup>, Lindsey A. Criswell<sup>2</sup> & Marina Sirota<sup>\*1</sup>

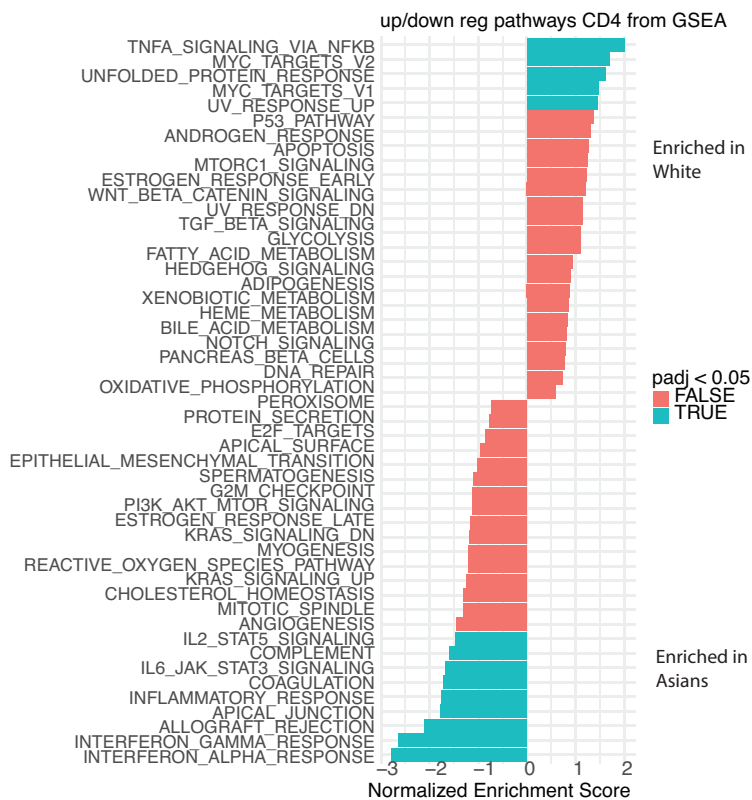
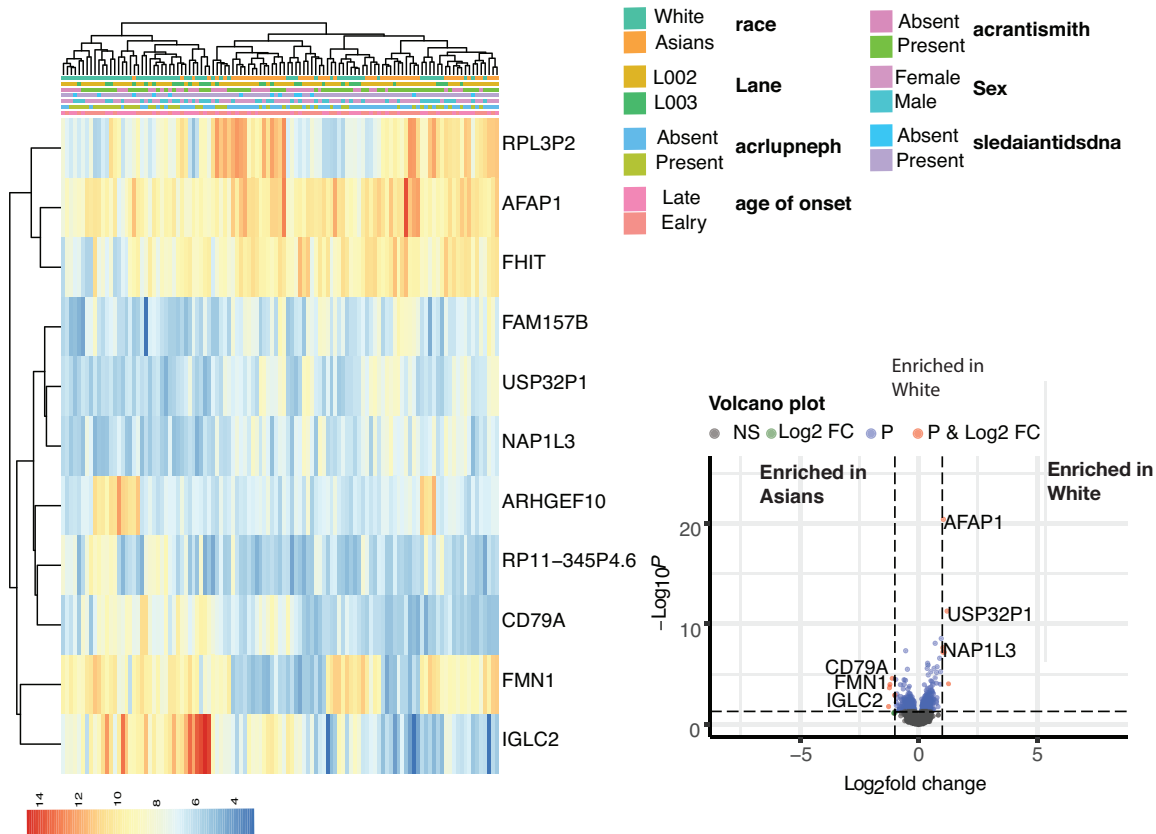
<sup>1</sup>Bakar Computational Health Sciences Institute, University of California, San Francisco, CA, USA. <sup>2</sup>Russell/Engleman Rheumatology Research Center, Department of Medicine, University of California San Francisco, San Francisco, CA, USA. <sup>3</sup> Department of Pathology, University of California, San Francisco, San Francisco, CA 94143, USA; UCSF Immunoprofiler Initiative, University of California, San Francisco, San Francisco, CA 94143, USA. <sup>4</sup>Department of Epidemiology and Biostatistics, University of California, San Francisco, San Francisco, CA, USA, <sup>5</sup>Division of Rheumatology, Department of Medicine at UCSF

\*Corresponding author: Marina Sirota [marina.sirota@ucsf.edu](mailto:marina.sirota@ucsf.edu), ORCID: 0000-0002-7246-6083

