Analysis of Gene Expression Patterns in Multiple Sclerosis

Insights from PBMC and CSF Cell Populations

Gabriele Benanti
g.benanti@studenti.unipi.it
Roll number: 550552

Giulia Ghisolfi
g.ghisolfi@studenti.unipi.it
Roll number: 664222

Alessandro Stefanelli
a.stefanelli3@studenti.unipi.it
Roll number: 686084

Master Degree in Computer in Science, University of Pisa
Computation Healt Laboratory course (755AA), Prof. Corrado Priami
Academic Year: 2023-2024

Outline

Introduction

Multiple Sclerosis: Overview

- **Multiple sclerosis** (MS) is a chronic inflammatory, neurodegenerative, and autoimmune disease.
- Primarily affects young adults, more frequently women.
- Involves complex gene-environment interactions.
- Despite extensive research, the underlying causes and mechanisms remain elusive.

Introduction to the Problem

- Studying MS is challenging due to lack of comprehensive data.
- Most studies have small, non-representative sample sizes.
- MS is a heterogeneous disease with a variable and unpredictable course.
- Our project aims to **identify biomarkers** related to **specific cell types** via **differential expression analysis**.
- Major obstacle: Scarcity of usable datasets.

Datasets

Datasets (1)

- Data from PBMC and CSF Cell Populations.

- **Integration of data** from multiple studies to **enhance analysis**:
  - Complemented our gene expression analysis with data from GSE173787 (transcriptomic analysis of CD19+ B cells).
  - 547 genes differentially expressed between Multiple Sclerosis (MS) patients and Healthy Controls (HC).

Datasets (2)

	MS			HC		
	patients	**PBMC**	**CSF**	**patients**	**PBMC**	**CSF**
GSE239626	10	72,317	-	-	-	-
GSE194078	3	17,083	17,133	9	44,398	31,277
GSE138266	6	25,553	19,306	3	15,604	5,641
Total	19	114,953	36,389	12	60,002	36,918

Table 1: Summary of the **number of cells** we have **available** for differential expression analysis and the **number of patients from whom they were extracted**. PBMC refers to Peripheral Blood Mononuclear Cells, and CSF refers to Cerebrospinal Fluid. MS stands for Multiple Sclerosis, and HC stands for Healthy Control.

Cell Type Extraction from Antibody Clustering

Cell Type Extraction from Antibody Clustering (1)

- Data dimensionality reduced using PCA.
- Utilized **Leiden algorithm** for **clustering** based **on antibody features**.

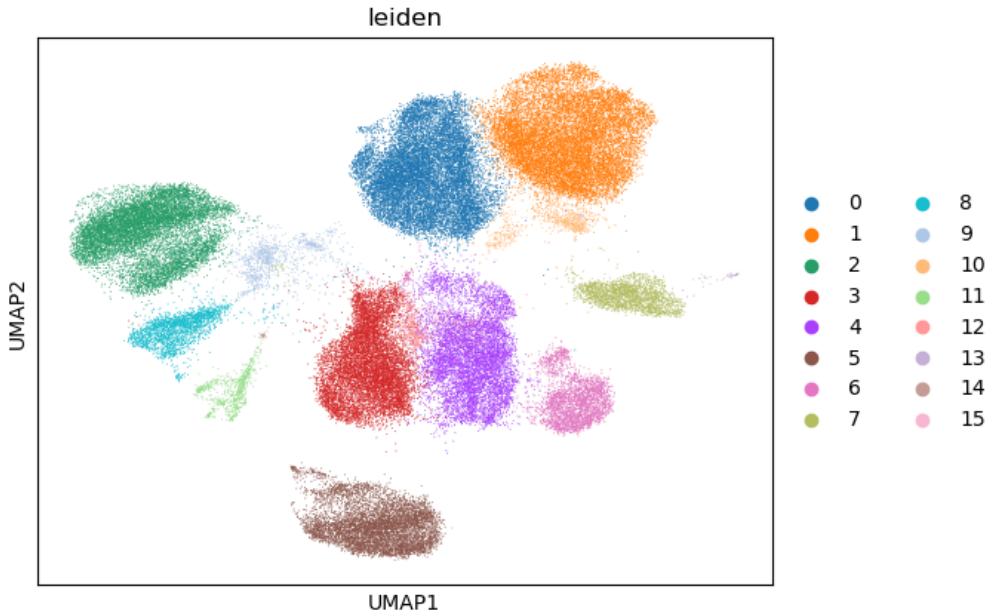
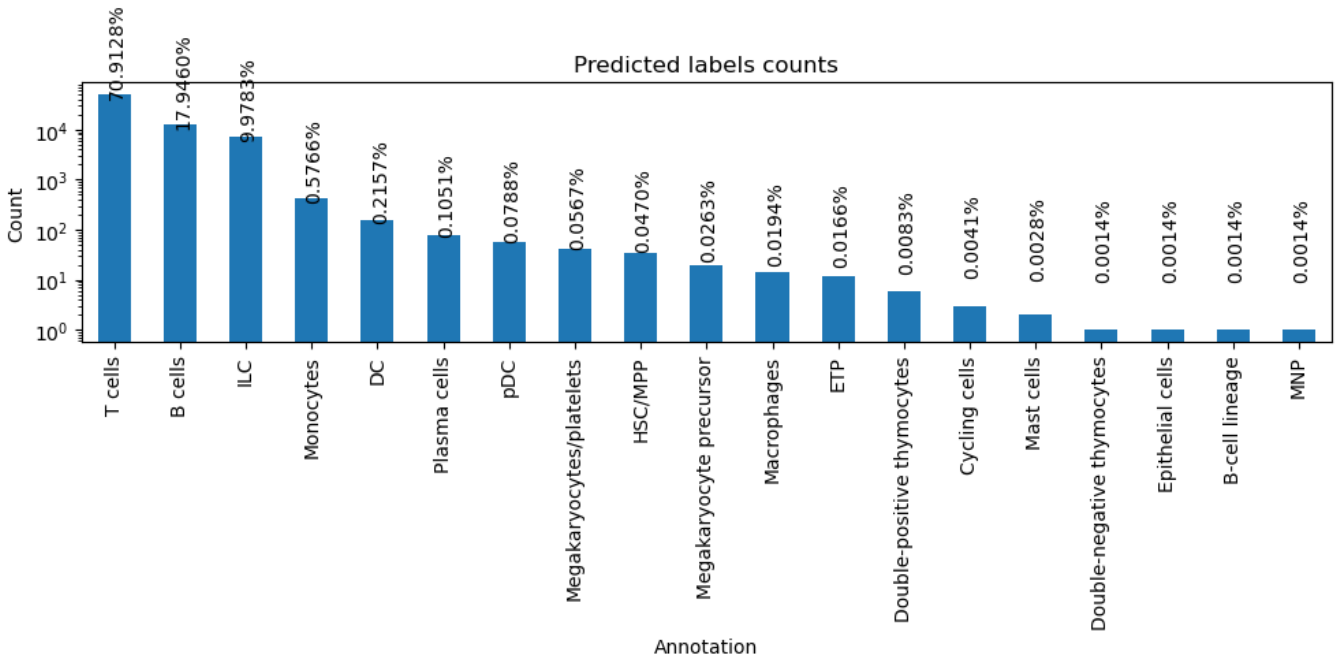


Figure 1: Figure showing the distribution of clusters in the two principal component UMAP space. The clusters are defined by the Leiden algorithm applied to antibody features.

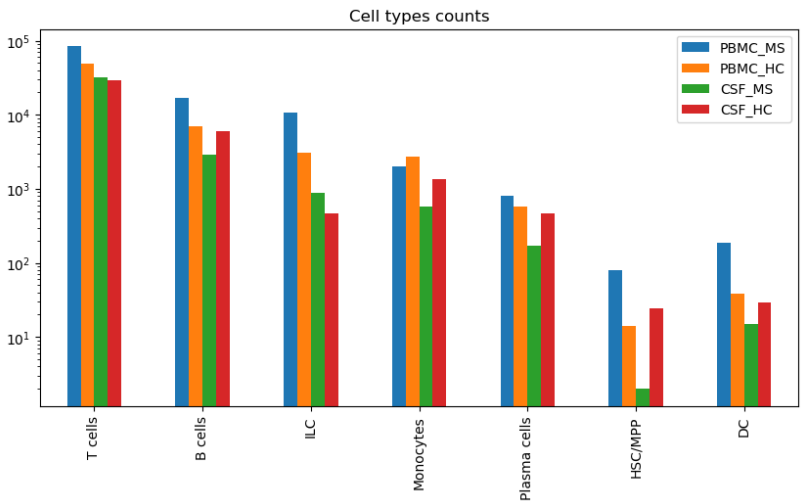
Cell Type Extraction from Antibody Clustering (2)

- `Immune_All_High` pre-trained model from `CellTypist` was used for annotate data.
- **Incorporating clustering labels** and gene counts from GSE239626 into the model **improved annotation accuracy**.
- Most common cell types identified: T cells (70.91%), B cells (17.95%) and ILC (9.98%).



Cell Type Labels Transfer

- GSE194078 and GSE138266 **dataset labelled** using `ScanPy` built-in function **Ingest**, transferring cell type labels from GSE239628.
- **BBKNN algorithm** applied to the previous two datasets, concatenated with the reference dataset.



- Number of cells reported in MS samples is generally higher than in HC samples.
- Pronounced difference in T cells in PBMC cells.

