IJHSR
ijhighschoolresearch.org
RESEARCH ARTICLE

# Machine Learning-Driven Meta-Analysis Reveals Shared B Cell Repertoires linked to Multiple Sclerosis

Ayaan Agarwal[1], Aveer Singh[2]

1) American School of Bombay, SF2, G Block, Bandra-Kurla Complex Road, Mumbai, Maharashtra, 400051, India; ayaanagarwal209@gmail.com
2) Vasant Valley School, Sector C, Vasant Kunj, New Delhi, Delhi, 110070, India

ABSTRACT: Multiple sclerosis (MS) is an autoimmune, neurodegenerative disease where the immune system damages myelin sheaths in the central nervous system. MS pathogenesis was attributed to auto-reactive T cells. However, recent studies suggest that B cells also play a vital role, including antibody-dependent and independent effects in compartmentalized inflammation. Clonotypes are groups of B cells with the same receptor. To explore the role of B cell clonotypes in MS pathogenesis, we conducted a meta-analysis of raw high-throughput sequencing data of the B cell repertoire of 82 subjects (48 MS patients and 34 controls). A novel combination of density-based clustering (DBSCAN) and supervised ML (random forest) was applied to cluster clonotypes into unique clonotype groups (UCG), and their disease relevance was assessed. Random forest feature importances, Mann-Whitney tests, and empirical power estimation revealed six significant, potentially pathogenic UCGs. Additionally, tertiary analyses revealed genetic variability and clonal expansion as hallmarks of MS-associated B cells. Interestingly, new genomic motifs among our potentially pathogenic UCGs aligned with known IGHV4 family biases. Overall, the study highlights the utility of machine learning analysis in decoding immune repertoires and contributes to our understanding of B-cell involvement in MS pathogenesis, paving the way for future therapeutic exploration.

KEYWORDS: Biology, Computational Biology and Bioinformatics, Multiple Sclerosis, Pathogenic B cells, Machine Learning.

## Introduction

Multiple sclerosis (MS) is a chronic, demyelinating disease affecting the body's central nervous system. Currently incurable, the estimate for global MS prevalence is around ~2.8 million people.[1] Symptoms include a lack of coordination, partial or complete loss of vision, fatigue, trouble with memory, and more.[2] While MS is a B- and T-cell-mediated disease, and B-cell-targeting therapies have demonstrated clinical success, specific pathogenic B-cell clonotypes and BCR sequence patterns that drive disease pathogenesis remain underexplored in current research.[3]

B cells or lymphocytes are key cells of the adaptive immune system. Primary functions of B cells are antibody production, immunological memory, antigen presentation, and cytokine secretion. Antigen binding to B cell receptors (BCRs) and subsequent activation by helper T cells lead to B cells undergoing clonal expansion, where they generate genetically identical clones. Clonotypes are groups of genetically identical B cells. B cells also offer antigen recognition diversity by undergoing somatic hypermutation. The BCR is a 'fingerprint' for B cells, which provides representative information of the B cell for classification and genetic analyses. The V(D)J regions of the immunoglobulin gene shuffle to generate genetic diversity amongst the BCR.

In multiple sclerosis, research suggests autoantibodies targeting self-antigens are contributors to MS pathogenesis;[3] however, it is unclear which B cells are associated with the disease. Anti-CD20 therapies have been proven effective in the treatment of multiple sclerosis,[4] further rationalizing the additional exploration of B cells as drug targets.

Machine Learning (ML), which is a subset of artificial intelligence, excels at extracting features from high-dimensional BCR genetic data for clustering and identifying non-linear patterns indicative of pathogenicity that might be missed by conventional analyses. In this study, we performed an ML-driven meta-analysis of the B cell receptor sequencing repertoire of MS patients and healthy controls (HCs) to identify potentially pathogenic B cells for therapeutic targets.

## Methods

Figure 1 is a graphical representation of our methodology.

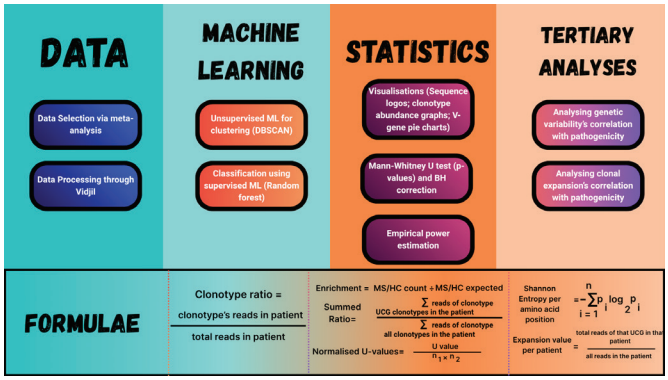


**Figure 1:** Graphical representation of our methodology; chronology: left to right. Highlights steps for data processing, machine learning, statistics, and tertiary analyses. All formulae used are summarized here.