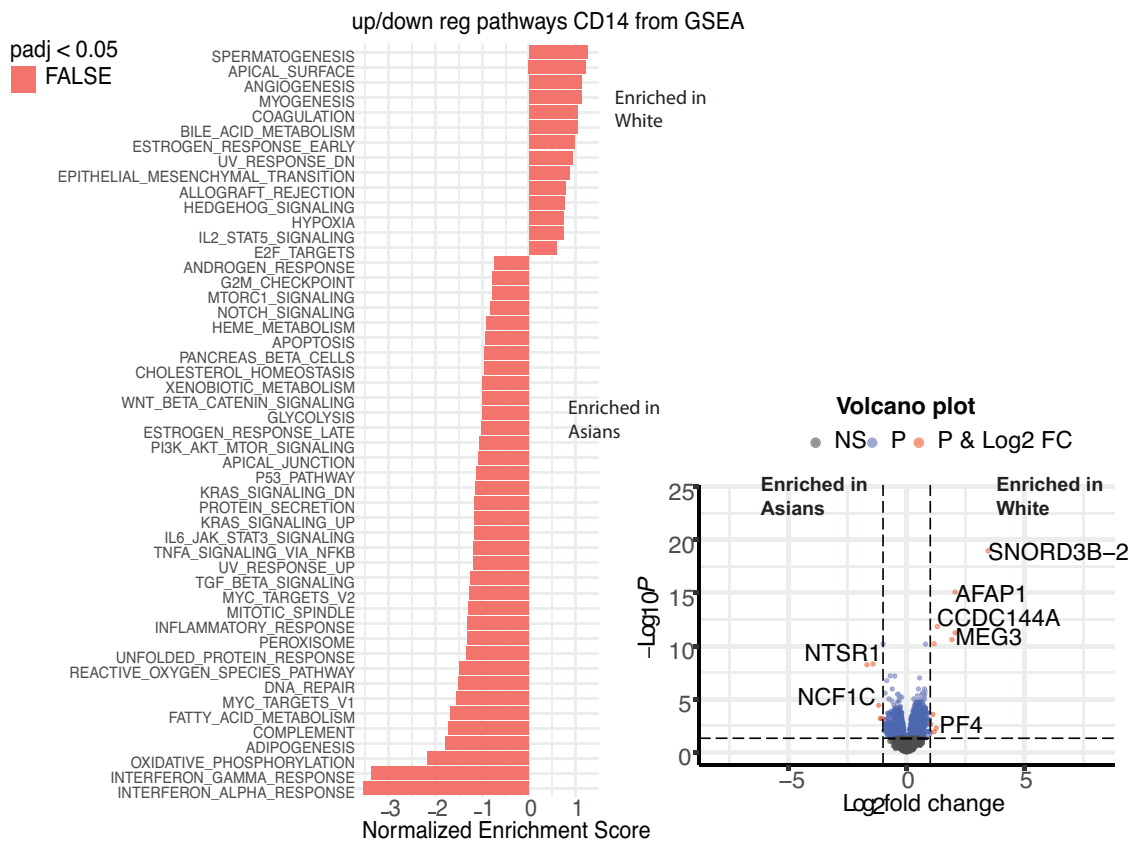
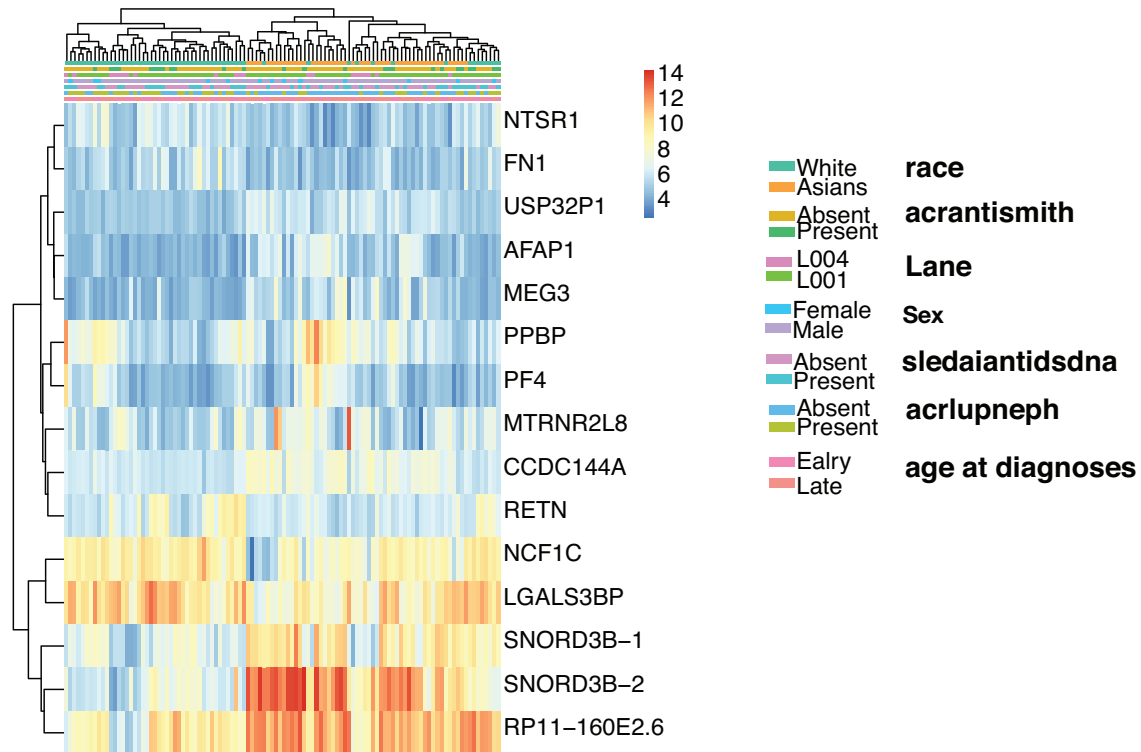
Supplementary Figures:



**Supplementary Figure 1: PCA visualization and quality control.** A) Data before and after batch correction using limma. Batch effect was observed only in CD4<sup>+</sup> T cells and CD14<sup>+</sup> monocytes. K-means clustering after batch correction. B) Principal component analysis (PCA) on the batch corrected data. C) Table detailing the number of samples removed per immune-cell type after QC.