Differential Expression Analysis

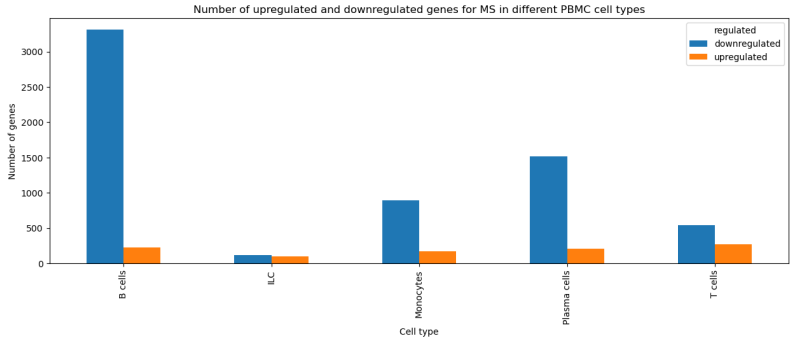
Differential Expression Analysis

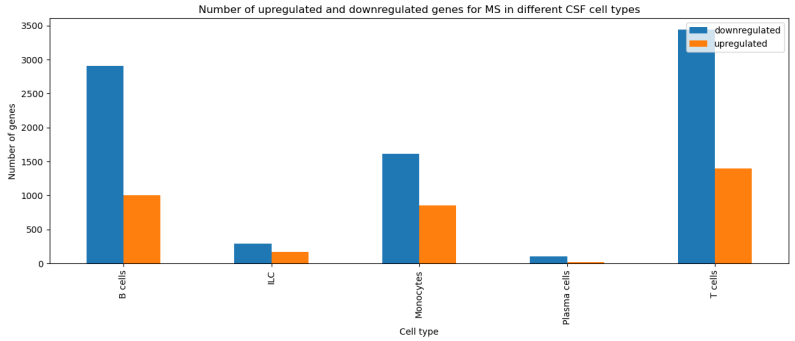
- Conducted **differential expression analysis** comparing MS patients and healthy controls.
- Focus on PBMC and CSF cell populations.
- Analyzed gene expression profiles across **T cells**, **B cells**, **Monocytes**, **Plasma cells**, and **ILC**.

- Utilized **Wilcoxon rank-sum test** from `rank_genes_groups` method implemented in the `ScanPy` library.
- Applied thresholds for significance: **adjusted p-value of** 10^{-30} and **log fold change** > 0.5.

Results

Differential Expression Results (1)





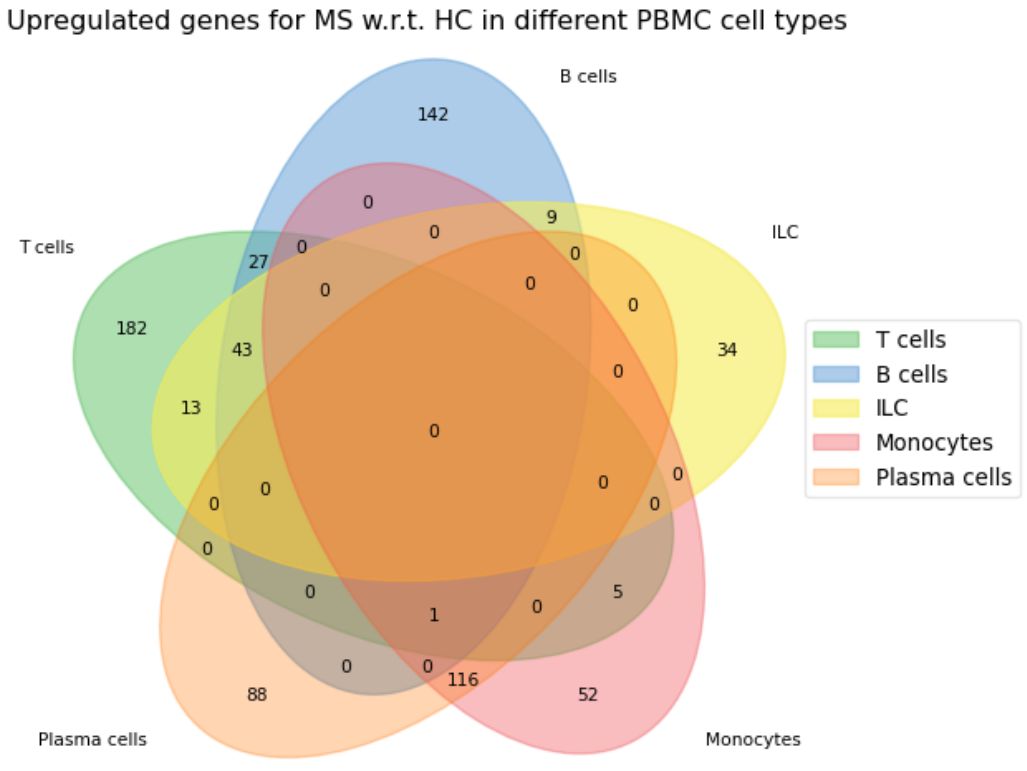
- Significant **differences** in **gene expression profiles** were identified.
- Notably **higher number** of **downregulated genes** in **MS patients**, compared to healthy controls, especially in CSF.
- **T cells** and **B cells**, especially in the **CSF**, exhibit **extensive differences in gene expression**, both in terms of upregulation and downregulation.

Differential Expression Results (2)

	PBMC		CSF	
Cell Type	**Downregulated**	**Upregulated**	**Downregulated**	**Upregulated**
B cells	**3309**	222	2907	1008
ILC	117	99	287	165
Monocytes	891	174	1610	853
Plasma cells	1513	205	108	19
T cells	543	**271**	**3436**	**1395**

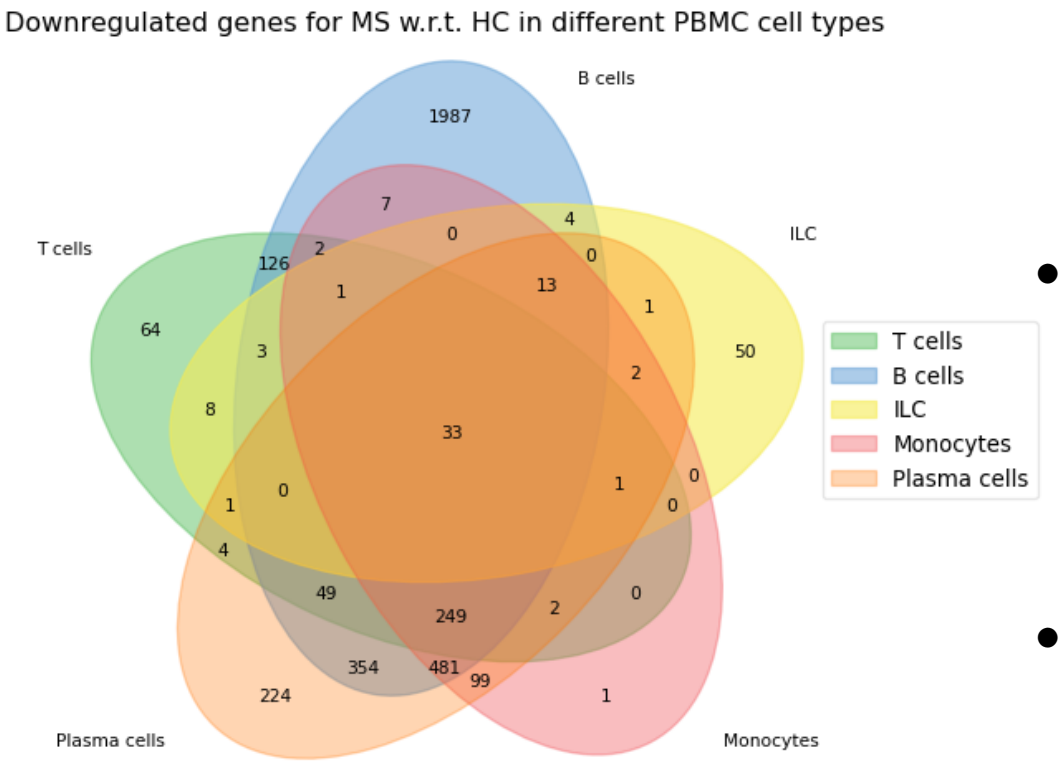
Table 2: Number of upregulated and downregulated genes for each cell type in PBMC and CSF samples.

Differential Expression Results: Upregulated in PBMCs



- **T cells** exhibit the **highest number** of **unique upregulated genes** (182 genes).
- B cells showing a count of 142 unique upregulated genes.
- 71 **upregulated genes commonly** expressed by **T cells** and **B cells**.
- **Plasma cells** and **Monocytes** shared **upregulated pathways**.
- No gene is found to be common across all five PBMC cell types: specific **upregulation profiles depending on the cell type**.

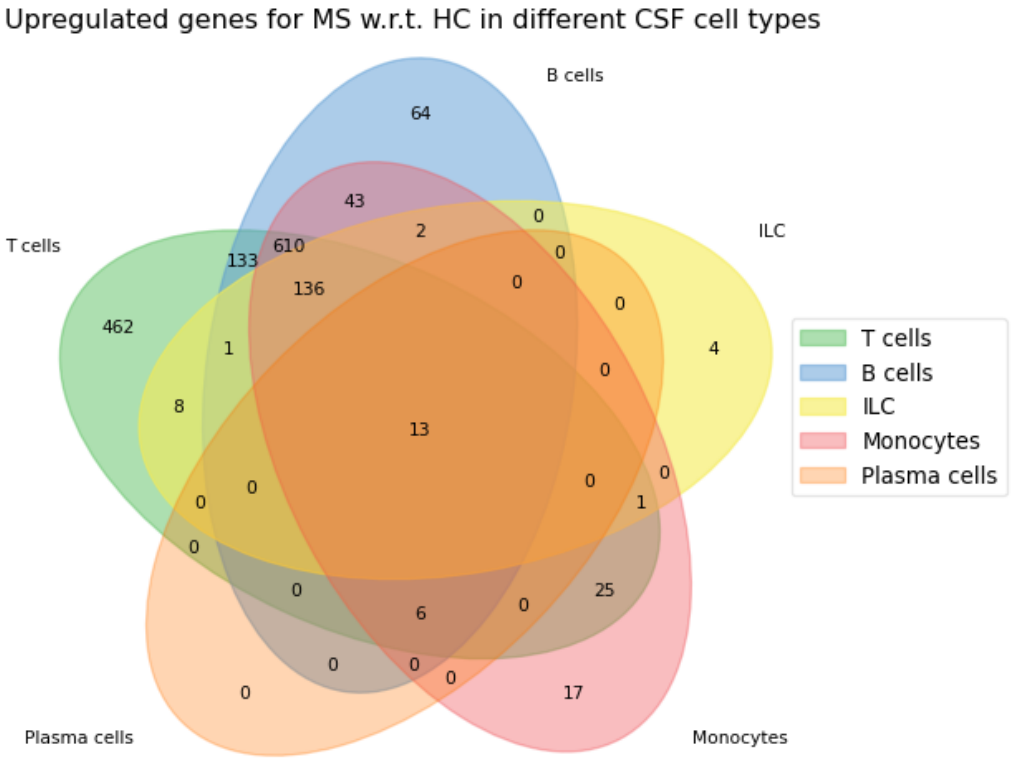
Differential Expression Results: Downregulated in PBMCs



- **B cells** exhibit the **highest number** of **unique downregulated genes** (1987 genes).
- A distinct **overlap of 126 genes** is observed uniquely between **T cells** and **B cells**, indicating **shared suppression mechanisms**.
- **B cells**, **Monocytes**, and **Plasma cells**, highlighting potential **interconnected pathways** affected by **downregulation**.