DOI: 10.36838/IJHSR816.35

***Dataset & Data Selection:***

A traditional limiting factor for RNA-seq meta-analysis studies of this type is the high cost and effort associated with large datasets;[5] nevertheless, combining datasets of existing studies can eliminate this constraint. This study incorporates a dataset of MS patients larger than any prior study of this nature reviewed by us.

The meta-analysis is performed using diverse search criteria on research databases (such as Google Scholar), including "BCR RNA-seq (MS/HC)," "Immunoglobulin RepSeq (MS/HC)," "AIRR seq (MS/HC)," and "Effect of [MS immune drug] on BCR RNA-seq repertoire of MS patients." Out of the 8 identified datasets, 2 were exclusively HC datasets,[6,7] 5 were MS datasets,[8–12] and 1 was mixed.[13] A total of 82 subjects (48 MS patients and 34 HCs) were identified for processing.

***Data Processing:***

Vidjil is a platform for clonal analysis of lymphocyte high-throughput sequencing data of the immunoglobulin region. Downloaded FastQ files from the Sequence Read Archive were uploaded to the Vidjil web server for clonotype analysis. Resultant CSV files provided information regarding each clonotype's member count (or 'reads') and the ratio between the clonotype's reads in the patient and the total reads of all clonotypes in that patient. CSV files were merged, and unproductive clonotypes (such as those with stop codons in the nucleotide junction) were filtered out. Irrelevant data columns were removed. For patients whose sequencing was performed across various samples, clonotype reads were summed across samples. A new source column was added to specify the patient of origin. Refer to the supplementary manuscript file under 'S1' for more details.

***Clustering:***

We identify thousands of clonotypes, although each clonotype has low frequency counts (reads). This is because B cells undergo somatic hypermutation, where they rapidly mutate their DNA in an immune response. This creates a diverse repertoire of similar B cells but of distinct clonotypes. Consequently, utilizing supervised ML classification on our raw clonotype data is ineffective, as each clonotype becomes a feature for the ML algorithm, and a large number of features with very low variance (most clonotypes have close to 0 reads) massively impacts model performance. Therefore, to counter the high-dimensional clonotype makeup, we cluster similar clonotypes into unique clonotype groups (UCGs); these UCGs are used downstream as our features for classification.

To cluster clonotypes into UCGs, we needed parameters to classify similar clonotypes. We formulate these conditions based on a letter in Leukemia and use them to develop a novel, custom distance function to compare the similarity between two clonotypes.[14] The distances were integrated with density-based clustering. Parameters for similarity are as follows: two clonotypes are in the same cluster when they have a minimum 80% identity in their CDR3 amino acid sequence, similar CDR3 nucleotide sequence (junction) lengths, and the same V gene. We calculated percent identity using the Needleman-Wunsch algorithm, implemented through the Biopython library in Python, to align the amino acid sequences. Then, we developed the distance function, which heavily penalizes the distance between clonotypes with a different V gene or with a difference in length of the AA sequence greater than 3. The combination of these three criteria provided a robust, novel metric to compute distances in "biological" space between two clonotypes.

We ran this algorithm on approximately 36 million clonotype pairs [(~8000 × ~8000)÷2] to create a distance matrix. The matrix was a clonotype X clonotype file, providing a distance score for each comparison between 2 clonotypes. A distance of 0 means the clonotypes are identical; as the difference between the clonotypes increases, so does the distance. We used the DBSCAN clustering algorithm (an unsupervised ML algorithm) to cluster similar clonotypes into UCGs. We use DBSCAN as it is density-based and is compatible with non-standard distance functions, while alternatives (like k-means and k-medians) might require incompatible metrics such as Euclidean distance. Additionally, DBSCAN does not require us to define the number of clusters beforehand.

The clustered UCGs were transformed into a patient X UCGs matrix. In each cell, the ratio of the UCG for that patient is calculated using this formula:

$$ratio = \frac{reads\ of\ UCG\ in\ that\ patient}{reads\ of\ all\ UCGs\ in\ that\ patient}$$

**Equation 1:** Abundance ratio formula.

We use ratios instead of raw reads to normalize for potential study-based differences in sequence preparation (such as bulk versus single-cell sequencers or differing tissue origins). Further details for clustering parameters can be found in the supplementary manuscript file under 'S2' and 'S3'.