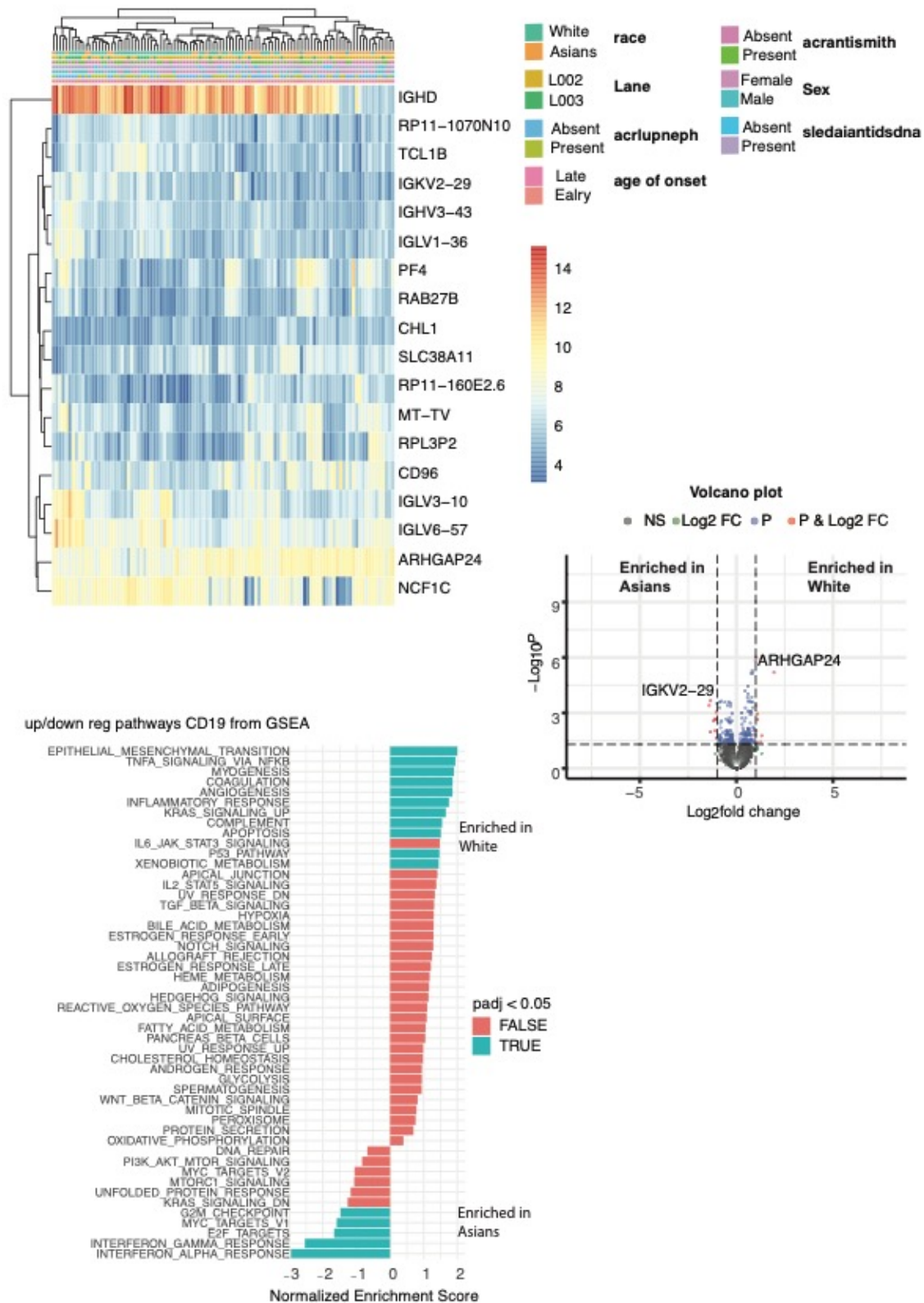


**Supplementary Figure 2: Differential Expression by race for CD4<sup>+</sup> T cell.** Heatmap, volcano plot and gene set enrichment analyses (GSEA) on the significant genes (FDR < 0.05 and abs(logFC) ≥ 1).



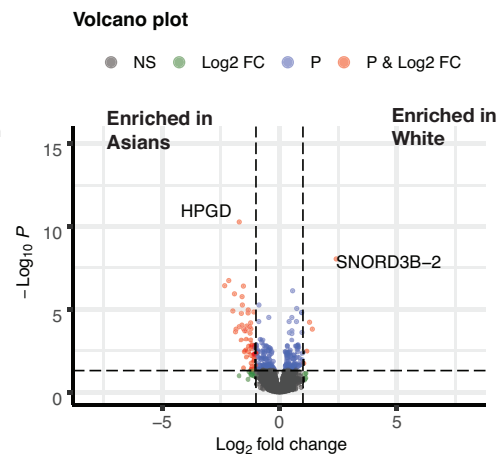
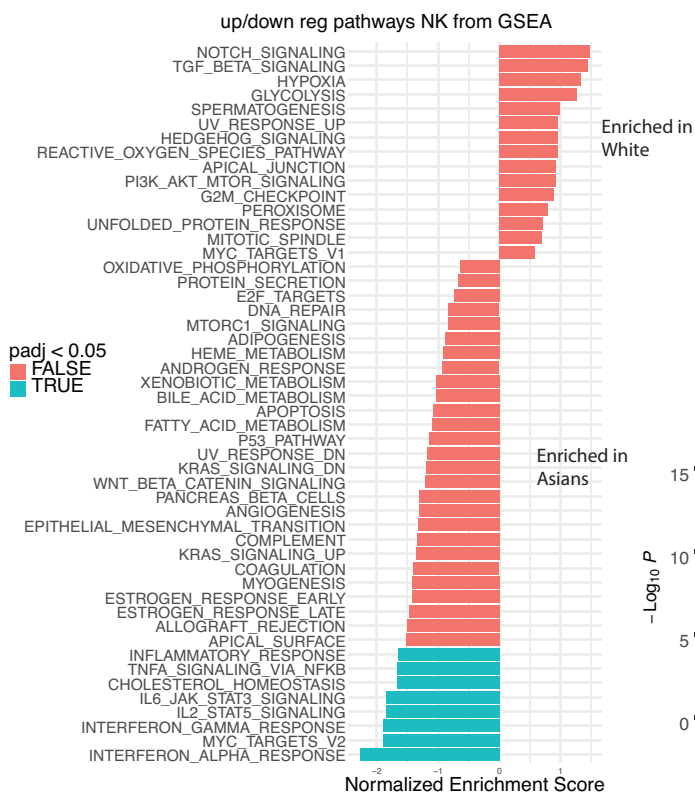
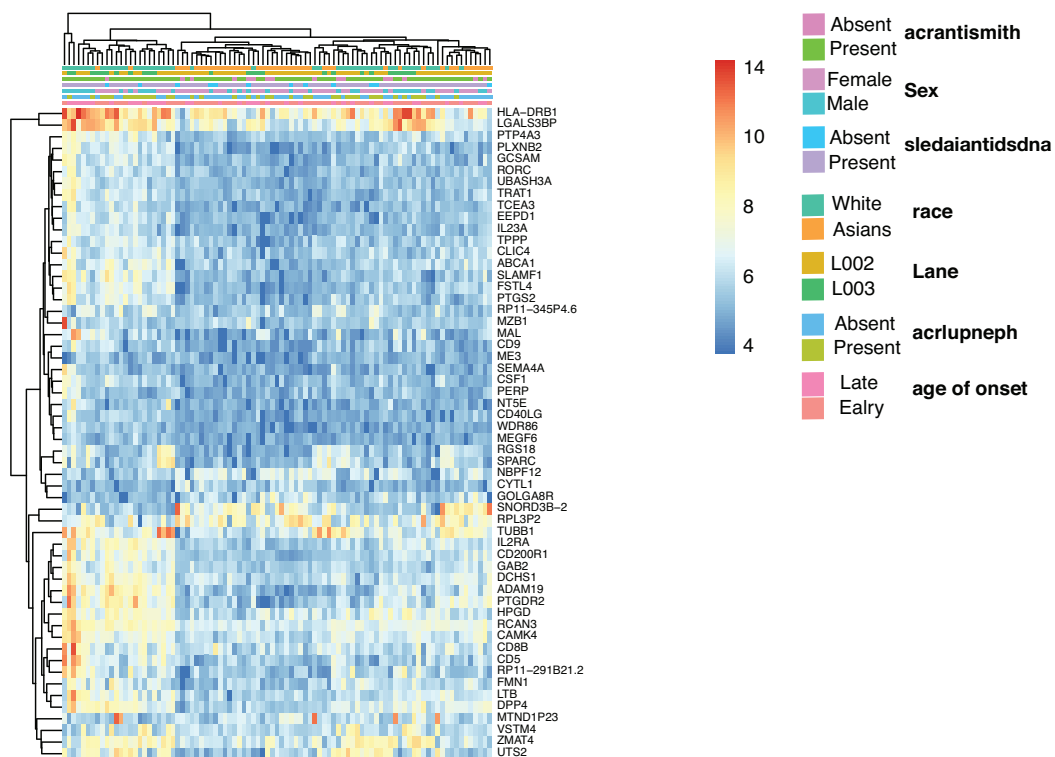
**Supplementary Figure 3: Differential Expression by race for CD14<sup>+</sup> monocytes.**