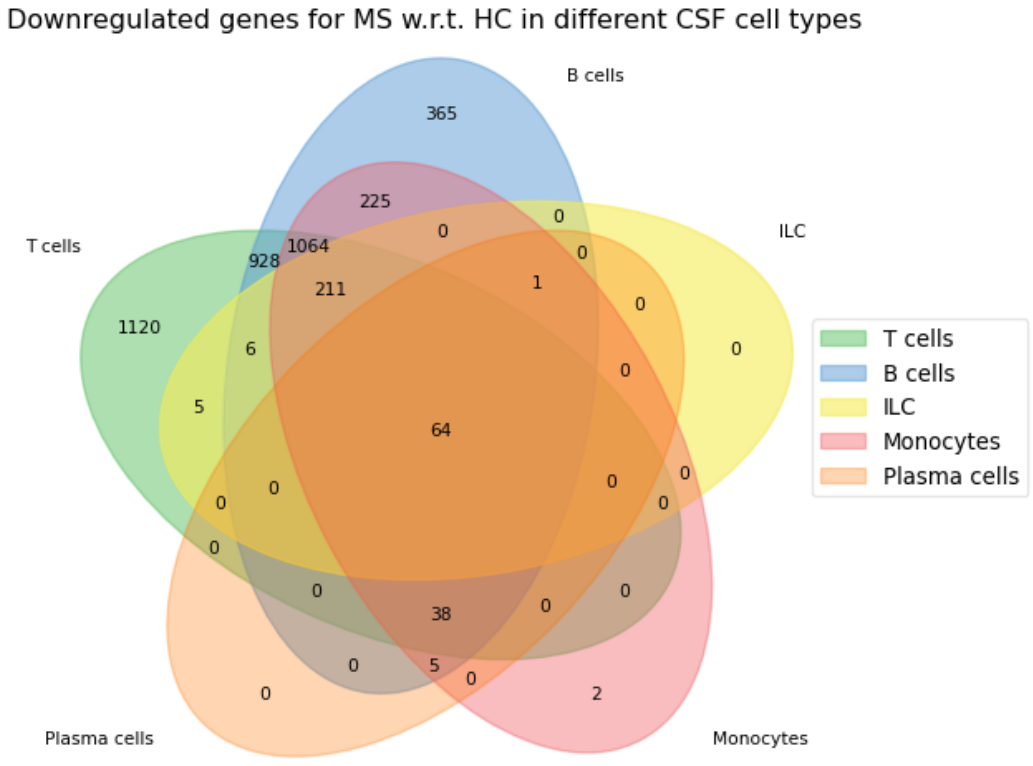
Differential Expression Results: Upregulated in CSFs

- **T cells** exhibit the **highest number** of **unique upregulated genes** (462 genes).
- Significant overlap of 610 genes shared between **T cells**, **B cells**, and **Monocytes**: potential **common pathways**.
- 13 genes being **common** to **all five cell types**, highlighting potential **universal mechanisms** involved i**n MS pathology**.



Differential Expression Results: Downregulated in CSFs

- **T cells** exhibit the **highest number** of **unique downregulated genes** (1120 genes).
- Overlap of 1064 genes shared between **T cells**, **B cells**, and **Monocytes** again suggests **shared pathways**.
- Between **T cells** and **B cells** highest number of **shared genes detected**, totaling 2311, of which 928 are uniquely shared.
- 64 genes are **downregulated across all five cell types**, indicating a **suppression** of certain functions or **pathways in MS**.



Comparative Analysis in B Cells

Comparative Analysis in B Cells (1)

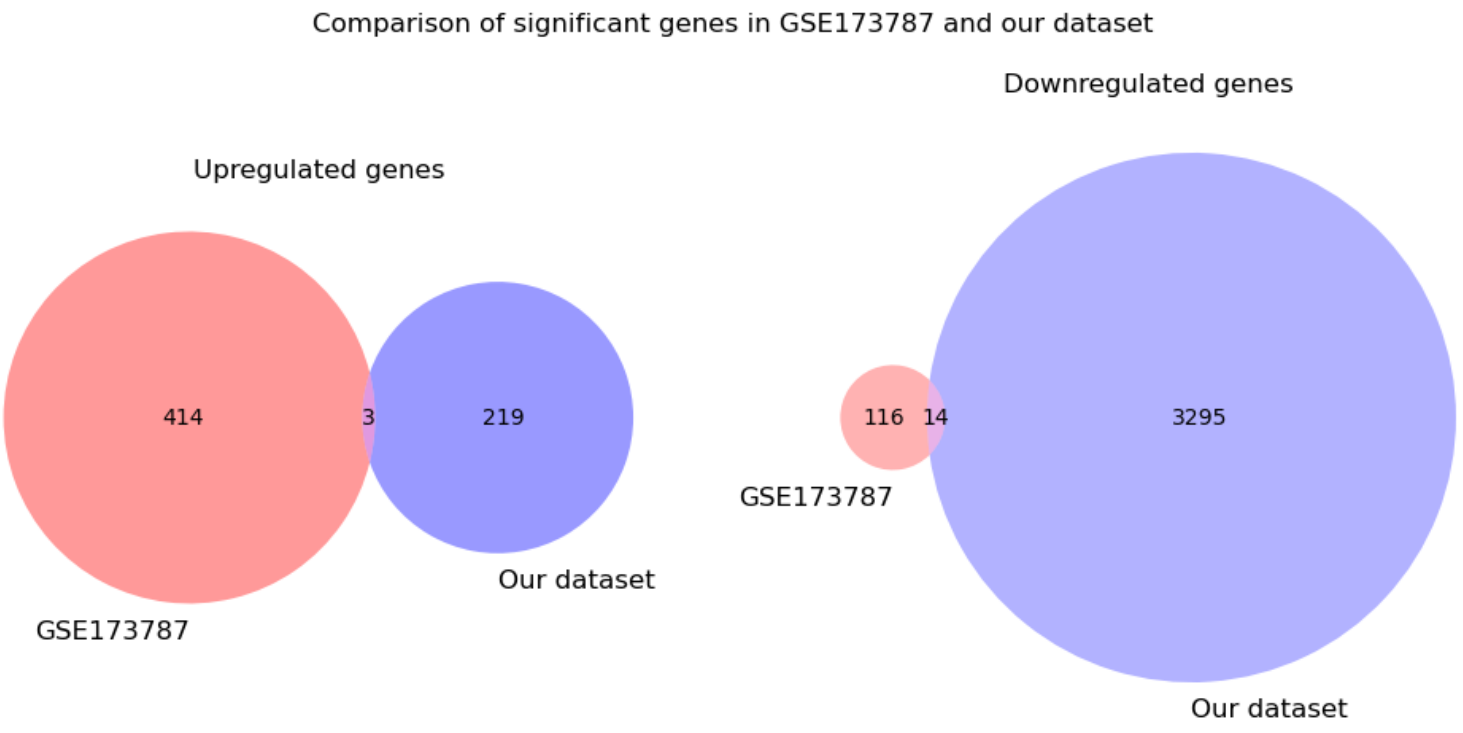
- Compared up and downregulated genes in B cells between GSE173787 dataset and our PBMC dataset.

Dataset	Upregulated Genes	Downregulated Genes
GSE173787	417	130
Our Dataset	222	3309

Table 3: Differentially Expressed Genes in B Cells

- Found **3 shared upregulated genes** and **14 shared downregulated genes**.
- Highlighted **variability** in gene expression patterns **across different studies**.

Comparative Analysis in B Cells (2)



Upregulated shared genes: IGLC3, IGLC2, and IGKC.
Downregulated shared genes: CCDC28A, OGFRL1, BRI3, HSPB1, IFNGR1, MID1IP1, CSNK1E, EMC6, DUSP6, NDUFV2, DHRS3, CLEC2B, CMTM3, and BEX4.

Conclusion

Limitation

- Limited data availability may impede comprehensiveness and introduce bias.
- **Absence of a representative statistical sample** can undermine validity and generalizability.
- Comparative analysis highlights variability across studies.
- **Further research needed** to address data gaps and enhance robustness.

Conclusion

- Significant **gene expression differences** observed **in MS patients** compared to HCs.
- Higher number of **downregulated genes** in MS, particularly **in CSF**.
- Our results suggest a suppression of pathways in MS, contributing to the disease's pathology.
- **Common pathways** identified between **T cells**, **B cells**, and **Monocytes**.
- Our study serves as a **starting point** for **further research** to better understand MS pathology and gain deeper insights into underlying mechanisms, potentially **leading to better diagnostic and therapeutic strategies**.