***Machine Learning to Identify Potentially Pathogenic UCGs:***

We devise an ML classification approach to assign weights to our features (UCGs). We use supervised ML algorithms such as Support Vector Machine (SVM), Random Forest, Decision Tree, Logistic Regression, and LightGBM to classify between MS and HC. Supervised ML predicts labels predefined by the user (in our case, MS and HC).[15] A stratified train-test split was run with three-fold cross-validation to ensure robust and stable results during hyperparameter tuning and to account for class imbalance. We generate a detailed classification report with performance scores for each model. A confusion matrix is generated for the highest-performing model (Random Forest).

Feature importances from the best-performing model are extracted. Higher importances relative to the other features entail that the UCG was highly influential to the model's classification decisions. The UCGs with the top 10 feature importances were shortlisted for further analysis. Specifications for classification can be found in section 'S4' of the supplementary manuscript file.

***Visualizations:***

Within each UCG, each B cell clonotype had a corresponding CDR3 amino acid sequence. Sequence logos for each UCG (representative of their clonotype CDR3 makeup) and consensus sequences were generated by i) aligning CDR3 sequences using MUSCLE,[16] ii) extracting sequences from generated FastAs using SeqIO from Biopython, and iii) generating sequence logos and consensus sequences using the logomaker module (see section 'S5' of the supplementary manuscript file). We identify the common parts of the V gene makeup in each UCG. Distribution pie charts for the clonotypes' V gene family, subgroup, and position were generated. Per UCG, we generate an enrichment vs observed bar graph. Values above 1 indicate that the class (healthy or MS) is over-represented in the cluster, values equal to 1 indicate that the class is expressed as expected, and values lower than 1 indicate that it is under-represented. To calculate this, we first calculate the expected count for each class and then calculate the enrichment value.

$$\text{A } Expected\ Count_{MS\ or\ HC} = (total\ clonotypes\ in\ UCG) \times \frac{total\ MS\ or\ HC\ clonotypes}{total\ clonotypes\ overall}$$

$$\text{B } Enriched = \frac{MS/HC\ count}{MS/HC\ expected}$$

**Equation 2:** Enrichment calculation: A. Expected count calculation, B. Enriched score calculation.

For each patient in which the UCG is detected, we generate a summed clonotype ratios histogram. Patients are color-coded in accordance with their disease status.

$$summed\ ratio = \frac{\sum_{UCG\ clonotypes\ in\ the\ patient} reads\ of\ clonotype}{\sum_{all\ clonotypes\ in\ the\ UCG} reads\ of\ clonotype}$$

**Equation 3:** Summed ratio formula.

***Statistical Analysis:***

A non-parametric Mann-Whitney U test is performed on the summed ratios for each MS patient in the UCG and those for each HC patient in the UCG. Empirical power estimation with 10,000 resamples is run for each test. Benjamini-Hochberg adjusted p-values are calculated for each UCG to adjust for false positives; these values adjust the raw p-values based on the number of hypotheses (in this case, the top 10 UCGs being tested for significance). UCGs with a p-value < 0.05, a power estimation between 0.800 and 1.000, and a Benjamini-Hochberg adjusted p-value < 0.05 are considered statistically significant. U-values from the Mann-Whitney test are extrapolated and normalized using this formula:

$$normalized\ U = \frac{U\ value}{n_1 \times n_2}$$

**Equation 4:** U-value normalization.

Here, $n_1$ and $n_2$ are the sizes of the MS and HC groups. Normalized U values ≥ 0 and ≤ 0.5 indicate HC skew, normalized U values = 0.5 show balanced data, and normalized U values ≥ 0.5 and ≤ 1.0 demonstrate an MS skew. Additional information can be found in section 'S6' of the supplementary manuscript file.

***Tertiary Analysis:***

***Genetic Variability:***

To understand whether genetic variability is correlated with UCG pathogenicity, UCGs with the top 20 feature importances and the bottom 20 feature importances with available sequence data were selected for comparison. Shannon entropy per position for the aligned CDR3 sequences per UCG was calculated using this formula:

$$Entropy = -\sum_{i=1}^{n} p_i \log_2 p_i$$

**Equation 5:** Shannon entropy formula.

Here, $i$ is indexing over all amino acids at that position and $p_i$ is the probability (frequency) of the amino acid 'i' being at that position. Shannon entropy is averaged across positions for each UCG. A Mann-Whitney U test is performed on the mean entropies of UCGs with high importances against those of UCGs with low importances. The results are illustrated on a box plot along with indicators of statistical significance (e.g., asterisks based on the Mann-Whitney p-value).