Heatmap, volcano plot and gene set enrichment analyses (GSEA) on the significant genes (FDR < 0.05 and  $\text{abs}(\log\text{FC}) \geq 1$ ).



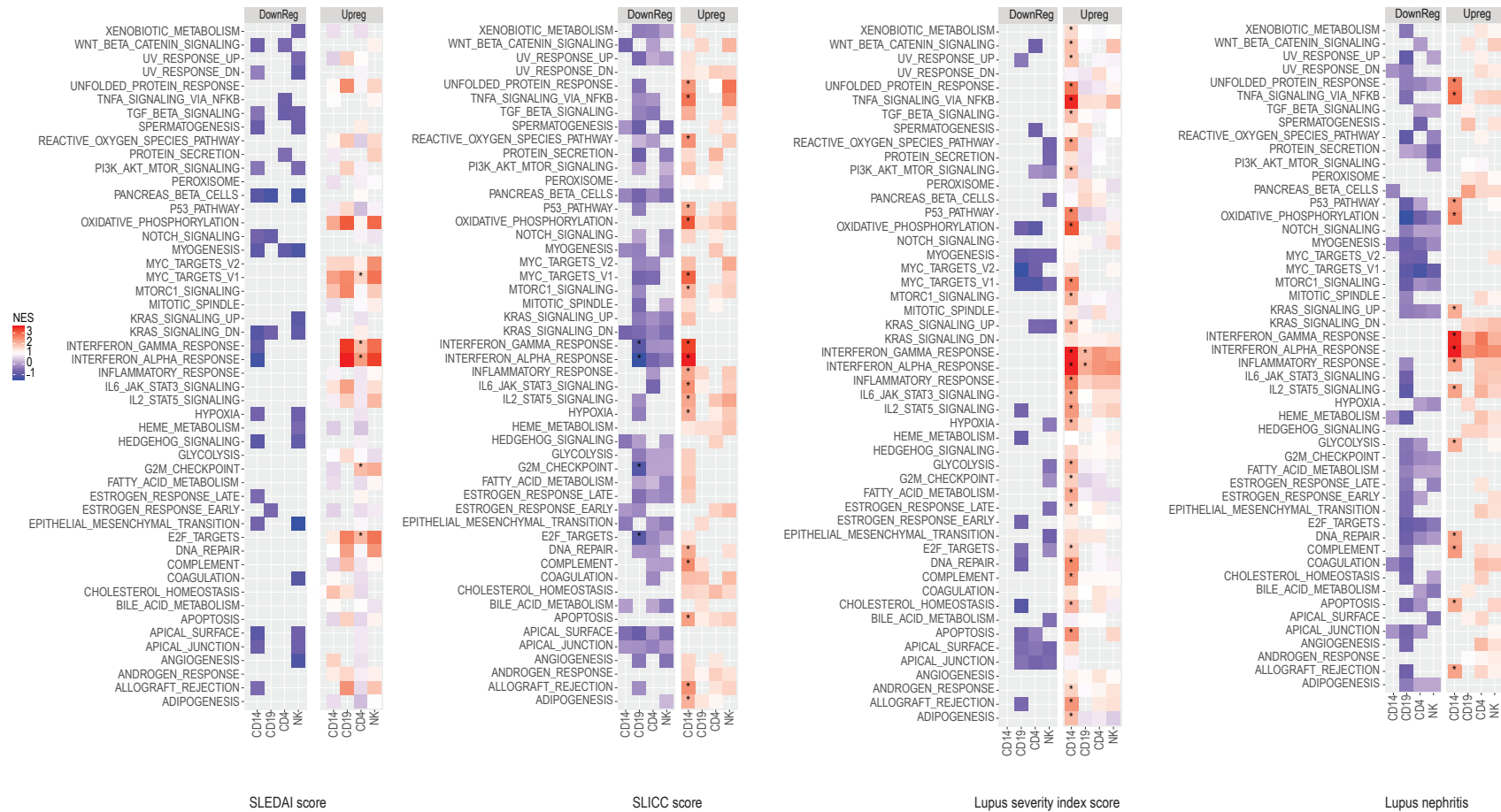
Supplementary Figure 4: Differential Expression by race for CD19 B cell. Heatmap,

volcano plot and gene set enrichment analyses (GSEA) on the significant genes ( $\text{FDR} < 0.05$  and  $\text{abs}(\log\text{FC}) \geq 1$ ).