Supplementary Material

Dataset: GSE239626

- 72,317 PBMCs from 10 MS patients (two samples each, 3 months apart)
- Only dataset with antibody data, used for cell type labelling
- Vitamin D supplementation; no positive outcomes

- 36,601 genes per patient
- 35 antibodies per patient
- Top Gene: MALAT1
- Top Proteins: CD11a, CD3, CD8

Dataset: GSE239626

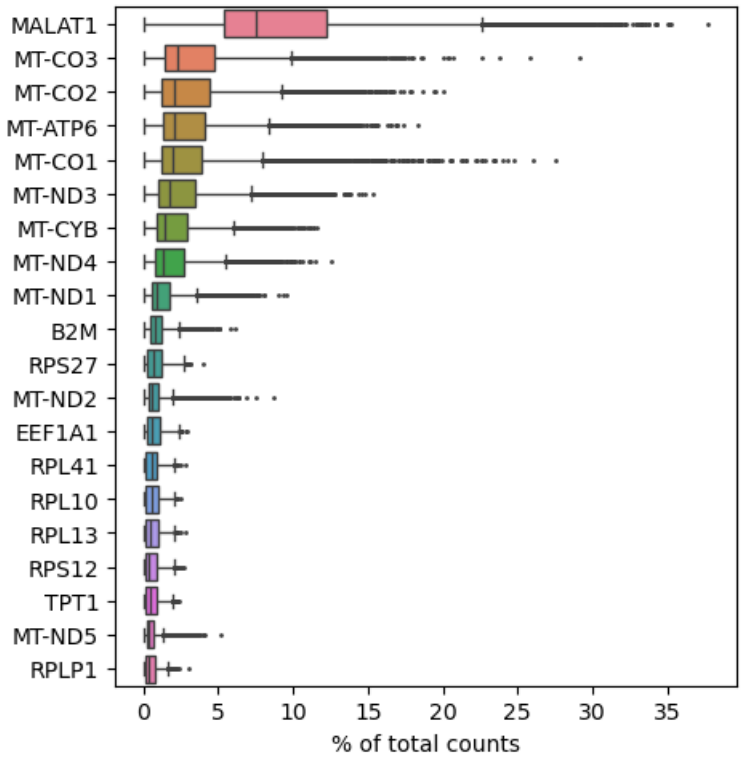


Figure 2: Box plot illustrating the distribution of genes expression levels for the top 20 genes in dataset GSE239626 with the highest average expression across all cells in the dataset. The gene with the highest expression is MALAT1, followed by mitochondrial genes such as MT-CO3, MT-CO2, MT-ATP6, and others. Other notable genes include B2M, RPS27, EEF1A1, and several ribosomal genes like RPL41, RPL10, RPL13, RPS12, and RPLP1.

Dataset: GSE239626

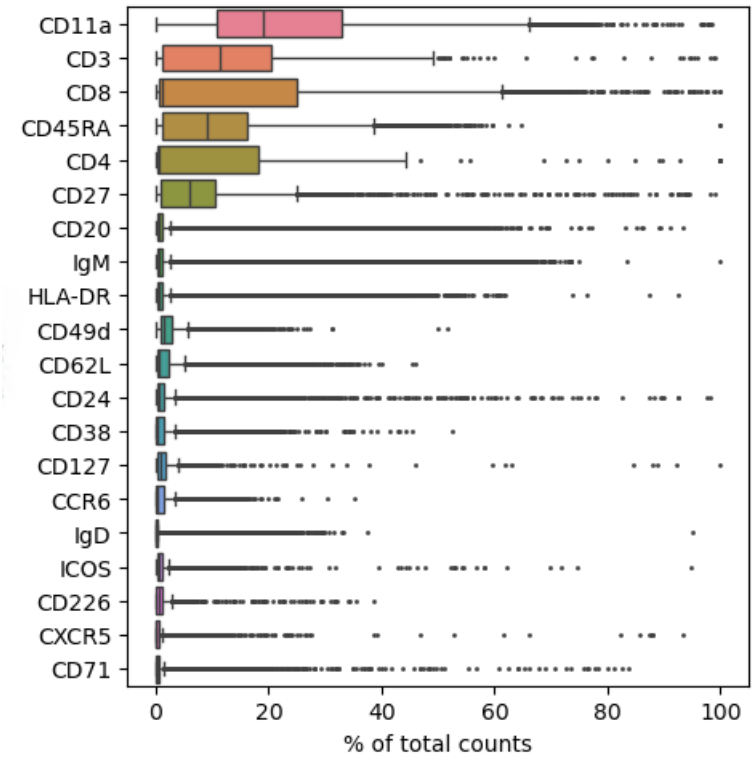


Figure 3: Box plot illustrating the distribution of antibodies expression levels for the top 20 proteins in dataset GSE239626 with the highest average expression across all cells in the dataset. The antibody with the highest expression is CD11a, followed by CD3, CD8, CD45RA, and CD4. Other notable proteins include CD27, CD20, IgM, HLA-DR, as well as CD49d, CD62L, CD24, and CD38.

Dataset: GSE239626

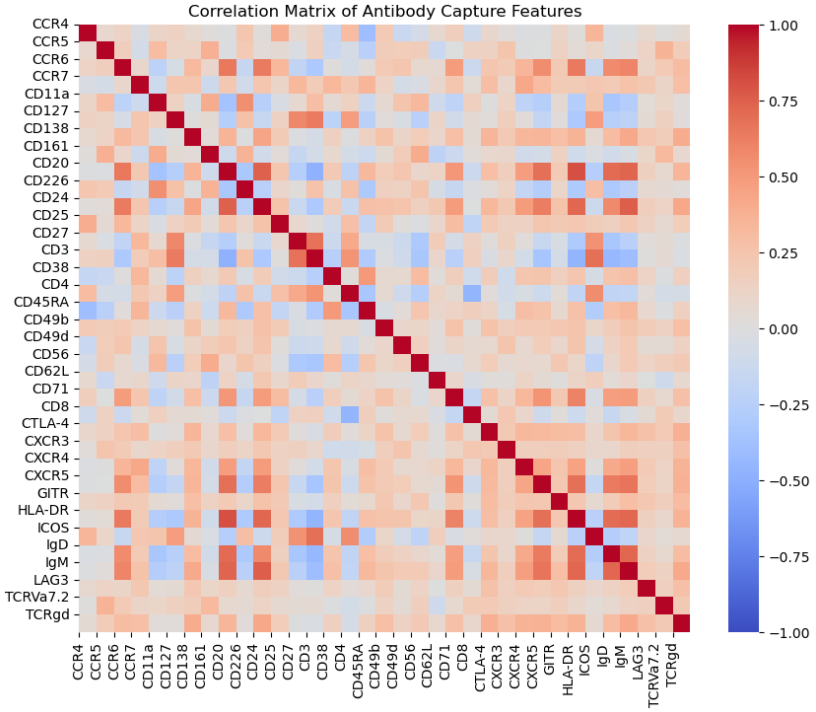


Figure 4: Correlation matrix of antibody features, indicating the degree of correlation among the 35 antibody features used in the study, providing a comprehensive view of the relationships between various antibody capture features. Correlation matrix showcases a wide range of correlation values, from strongly positive (red) to strongly negative (blue).

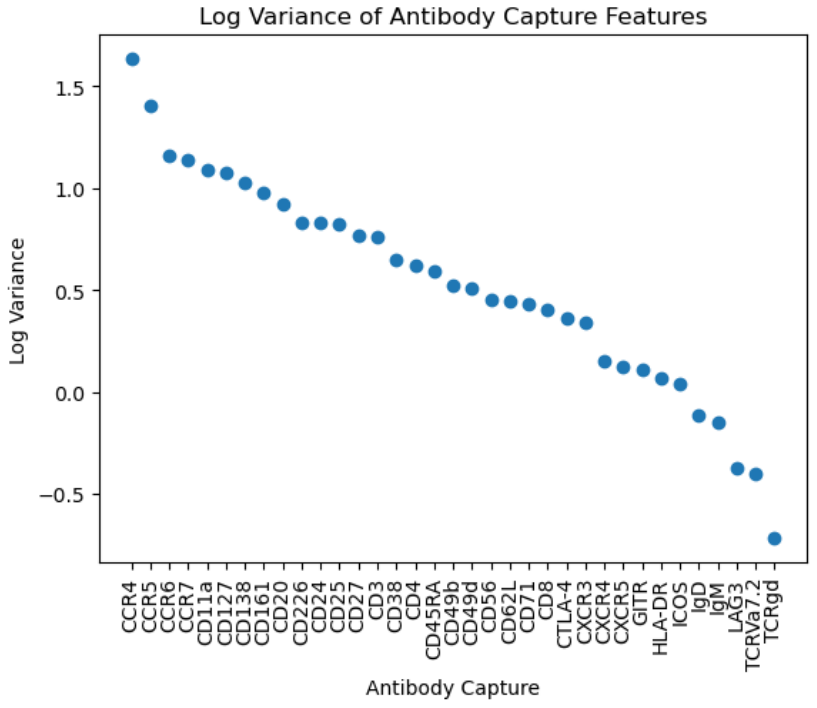


Figure 5: Logarithmic variance of the antibody capture features. The figure highlights the variance of each feature. The features CCR4, CCR5, and CCR6 show the highest logarithmic variances, indicating significant variability in their expression levels across samples. Conversely, features such as TCRV7.2, TCRgd, and LAG3 exhibit very low logarithmic variances, suggesting minimal variability in their expression levels.

Dataset: GSE194078

- Transcriptomic data from 12 patients: 3 with MS, 9 healthy controls.
- Total of 109,891 cells analyzed: 75,675 healthy, 34,216 MS.

- Individual variations observed; unclear distinction between MS patients and controls.
- MALAT1 predominant, with high expression of mitochondrial and ribosomal genes.