***Clonal Expansion:***

In addition to genetic variability, we wanted to learn whether clonal expansion in MS patients is correlated with UCG pathogenicity. Once more, UCGs with the top 20 and bottom 20 feature importances with available sequence data were chosen for comparison. Clonal expansion per MS patient was calculated as follows:

$$Expansion_{patient} = \frac{Total\ reads\ of\ that\ UCG\ in\ that\ patient}{All\ reads\ in\ that\ patient}$$

**Equation 6:** Expansion formula.

Expansion values for all MS patients are averaged per UCG. A Mann-Whitney U test is performed on the mean expansion values for the highest-performing UCGs against those of the lowest-performing UCGs.

## Results

***Segregation in Principal Component Analysis (PCA) Between Healthy and Diseased Classes:***

A qualitative analysis of the PCA plots reveals a promising segregation between the healthy and diseased classes, indicating a meaningful difference for classification (Figure 2). The relative homogeneity of data from different studies indicates little to no source bias. PC1, PC2, and PC3 explained 29.68%, 16.74%, and 6.51% of the total variance, respectively, based on their associated eigenvalues. They cumulatively account for 52.94% of the variance.

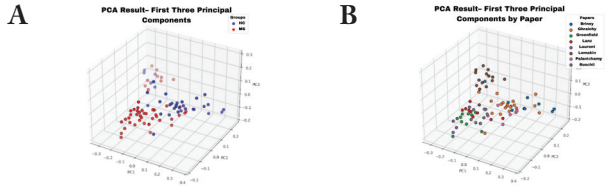


**Figure 2:** Principal Component Analysis (PCA) and ML accuracy. A. PCA colored by disease status shows clear separation between healthy and diseased classes. B. PCA colored by paper shows limited source-wise separation.


DOI: 10.36838/IJHSR816.35

***High ML accuracy during classification using Random Forest:***

The model with the highest weighted average F1-score was 87.79%; this was achieved by the random forest model. The confusion matrix for the random forest model (Figure 3) reveals 0 true negatives and only 2 false positives.

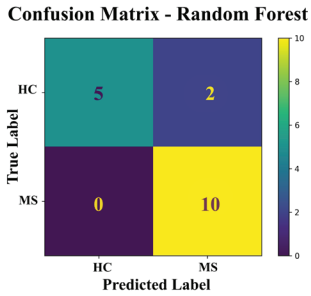


**Figure 3:** The random forest model confusion matrix. Outlines classification performance, showing 5 correctly predicted HC, 2 HC misclassified as MS, 0 MS misclassified as HC, and 10 correctly predicted MS.

***Identification of Statistically Significant UCGs:***

The classification model was trained on our unique clonotype groups – UCGs that contributed more to the model's decision between healthy and diseased were assigned higher feature importances. We extracted these importances to rank our UCGs. Figure 4 represents the importance of the 20 UCGs with the highest feature importances.

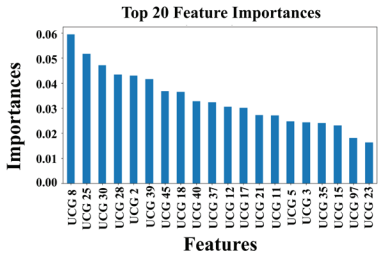


**Figure 4:** Top 20 feature importances extracted from the random forest model.

Table 1 summarizes the feature importances, Mann-Whitney p-values, Benjamini-Hochberg corrected p-values, and empirical power estimation values for the UCGs with the top 10 importance values.

**Table 1:** Comparison table of UCGs with top 10 feature importances. Statistically significant UCGs have a p-value < 0.05, an empirical power estimation >0.8, and a corrected p-value < 0.05. Statistically significant UCGs are highlighted in green.