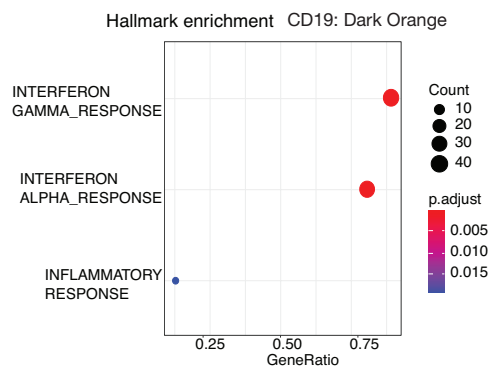
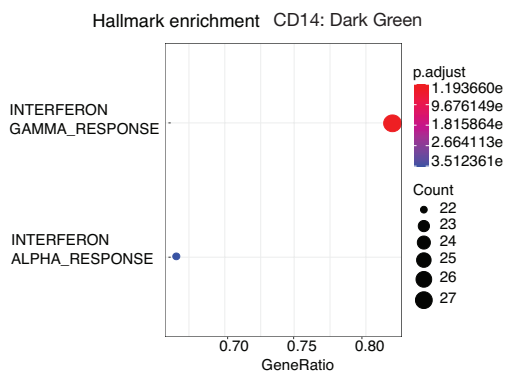
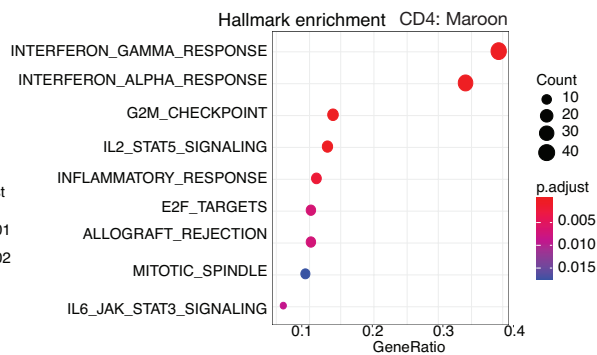
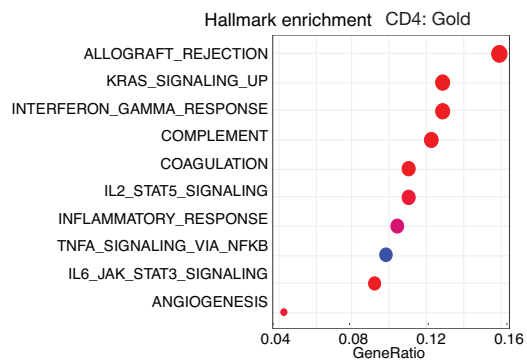
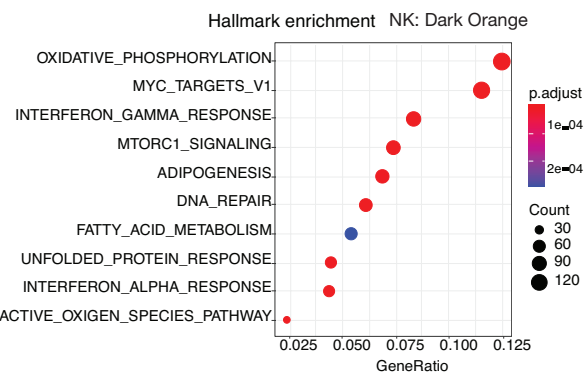
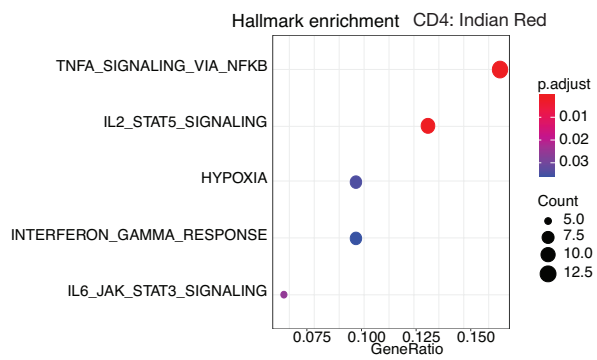
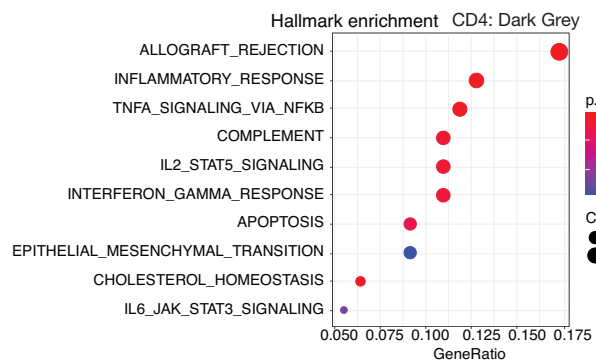
**Supplementary Figure 5: Differential Expression by race for NK cells.** Heatmap,

volcano plot and gene set enrichment analyses (GSEA) on the significant genes ( $\text{FDR} < 0.05$  and  $\text{abs}(\log\text{FC}) \geq 1$ ).

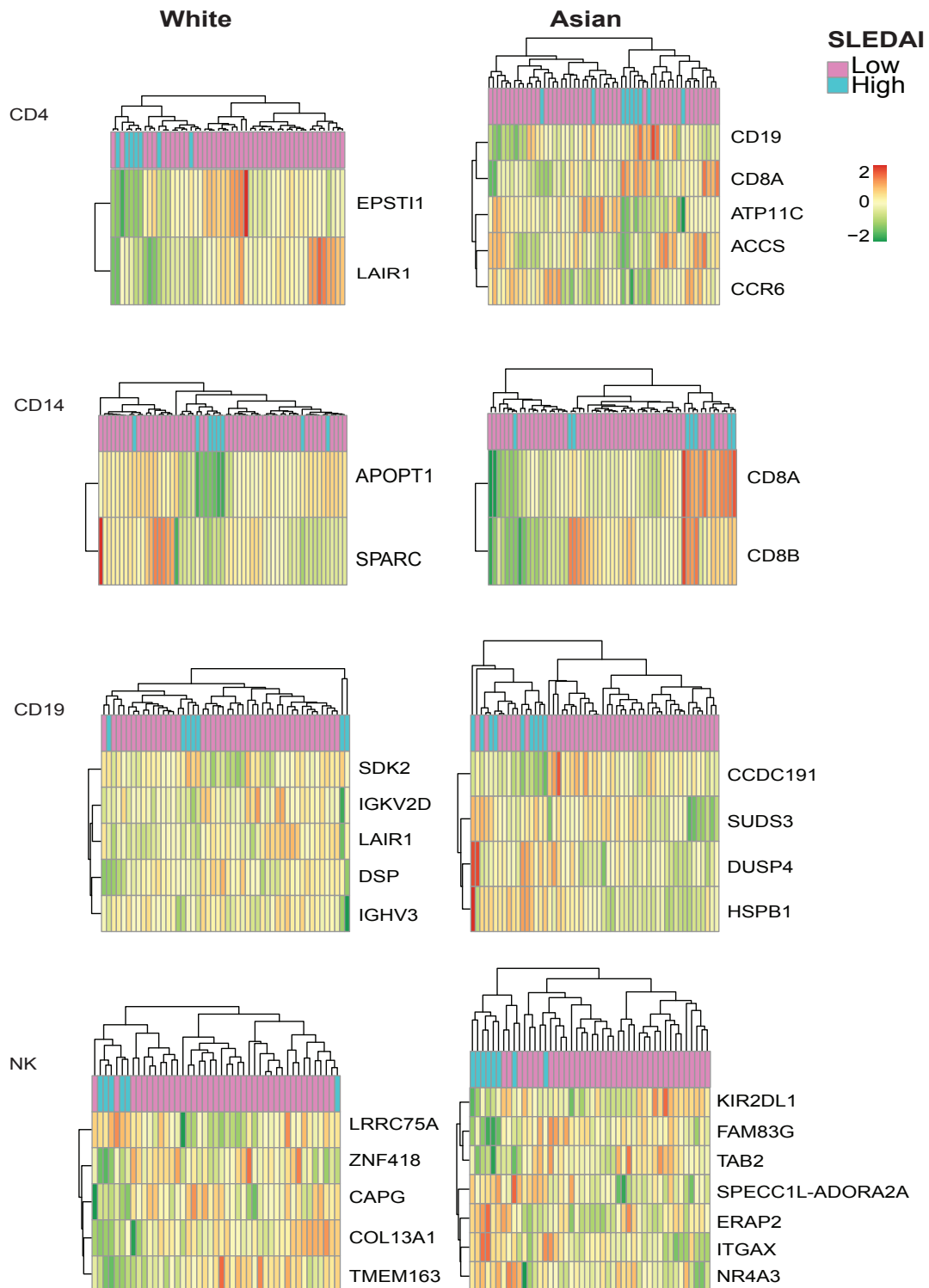


**Supplementary Figure 6: Heatmaps displaying pathways analyses across clinical features.** Pathway Analysis using GSEA

from DE conducted on other clinical criteria Sledai score, slicc score, lupus severity index, and lupus nephritis