Dataset: GSE194078

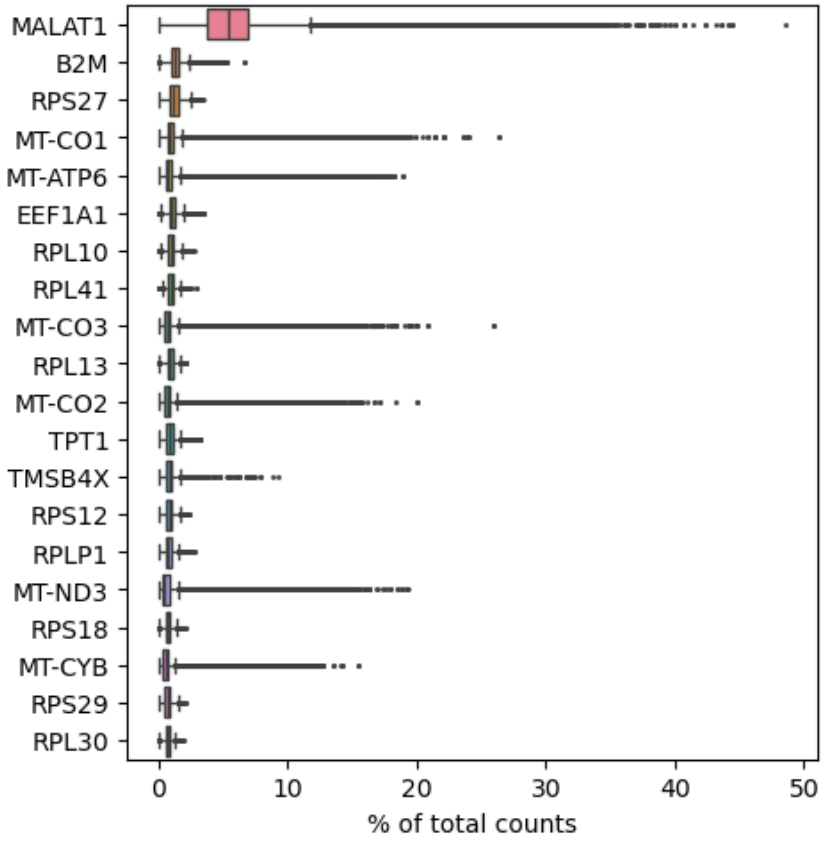


Figure 6: Box plot illustrating the distribution of gene expression levels for the top 20 genes in dataset GSE194078 with the highest average expression across all cells in the dataset. This plot effectively visualizes the distribution of total counts among the most highly expressed genes, with MALAT1 standing out as the most predominant. The remaining genes exhibit much lower percentages of the average fraction of counts assigned to each gene across all cells, all staying below 10%. This distribution showcases the dominance of one highly expressed gene while the majority maintain lower expression levels.

Dataset: GSE138266

- CSF and PBMC samples from MS patients and healthy controls.
- 6 HC and 6 MS patients for CSF; 5 HC and 5 MS for PBMC.
- Total cells: 814,177.

- Data from 3 HC (PST) discarded after evaluating their distribution.
- Used 25,553 PBMC cells and 19,306 CSF cells (MS); 15,604 PBMC and 5,641 CSF (HC).

Dataset: GSE138266

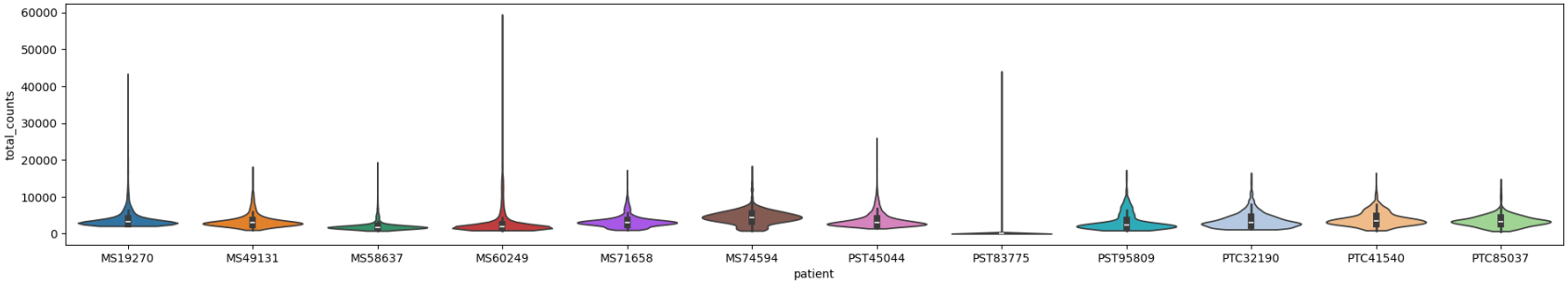


Figure 7: Violin plots illustrate the distribution of the total number of genes detected per CSF cell for each patient in the GSE138266 dataset. The dataset includes samples from 6 MS patients (identified by codes starting with MS) and 6 healthy controls (identified by codes starting with PST and PTC).

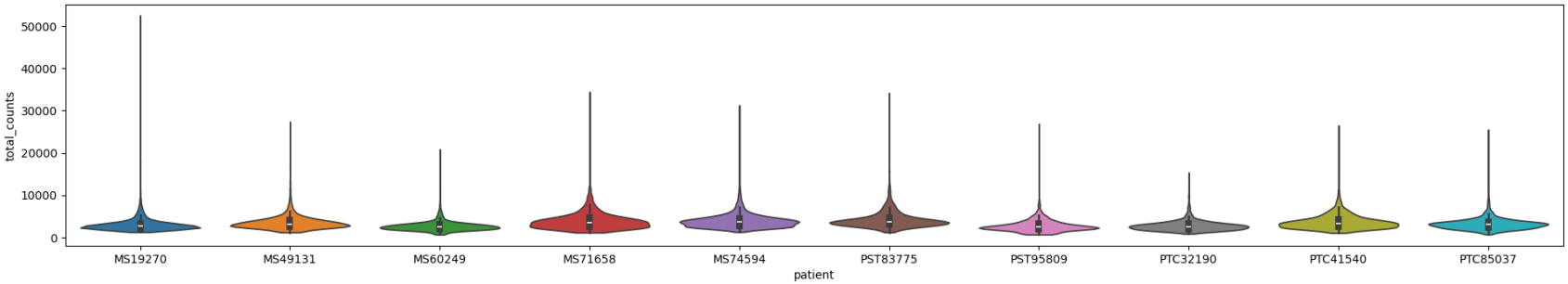


Figure 8: Violin plots illustrate the distribution of the total number of genes detected per CSF cell for each patient in the GSE138266 dataset. The dataset includes samples from 5 MS patients (identified by codes starting with MS) and 5 healthy controls (identified by codes starting with PST and PTC).

Datasets: GSE173787

Integration of data from multiple studies to enhance analysis:

- Complemented our gene expression analysis with data from GSE173787.
- Identifies Differentially Expressed Genes (DEGs) through transcriptomic analysis of CD19+ B cells.
- 547 genes differentially expressed between Multiple Sclerosis (MS) patients and Healthy Controls (HC).
- Highlights alterations in complement activation pathway and B lymphocyte differentiation and activation pathways.

Dataset: GSE173787

	logFC	logCPM	P value	FDR
mean	0.8677	3.6941	6.74×10^{-4}	1.53×10^{-2}
std	1.1687	2.2032	8.42×10^{-4}	1.38×10^{-2}
min	-3.1006	-0.3859	6.06×10^{-13}	1.01×10^{-8}
25%	0.5944	1.8440	3.57×10^{-5}	3.32×10^{-3}
50%	1.0172	3.5614	2.64×10^{-4}	1.07×10^{-2}
75%	1.6418	5.1479	1.07×10^{-3}	2.52×10^{-2}
max	6.6208	12.0595	3.26×10^{-3}	4.99×10^{-2}

Table 4: Summary statistics of logFC (log fold change), logCPM (log counts per million), p-value, and FDR (false discovery rate) for the 547 differentially expressed genes in dataset GSE173787.

Performance evaluation Leiden clustering

neighbors	resolution	Silhouette	Calinski−Harabasz	Davies−Bouldin
5	0.5	0.0997	3937.4035	1.9732
5	0.8	0.0636	3114.0382	2.1261
5	1.0	0.0684	2943.2850	2.2308
10	**0.5**	**0.1255**	**4947.5156**	**2.0338**
10	0.8	0.1082	4214.3011	2.2261
10	1.0	0.0829	3850.6357	2.1951

Table 5: Performance evaluation of the Leiden clustering algorithm applied to the selected five principal components for antibody data analysis. The table displays clustering performance metrics for various parameter combinations, encompassing Silhouette score, Calinski−Harabasz score, and Davies−Bouldin score.