| UCG Designation | Feature Importance | Mann-Whitney p-value (<0.05 is significant) | Empirical Power Estimation Value (>0.8 is significant) | Benjamini-Hochberg p-value (<0.05 is significant) |
|---|---|---|---|---|
| UCG 8 | 0.0595 | <0.0001 | 1.000 | <0.0001 |
| UCG 25 | 0.0516 | 0.0050 | 0.952 | 0.0125 |
| UCG 30 | 0.0471 | 0.3114 | 0.161 | 0.3460 |
| UCG 28 | 0.0434 | 0.4140 | 0.001 | 0.4140 |
| UCG 2 | 0.0430 | 0.0002 | 0.989 | 0.0007 |
| UCG 39 | 0.0417 | <0.0001 | 1.000 | <0.0001 |
| UCG 45 | 0.0367 | 0.0068 | 0.822 | 0.0127 |
| UCG 18 | 0.0366 | 0.0076 | 0.801 | 0.0127 |
| UCG 40 | 0.0327 | 0.1730 | 0.263 | 0.2163 |
| UCG 37 | 0.0324 | 0.1085 | 0.362 | 0.1550 |

Mann-Whitney p-values (and Benjamini-Hochberg corrected p-values) are calculated from the summed ratios for the MS and HC classes in each UCG. Empirical power estimation validates the accuracy of the p-value. From this table, we extrapolate 6 statistically significant and potentially pathogenic UCGs.

***Tertiary Analyses Reveal Genetic Variability and Clonal Expansion as Markers for Pathogenic Clonotype Groups:***

Box plots are generated from the mean Shannon entropy scores and mean expansion scores from the best- and worst-performing UCGs alongside statistical analyses. The genetic variability box plot (Figure 5a) reveals a statistically significant difference (p-value = 0.0004 [<0.0005]) between the mean Shannon entropy from potentially pathogenic UCGs and those with low feature importances. Comparing medians shows increased variability in the top 20 UCGs. In the clonal expansion box plot (Figure 5b), the median comparison and low Mann-Whitney p-value suggest that clonal expansion in MS patients is positively correlated with disease-causing effects.

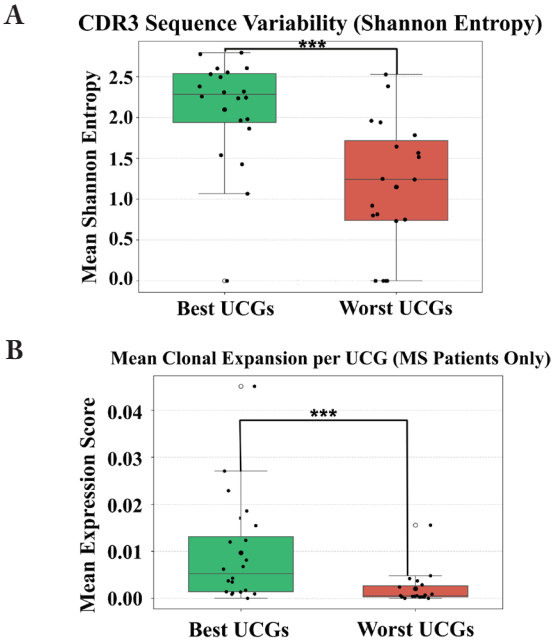


**Figure 5:** Tertiary analysis box plots comparing potentially pathogenic UCGs and those with low feature importances. A. The CDR3 sequence variability comparison plot. The green box plot represents the mean Shannon entropy for the top 20 UCGs, while the red box plot displays the same for the bottom 20 UCGs. A p-value of 0.0004 indicates a statistically significant difference between the two classes. B. The mean expansion values comparison plot. The green plot represents the top 20 UCGs, and the red plot represents the bottom 20 UCGs. A p-value of 0.0008 suggests a statistically significant difference between the data. A comparison of the medians reveals higher MS clonal expansion in the top 20 UCGs.

***Per-UCG Visualizations Display Genetic Characteristics and Abundance of Each UCG:***

Visualizations summarizing the relevant genetic characteristics of each UCG were generated. Figure 6 is a graphical representation of the components of these UCG visualizations.

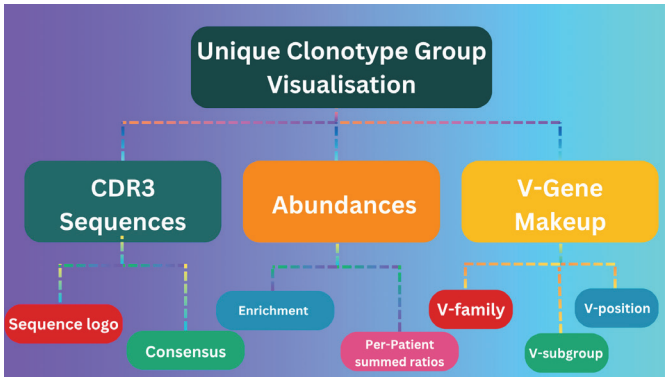


**Figure 6:** UCG visualization flow. A UCG visualization contains CDR3 sequence, UCG abundance, and V-gene makeup details.