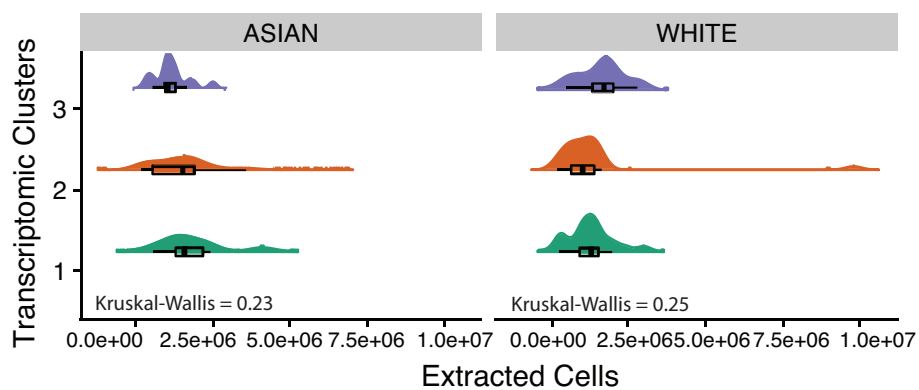


**Supplementary Figure 7: Functional enrichment (Hallmark pathways) of co-expressed modules enriched for interferon pathways**

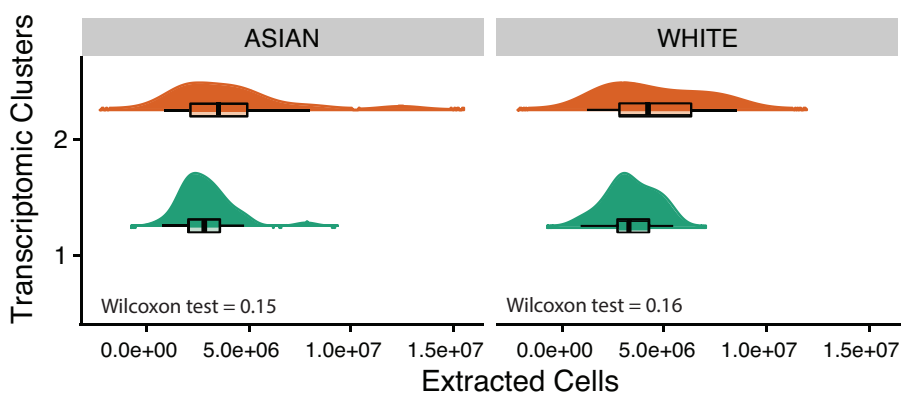


**Supplementary Figure 8: Machine learning heatmaps.** Clustergrams by cell type for each ethnic group by using the expression values of the top predictors.