Predicted cell type labels using `CellTypist`

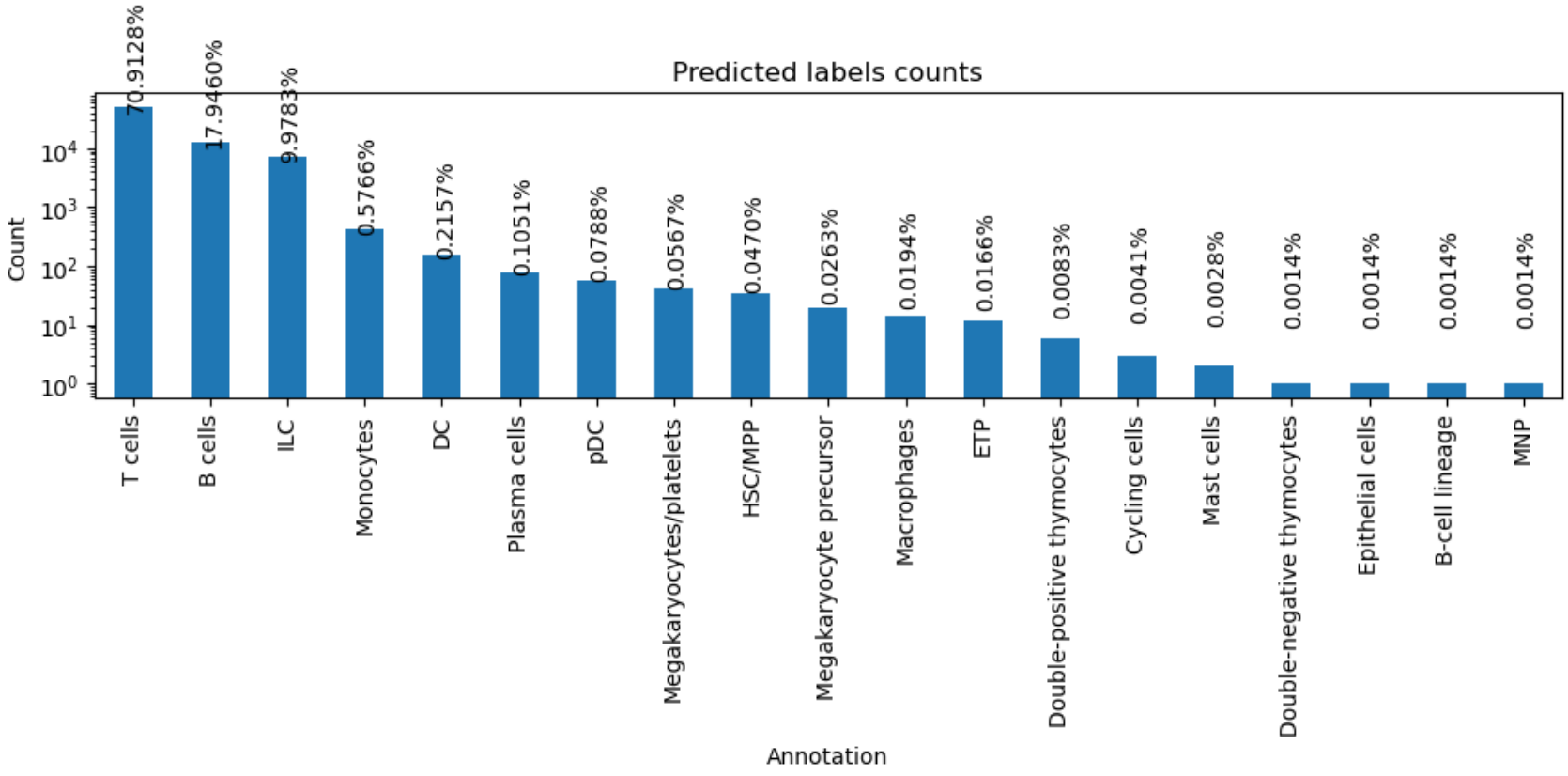


Figure 9: Barplot illustrating the distribution of predicted cell type labels across the GSE239626 dataset. The majority of cells were assigned the label T cells (70.91% of the dataset), followed by B cells (17.95%) and ILC (9.98%). Other cell types had much lower representation.

Predicted cell type labels using `CellTypist`

Cell Type	PBMC MS	PBMC HC	CSF MS	CSF HC
T cells	84479	48474	31911	28825
B cells	16866	7020	2905	6088
ILC	10672	3104	889	471
Monocytes	2003	2710	579	1357
Plasma cells	809	570	172	461
HSC/MPP	81	14	2	24
DC	187	38	15	29

Table 6: Cell counts across different conditions and cell types. PBMC MS and PBMC HC represent peripheral blood mononuclear cells in multiple sclerosis patients and healthy controls, respectively. CSF MS and CSF HC represent cerebrospinal fluid cells in multiple sclerosis patients and healthy controls, respectively. The most frequent cells are T cells, followed by B cells. The number of cells varies between different conditions, but the proportions of increase or decrease remain consistent across conditions.

Cell Type Labels Transfer

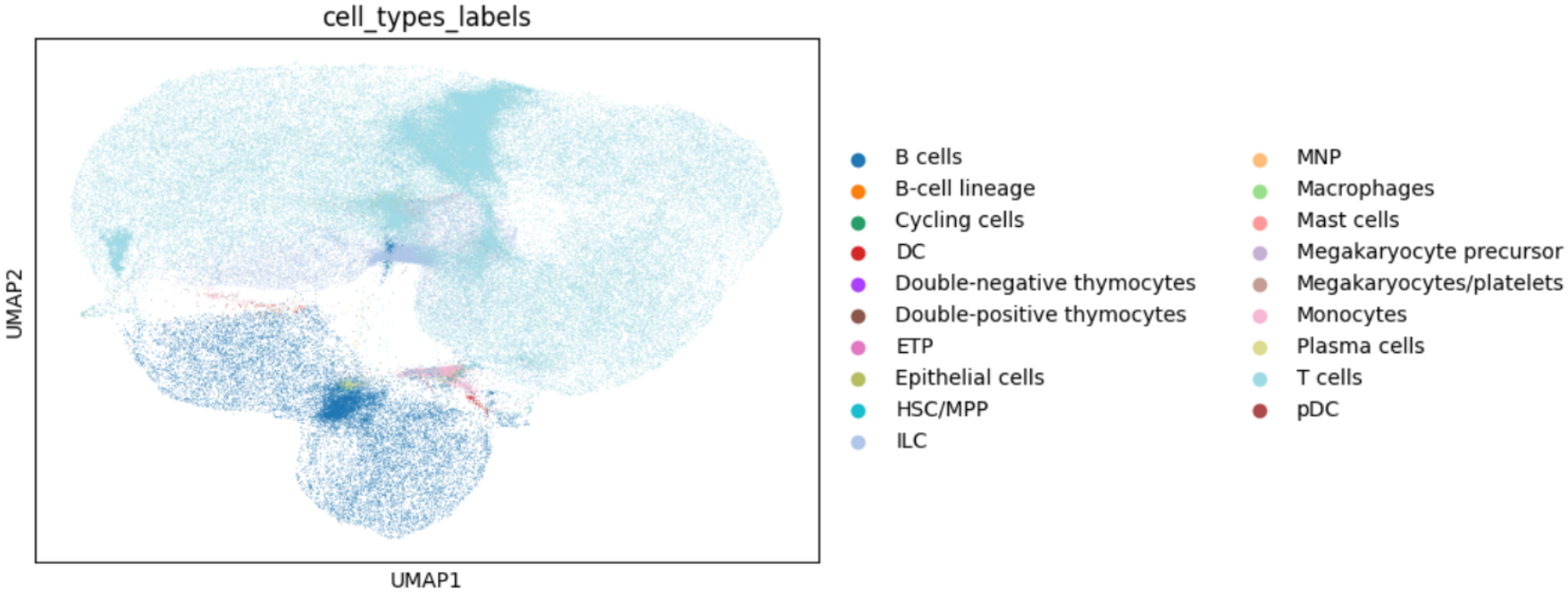


Figure 10: Representations of the merged dataset comprising the reference dataset GSE239626 with the additional datasets GSE138266 in the two principal component UMAP space. In the figures, distinct cell types are highlighted with different colors.

Cell Type Labels Transfer

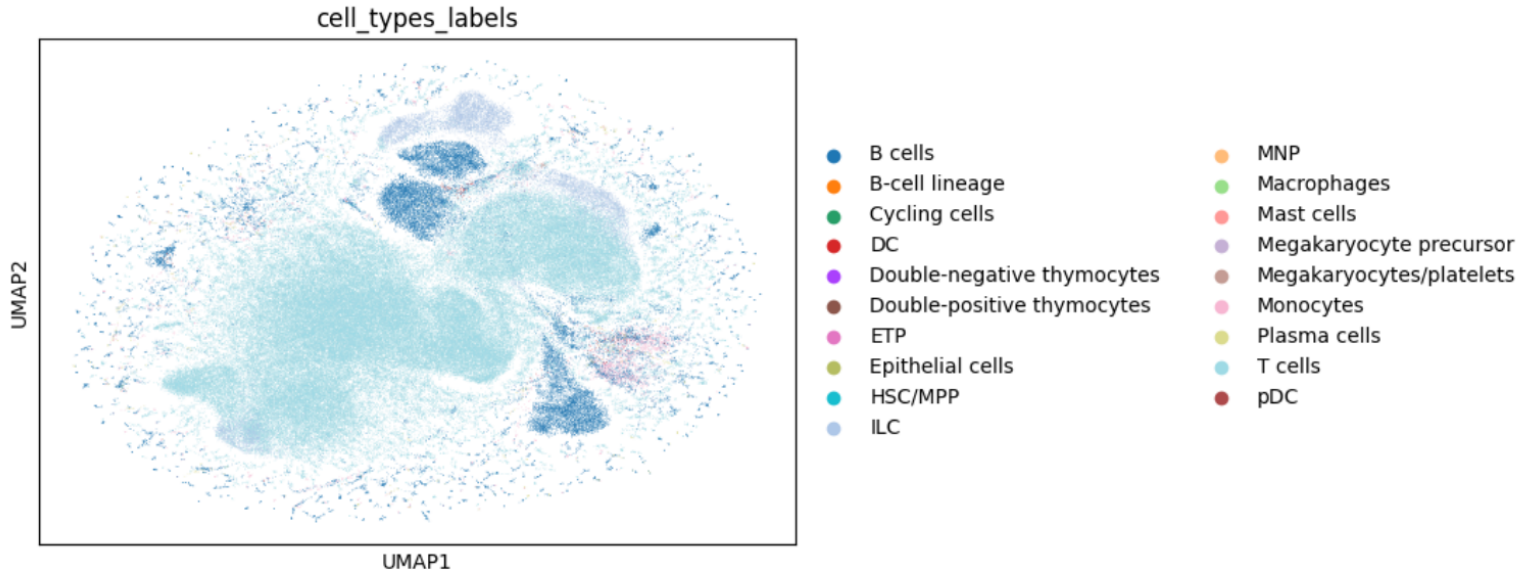


Figure 11: Representations of the merged dataset comprising the reference dataset GSE239626 with the additional datasets GSE194078 in the two principal component UMAP space. In the figures, distinct cell types are highlighted with different colors.