The per-UCG visualization includes a sequence logo, an enrichment bar graph, a summed ratio (abundance) histogram, and V gene distribution pie charts. An example of such a visualization is seen below (Figure 7).

The sequence logo, consensus sequence, and V-gene charts can tell us information about the genetic makeup of the clonotype group. While both the enrichment bar graph and the summed ratios histogram display clonotype makeup, the noteworthy difference is that the bar graph simply uses the frequency (count) of healthy and MS clonotypes in its calculations, while the histogram takes into account the abundance (reads) of individual clonotypes. Importantly, the graphs are not displaying raw counts or reads; rather, they are displaying normalized values (as outlined in methods).

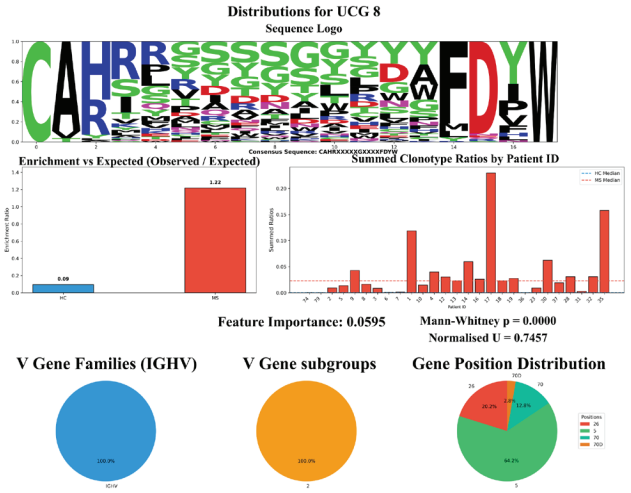


**Figure 7:** An example of a per-UCG visualization: UCG 8. The sequence logo represents conserved amino acids at each position across clonotypes. The enrichment bar graph presents the disease distribution of the overall clonotype makeup. The summed ratios represent UCG abundance per patient. The V gene pie charts display the V-gene makeup for the UCG. The Mann-Whitney p-value, normalized U value, and feature importance are displayed.

Among the six statistically significant UCGs, several genomic and abundance patterns emerged. UCG 8 demonstrated strong MS enrichment, with prominent clonal expansion in several MS patients; additionally, it is IGHV2-associated. UCG 25 is the most MS-exclusive UCG, with negligible HC enrichment. Similarly, UCG 2 and 39 possess a strong MS skew. UCG 45 is unique in that its HC enrichment exceeds MS enrichment despite higher MS abundance; it has a dominant IGHV4 signature. UCG 18 has an MS-associated abundance pattern. Overall, multiple UCGs showed conserved terminal motifs, variable CDR3 regions, and evidence of MS-associated clonal expansion.

Table 2 summarizes the data from these visualizations for the significant UCGs.

**Table 2:** Comparison table of the characteristics of the statistically significant UCGs. Refer to citation 5 for corroboration.[5]

| UCG | Sequence logo variability | Enrichment skew | U-value skew (<0.5 = HC; >0.5=MS) | Genomic family and subgroup | Gene position(s) |
|---|---|---|---|---|---|
| 8 | Variable | MS | MS | IGHV2 | 5, 26, 70D, 70 |
| 25 | Variable | MS | MS | IGHV1 | 2, 24, 3, 46, 69, 69D, 8 |
| 2 | Variable | MS | MS | IGHV5 | 51, 10-1 |
| 39 | Variable | MS | MS | IGHV1 | 18, 2, 24, 3, 46, 69, 69D, 8 |
| 45* | Variable | HC | MS | IGHV4 | 30-4, 31, 34, 38-2, 39, 4, 59, 61 |
| 18 | Variable | MS | MS | IGHV5 | 51, 10-1 |

## Discussion

Through a machine learning-mediated meta-analysis of B cell sequencing data paired with various statistical analyses, we discovered statistically significant characteristics of MS clonotype groups: genetic variability, through the amino acid sequence logos; clonal expansion, using normalized clonotype reads; and positional genomic data, such as the V region pie charts. To discuss the suitability of the supervised ML approach, we analyzed the PCA graphs, ML scores, and the confusion matrix.

First, by examining the 3D plot from Principal Component Analysis (PCA) (Figure 2), we visualized a separation between the healthy and diseased classes: an encouraging preliminary observation that stipulates a difference in the sequencing repertoire of the two classes that could be subject to further investigation through classification ML.