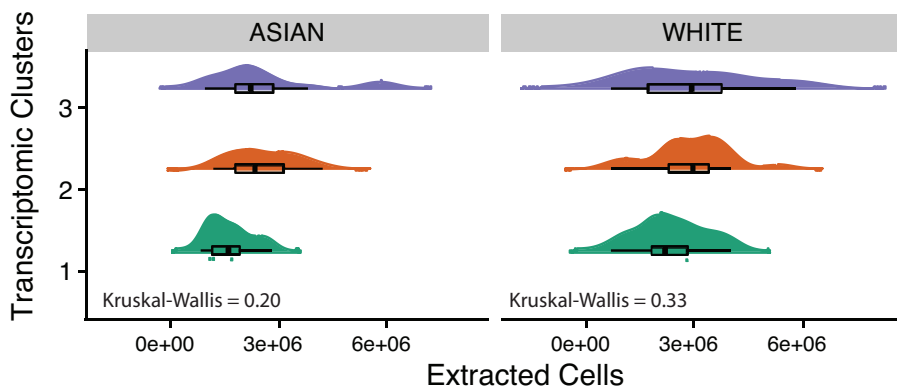
## CD14



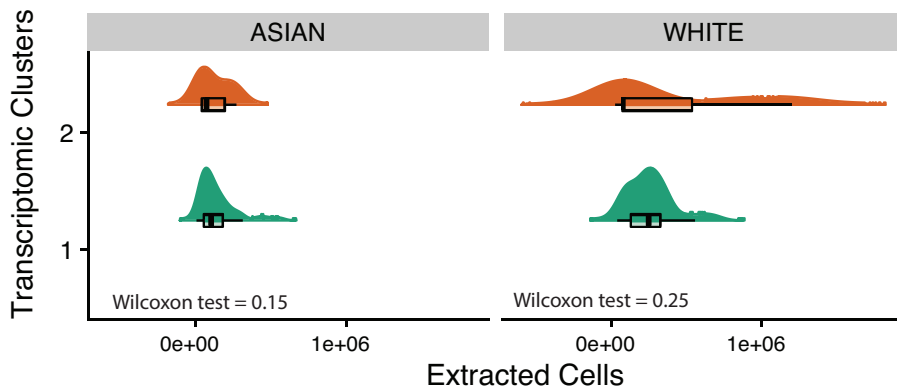
## CD19



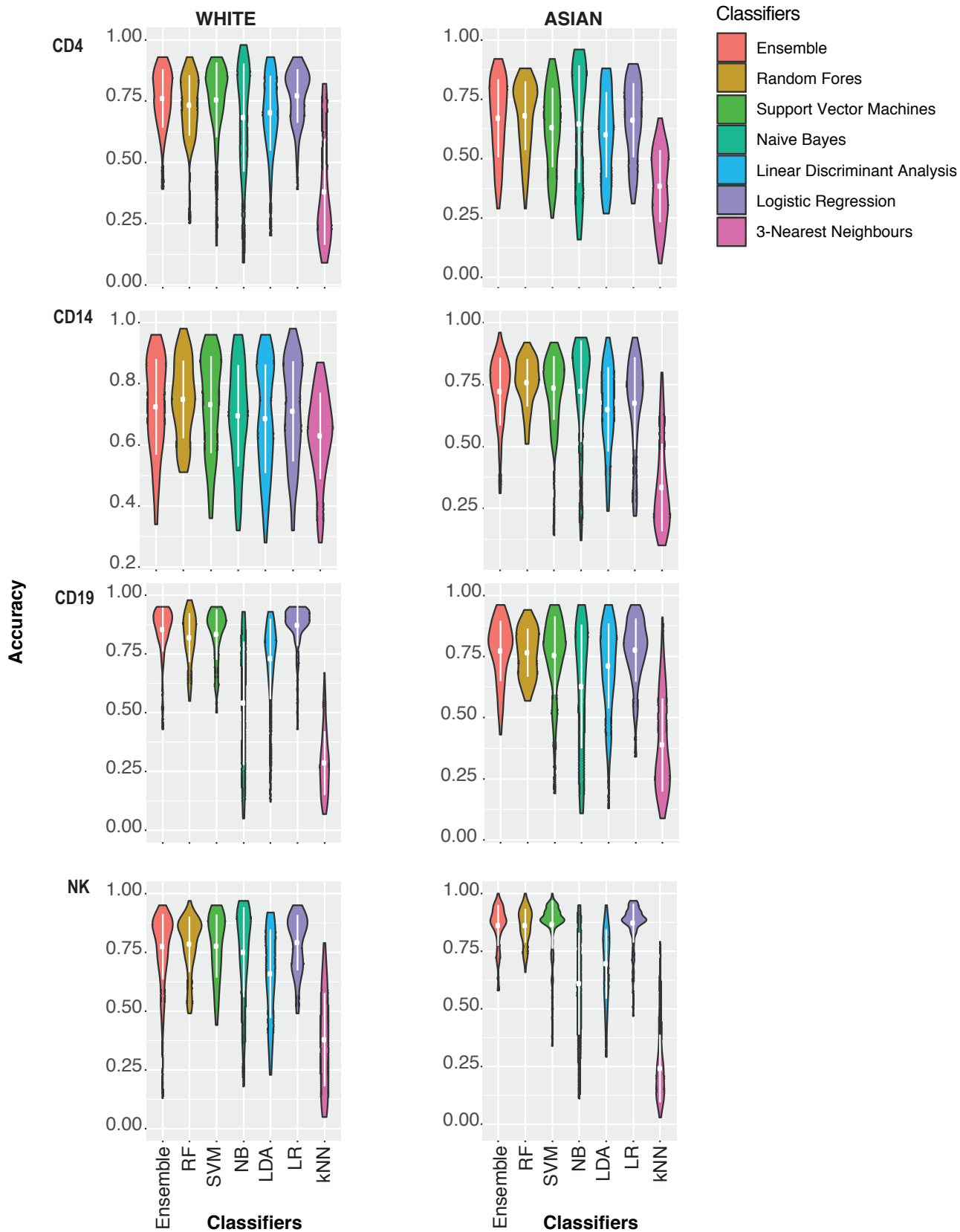
## CD4



## NK

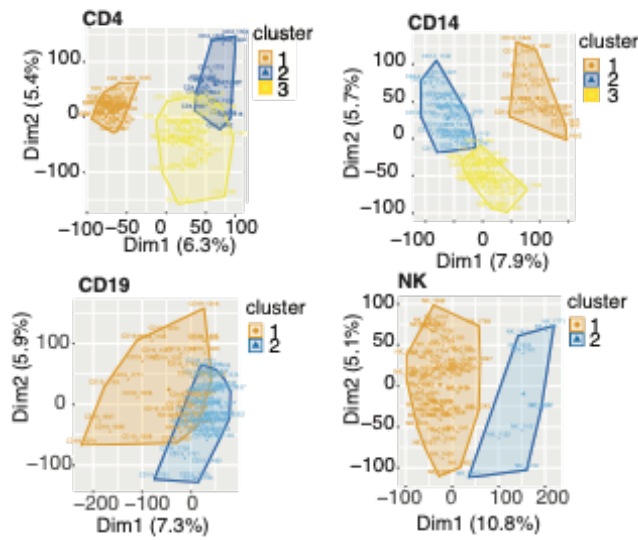


**Supplementary Figure 9: Cell counts distribution visualization.** Cell count distribution across the identified transcriptomic clusters for Asian and White Cohorts. None of the distributions are statistically significantly different between groups of interest.

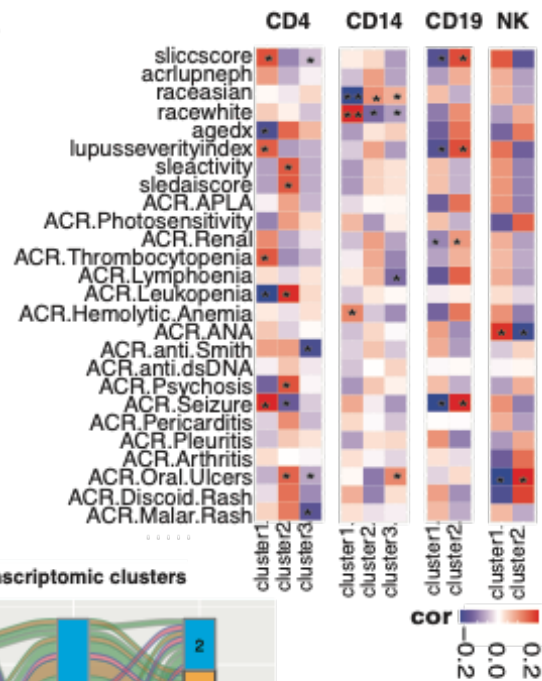


**Supplementary Figure 10: ML accuracies Comparison.** The violin plot highlights the classification accuracy of each classifier, computed at each iteration for each cell line within White and Asian cohorts. Random Forest is the top performer across all and therefore is presented in the main text.