Predicted cell type labels using `CellTypist`

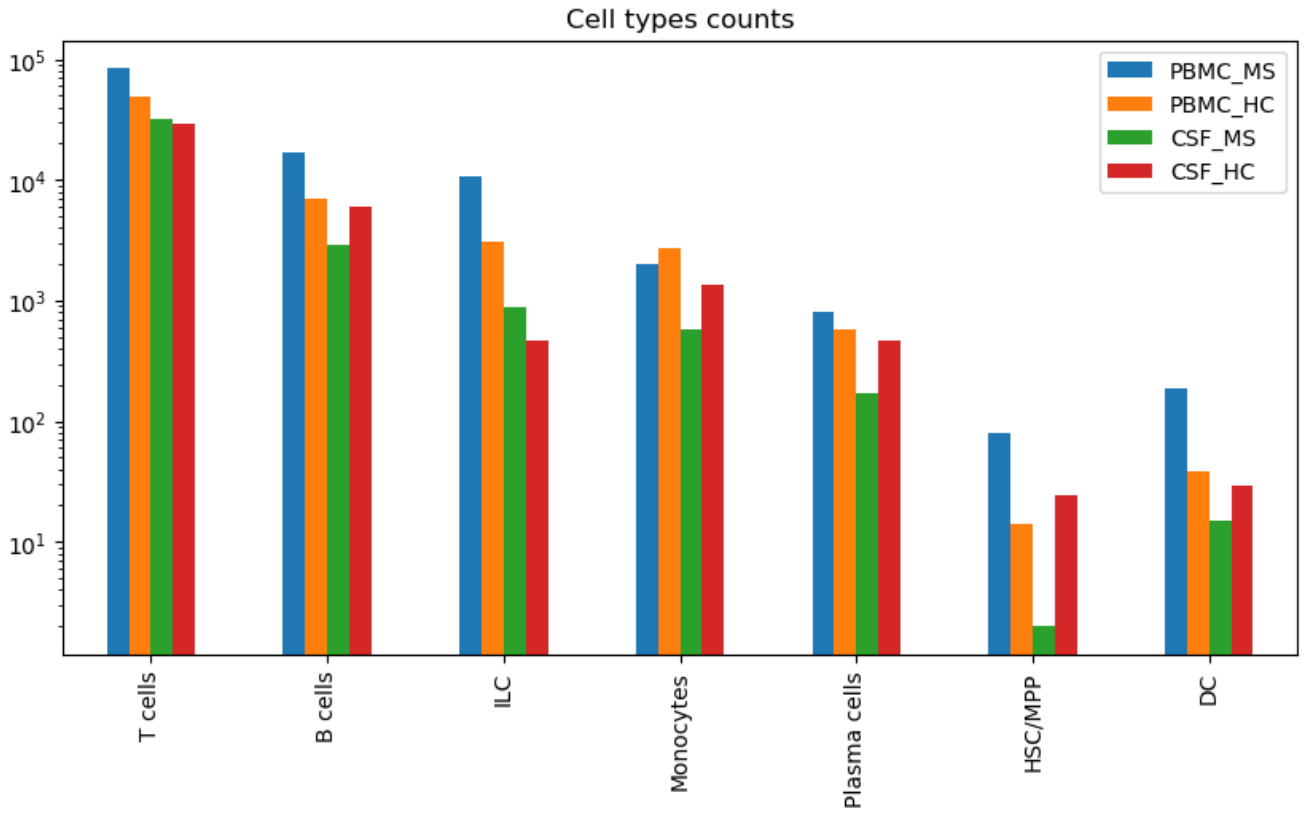
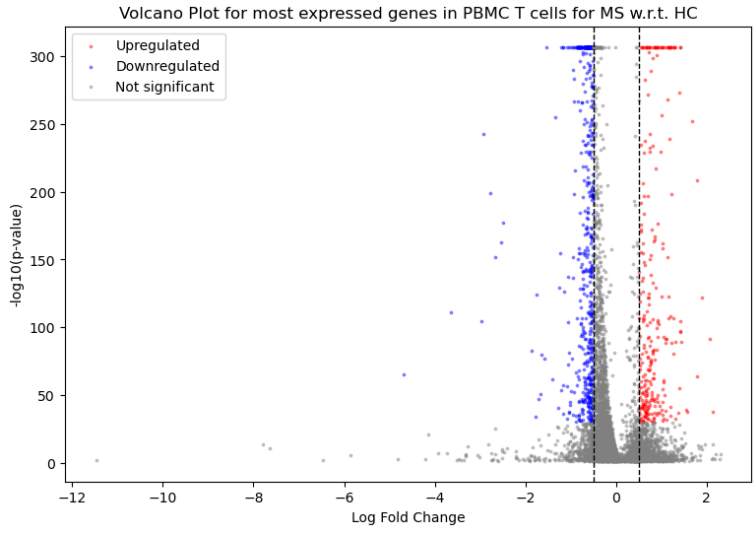
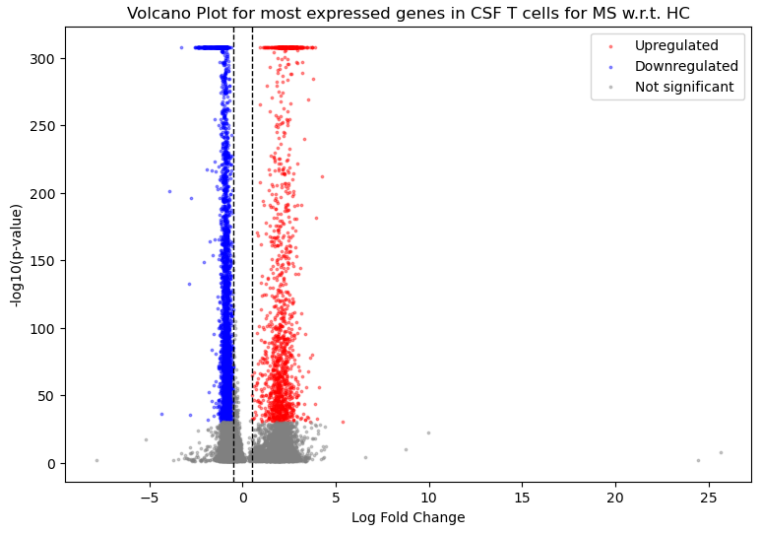
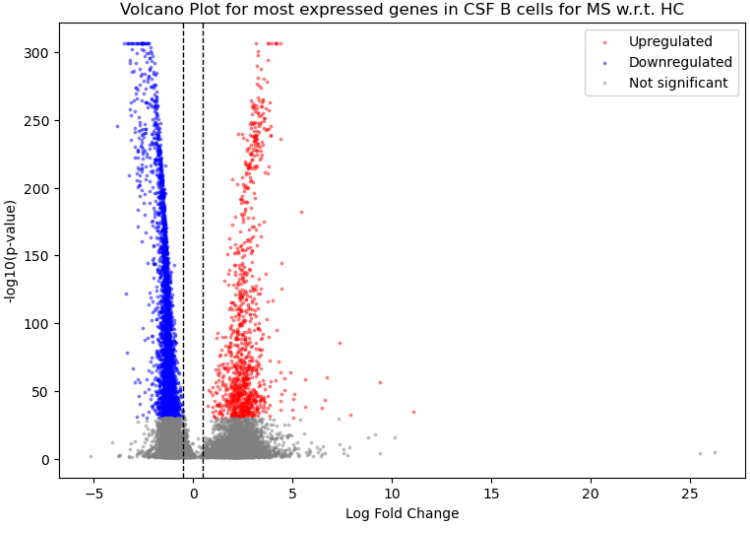


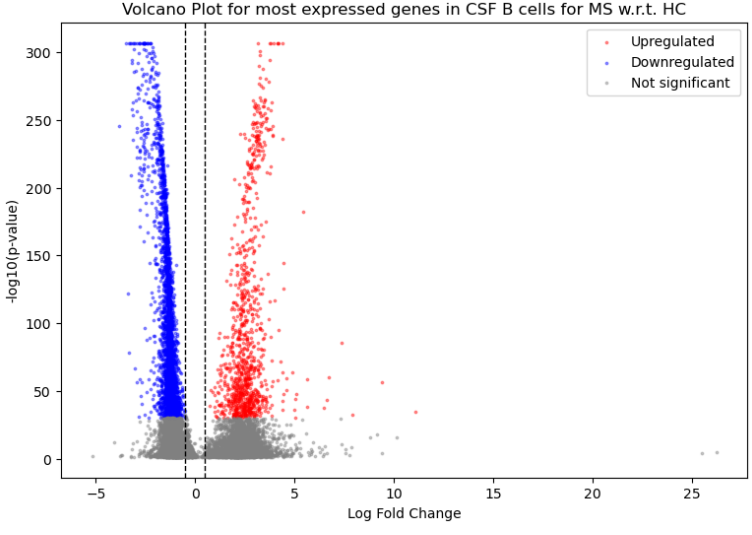
Figure 12: Bar plot showing the cardinality of each cell type with respect to PBMC or CSF and HC or MS samples. The cell types include T cells, B cells, ILC, Monocytes, Plasma cells, HSC/MPP, and DC. The y-axis is presented on a logarithmic scale to better visualize the differences in cell counts across the different conditions. Notably, the number of cells reported in MS samples is generally higher than in HC samples, with a particularly pronounced difference in T cells in the blood. This discrepancy reflects the imbalance in the number of samples between the MS and HC classes in our dataset.