Next, the high f1-score of 87.79% and high confusion matrix-derived accuracy bolster the significance and reliability of the feature weights extracted from the random forest model and further exemplify that BCR sequencing data, and by extension, B cells, are a suitable source of data to investigate the cause of multiple sclerosis.

Understanding the broader characteristics shared by pathogenic UCGs is one that holds merit in drug design. Our tertiary analysis focused on sequence variability and clonal expansion as characteristics of pathogenic UCGs. In this section, we discussed their viability as candidates for further research based on our statistical tests on their significance in MS pathogenesis.

In order to ascertain the importance of sequence variability in MS pathogenesis, we compared the classes of the impactful UCGs with those of low importance. The low p-value of 0.0004 indicated that there is a statistically significant difference between the CDR3 sequence variability of the two classes. By comparing medians, we observed that an increase in potential pathogenicity is positively correlated with increased


DOI: 10.36838/IJHSR816.35

sequence variability. We hypothesized that this increased variability is a result of somatic hypermutation in diseased B cells, a characteristic which has been previously observed in the CSF of MS patients.[17] B cells may undergo somatic hypermutation to refine their BCRs and, by consequence, their autoantibodies, to increase affinity for the body's myelin sheath; this could significantly increase their pathogenic success. Accordingly, we can say that sequence variability is an important characteristic of pathogenic UCGs, and its role and cause in MS pathogenesis should be further investigated in future studies.

With clonal expansion, once more, we observed a significant p-value of 0.0008. Combined with a median comparison, this p-value allowed us to conclude that potentially pathogenic B cells showed a statistically significant increase in clonal expansion in contrast to those with lower feature importances. This is concurrent with research in the past, where increased clonal expansion for B cells in the CSF has been reported in MS patients.[18] We hypothesized that the increase in potentially pathogenic B cells from clonal expansion could exacerbate MS-consequent neurodegeneration, and, therefore, clonal expansion is another crucial characteristic of interest for further investigation.

Ultimately, based on this tertiary analysis, we believe that UCGs displaying increased MS clonal expansion and increased genetic variability should be flagged as those of interest for research on future therapeutic targets; additionally, MS clonal expansion and genetic variability can be identified as global characteristics of potentially pathogenic MS B cell unique clonotype groups. Research on identifying or disproving causal links between these global characteristics and MS pathogenesis is required.

Lastly, we conducted an in-depth analysis of the visualizations of our significant UCGs. From Table 2, we can see that these potentially pathogenic UCGs all mostly have a skew towards MS patients, except for UCG 45, which has a higher HC enrichment score. This is because the enrichment score is based on unique clonotype frequency counts, and we hypothesized that once the MS antigen is encountered, clonal expansion (increasing clonotype abundance) may be prioritized over making unique clonotypes through somatic hypermutation, decreasing the frequency count. This may explain why, while the MS enrichment score in UCG 45 may have been reduced, the total MS clonotype abundance is more significant (indicated by the U-value >0.5). Further investigation is warranted for this observation.

Previous research has demonstrated a VH4 bias in MS B cell clonotypes;[5] notice that, accordingly, UCG 45 has a VH4 region majority. Moreover, UCG 45 incorporates 5 VH4 regions, which have previously been associated with MS B cells in the CSF (IGHV4-31, IGHV4-34, VH4-39, IGHV4-59, and IGHV4-61), along with 3 other new gene positions. This established a strong confidence in the validity of our pipeline. Additionally, we observed a motif for certain parts of IGHV1 (IGHV1-2, 24, 3, 46, 69, 69D, 8) and IGHV5 (IGHV5-51 and 10-1) across UCGs, which highlighted their importance as regions of interest in future therapeutic target studies. Finally, we also identified parts of the IGHV2 gene that may be of interest in pathogenicity studies.

## Conclusion

This study presented a novel approach to identify potentially pathogenic B cell clonotypes using an ML-driven meta-analysis of high-throughput sequencing data, and we identified six UCGs of this nature correlated with MS for future therapeutic research. Through a combination of supervised and unsupervised ML, we identified the motifs ('global characteristics') of genetic variability and clonal expansion among these UCGs: traits of interest for future research on the role of B cells in MS pathogenesis. Our findings exemplify the utility of applying machine learning to immune sequencing data to extrapolate meaningful conclusions from complex, high-dimensional sequences. Corroboration with previous studies confirms the merit and biological relevance of our results, revealing parallels with previously reported IGHV4 family bias. The recurrent clonotype architectures and IGHV-associated patterns we identify in this study could serve as biomarkers for immune repertoire-based MS profiling. Our observed links between MS pathogenicity, clonal expansion, and sequence variability may help identify B-cell subgroups of interest for future therapeutic targeting. In conclusion, our study provides insight into B-cell involvement in MS pathogenesis and creates a novel computational framework for future studies analyzing the MS immune repertoire.