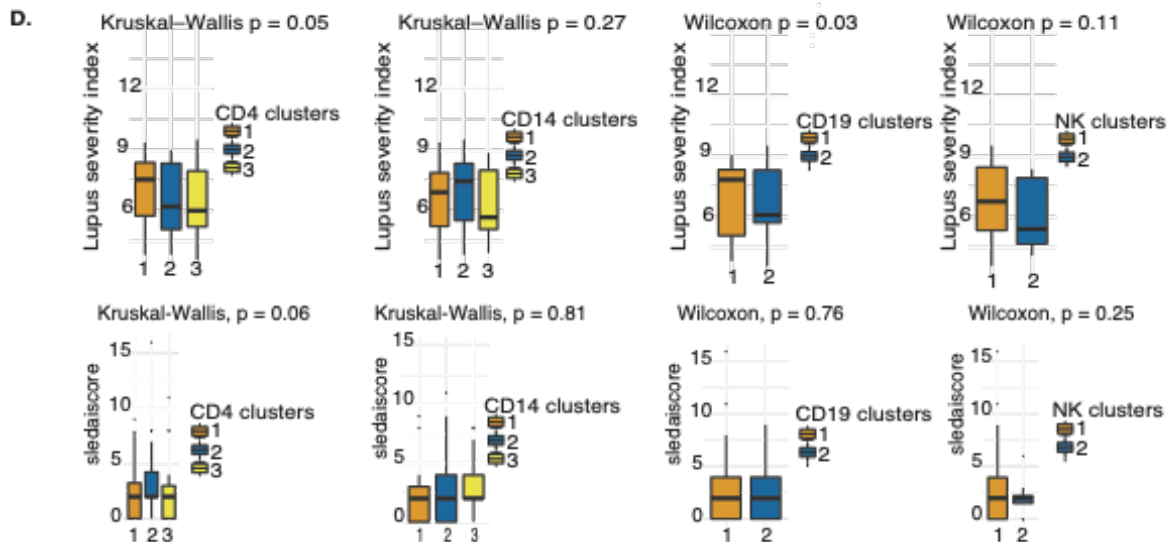
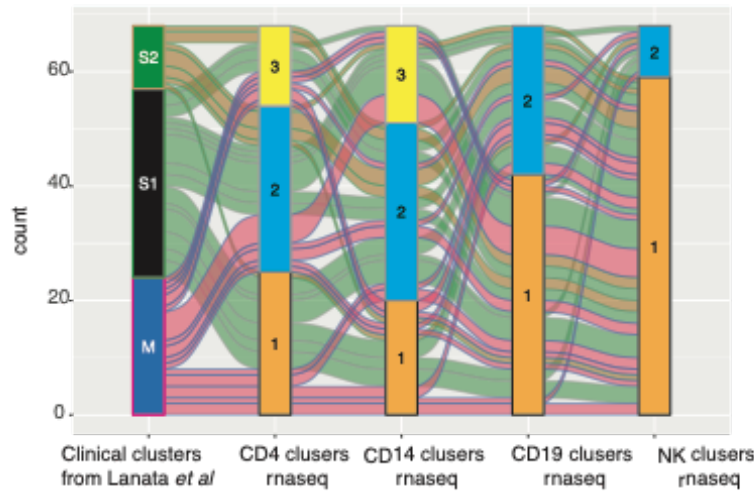
### A. K-means clustering



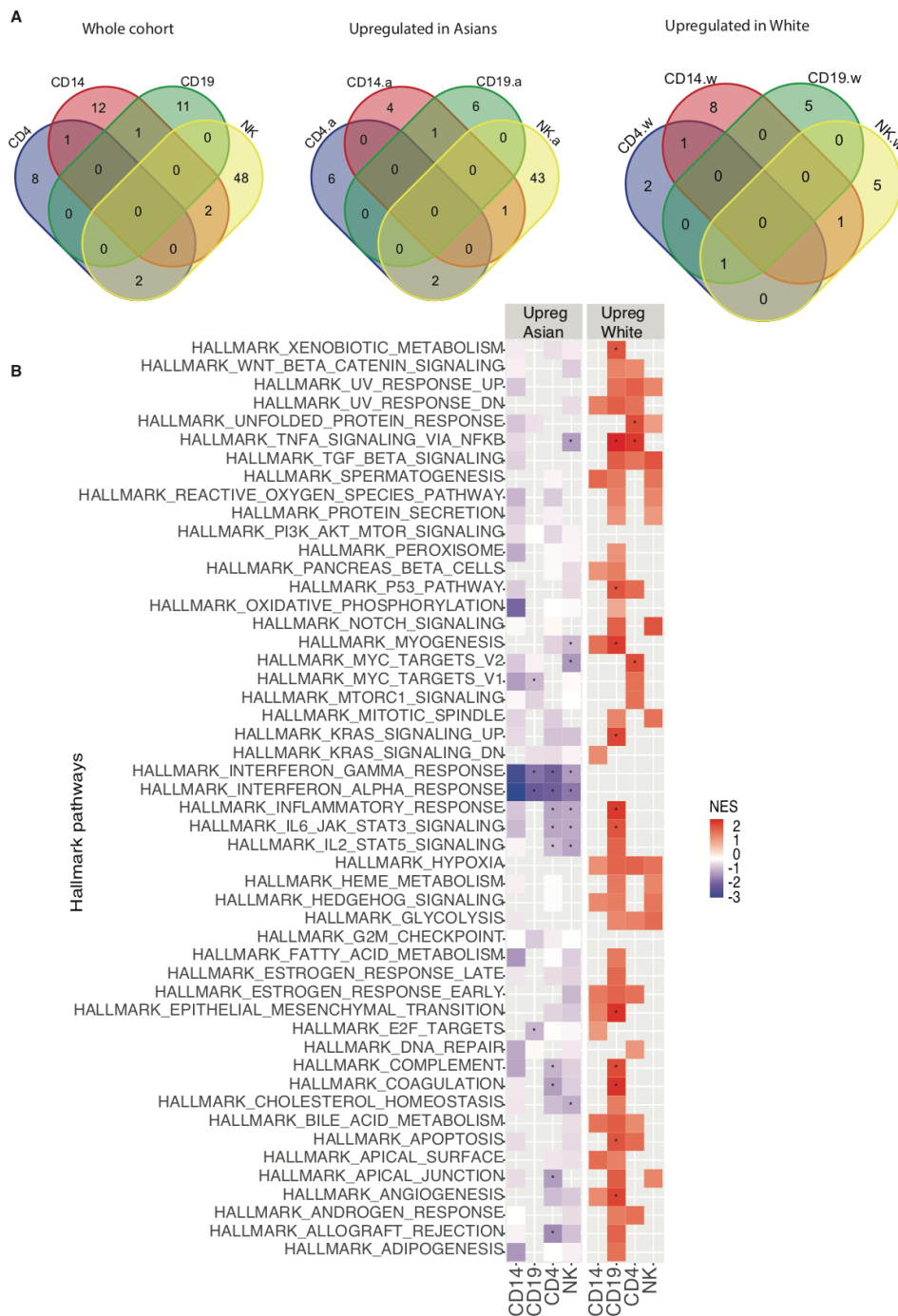
### B. Correlation analyses



### C. Alluvium plot - clinical and transcriptomic clusters



**Supplementary Figure 11: Sensitivity analyses without patients with ACR criteria < 4 and a confirmed diagnoses of lupus nephritis (n=21):** A) K-means clustering on the CD14<sup>+</sup> monocytes, CD4<sup>+</sup> T-cells, B cells and NK cells. B) Association analyses between clinical parameters and K-means s clustering. Red – positive association, blue – negative association. Number of stars indicate the level of significance. C) Alluvium plot visualizing the distribution of the samples according to different clusters. D) Distribution of lupus severity index and SLEDAI score across clusters with p-value computed using non-parametric tests. Findings between the sensitivity analysis and the main analysis are consistent.



**Supplementary Figure 12: Sensitivity analyses without patients with ACR criteria < 4**

**and a confirmed diagnoses of lupus nephritis: A) Identification of the common DE genes**

( $\text{padj} < 0.05$  &  $\text{abs log}_2\text{FC} > 1$ ) for White vs Asian cohorts across the different cell types. B)

Pathway enrichment analyses for the associated genes of White vs Asian cohorts. Findings between the sensitivity analysis and the main analysis are consistent.