Volcano plots illustrate upregulated and downregulated genes

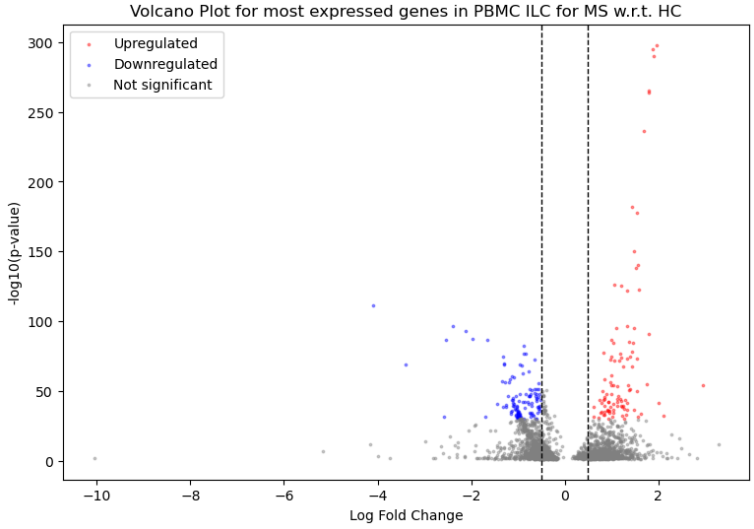


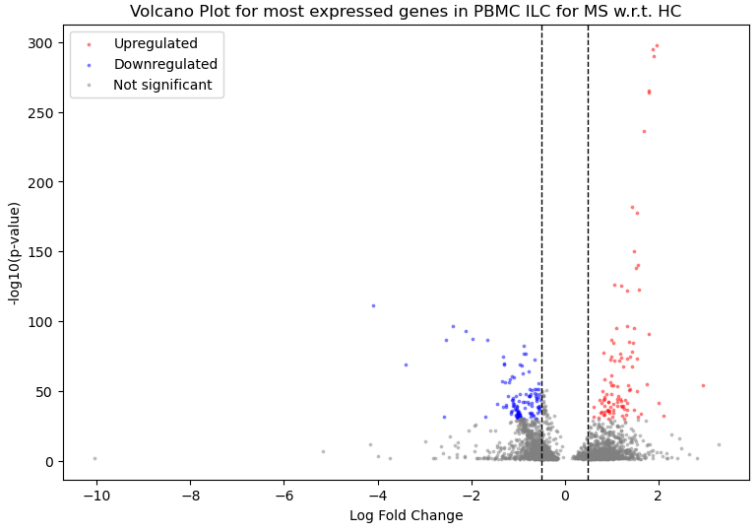


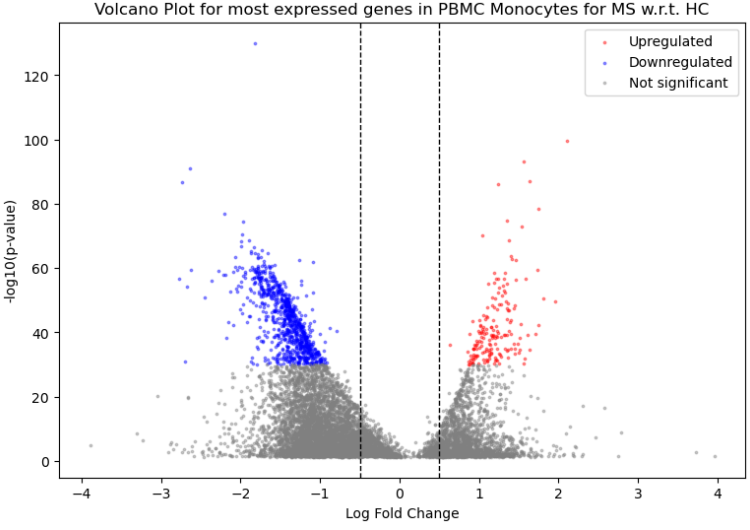


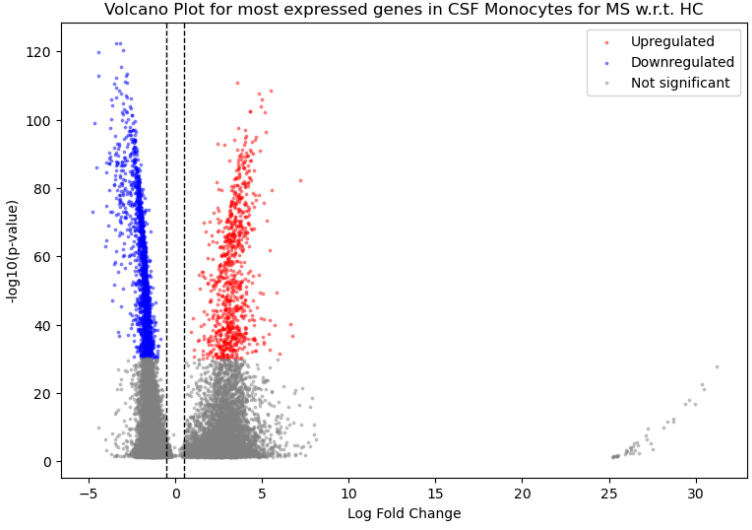


Volcano plots illustrate upregulated and downregulated genes









Volcano plots illustrate upregulated and downregulated genes

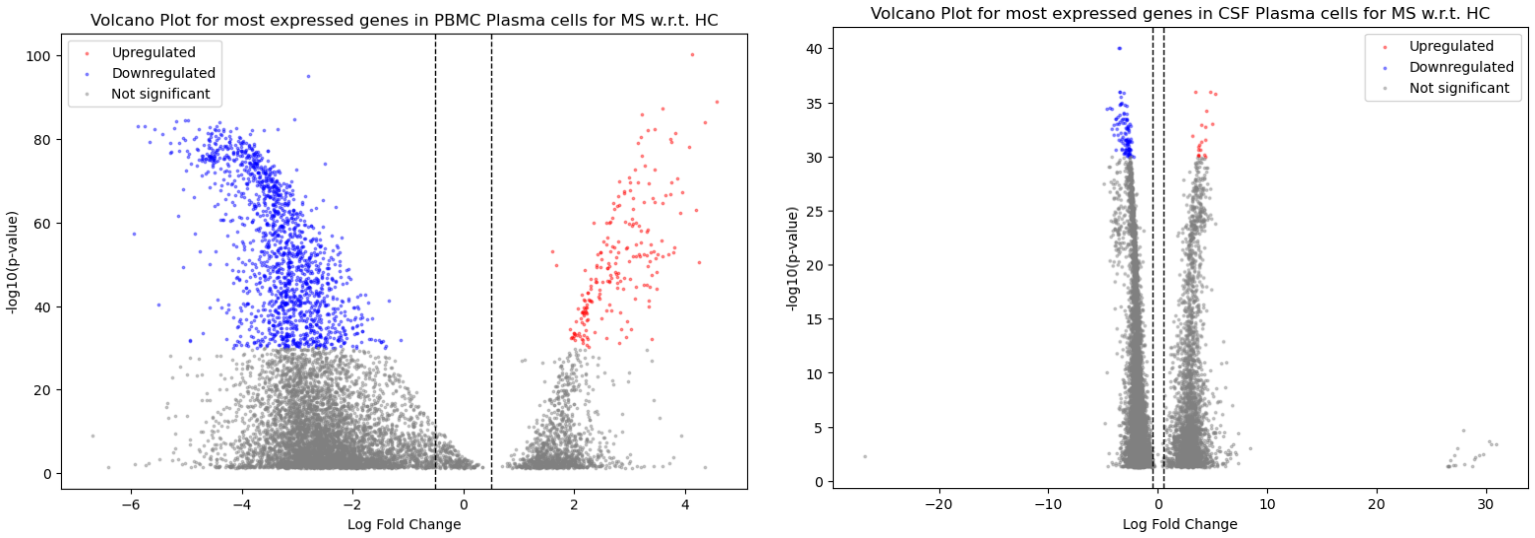


Figure 13: Volcano plots visually illustrate genes that are upregulated (in red) and downregulated (in blue) in the cell populations from MS patients compared to those from HC samples. The criteria for determining significant genes include a threshold value for the adjusted p-value of 10^{-30} and a threshold for log fold change of 0.5. Genes shown in gray in the plot correspond to non-significant genes. For each cell type, the volcano plot on the left represents PBMC cells, and the one on the right represents CSF cells.

Common biomarkers across all cell types

	List of Genes
Upregulated PBMC	-
Downregulated PBMC	RPL8, RPL38, CD48, HLA-DPB1, RPS13, RPL24, RPL12, FTL, UQCRB, FTH1, RPL27A, COX7C, RPS8, UBA52, NACA, SH3BGRL3, S100A6, RPL23, GAPDH, LGALS1, CCNI, RPS20, ATP5MC2, ATP5F1E, S100A4, BTF3, TMA7, RPL7, MYL12B, OAZ1, RPL31, HLA-DPA1, RPL29
Upregulated CSF	PCDHA2, CISD1, LINC00588, AJ003147.1, ETF1, AQP4-AS1, NOMO3, TEX52, AP001351.1, ESR2, AL359475.1, PCDHA4, LINC02417
Downregulated CSF	VAMP8, H3F3A, YBX1, MT-ND4, MT-ND5, HMGB1, SLC25A6, RHOA, TPM3, FTL, ITGB2, GABARAP, TLN1, SEPTIN7, PNRC1, RPL13A, CD99, TPI1, EMP3, ELOB, CDC42, RPS8, UQCRH, COX5B, SH3BGRL3, ARPC5, S100A6, MT-ND2, RPL23, GAPDH, COX6B1, S100A11, CYBA, ARPC1B, CFL1, TXNIP, COX8A, ARPC2, SERF2, IQGAP1, RPS10, RPS20, VIM, ATP5MC2, GNAI2, MTRNR2L12, SRP14, ATP5F1E, RPLP0, RPS17, S100A4, TMSB10, IFITM2, PTMA, GMFG, ITM2B, RPL7, OAZ1, CD63, ARPC3, RPL31, SET, HCST, RAC1

Table 7: Genes that are upregulated and downregulated in common across all cell types (T cells, B cells, Monocytes, Plasma cells, ILC) for PBMC and CSF. The tables demonstrate that 33 downregulated genes are common to all five cell types in PBMC. Furthermore, 13 upregulated genes and 64 downregulated genes are common to all five cell types in CSF, indicating potential universal mechanisms involved in MS pathology.