## Limitations and Future Steps

Our study is strengthened by our comparatively larger sample size than previous MS immune sequencing studies; however, correlating variables such as tissue origin and disease stage with B cells can produce valuable conclusions in the context of disease progression. Future research should prioritize further analysis into factors such as sequencing depth, age/sex, and tissue origin to discover important new insights about the role of B cells in MS pathogenesis while using rigorous normalization to limit the batch effect arising from data heterogeneity. Clinical verification and analysis of computational results are the next steps in developing B-cell-targeting therapeutics against MS.

## Acknowledgments

We thank Justine Shih for her contributions during the initial research phase. Additionally, we thank the Vidjil team for access to their platform.

## References

1. National Multiple Sclerosis Society. MS prevalence. https://www.nationalmssociety.org/about-the-society/who-we-are/research-we-fund/ms-prevalence (accessed March 21, 2026).
2. Mayo Clinic. Multiple sclerosis: Symptoms and causes. https://www.mayoclinic.org/diseases-conditions/multiple-sclerosis/symptoms-causes/syc-20350269 (accessed March 21, 2026).

3. Comi, G.; Bar-Or, A.; Lassmann, H.; Uccelli, A.; *et al.* Role of B cells in multiple sclerosis and related disorders. *Annals of Neurology* **2020**, *89* (1), 13–23. https://doi.org/10.1002/ana.25927.
4. Margoni, M.; Preziosa, P.; Filippi, M.; Rocca, M. A. Anti-CD20 therapies for multiple sclerosis: Current status and future perspectives. *Journal of Neurology 2021*, *269* (3), 1316–1334. https://doi.org/10.1007/s00415-021-10744-x.
5. Ruschil, C.; Kemmerer, C. L.; Beller, L.; Gabernet, G.; *et al.* Next generation sequencing of cerebrospinal fluid B cell repertoires in multiple sclerosis and other neuro-inflammatory diseases—A comprehensive review. *Diagnostics* **2021**, *11* (10), 1871. https://doi.org/10.3390/diagnostics11101871.
6. Briney, B.; Inderbitzin, A.; Joyce, C.; Burton, D. R.; *et al.* Commonality despite exceptional diversity in the baseline human antibody repertoire. *Nature* **2019**, *566* (7744), 393–397. https://doi.org/10.1038/s41586-019-0879-y.
7. Ghraichy, M.; Galson, J. D.; Kovaltsuk, A.; von Niederhäusern, V.; *et al.* Maturation of the human immunoglobulin heavy chain repertoire with age. *Frontiers in Immunology* **2020**, *11*, 1734. https://doi.org/10.3389/fimmu.2020.01734.
8. Greenfield, A. L.; Dandekar, R.; Ramesh, A.; Eggers, E. L.; *et al.* Longitudinally persistent cerebrospinal fluid B cells can resist treatment in multiple sclerosis. *JCI Insight* **2019**, *4* (6). https://doi.org/10.1172/jci.insight.126599.
9. Lanz, T. V.; Brewer, R. C.; Ho, P. P.; Moon, J.; *et al.* Clonally expanded B cells in multiple sclerosis bind EBV EBNA1 and GlialCAM. *Nature* **2022**, *603* (7900), 321–327. https://doi.org/10.1038/s41586-022-04432-7.
10. Laurent, S. A.; Strauli, N. B.; Eggers, E. L.; Wu, H.; *et al.* Effect of ocrelizumab on B- and T-cell receptor repertoire diversity in patients with relapsing multiple sclerosis from the randomized phase III OPERA trial. *Neurology: Neuroimmunology & Neuroinflammation* **2023**, *10* (4). https://doi.org/10.1212/nxi.0000000000200118.
11. Palanichamy, A.; Apeltsin, L.; Kuo, T. C.; Sirota, M.; *et al.* Immunoglobulin class-switched B cells form an active immune axis between CNS and periphery in multiple sclerosis. *Science Translational Medicine* **2014**, *6* (248), 248ra106. https://doi.org/10.1126/scitranslmed.3008930.
12. Ruschil, C.; Gabernet, G.; Kemmerer, C. L.; Jarboui, M. A.; *et al.* Cladribine treatment specifically affects peripheral blood memory B cell clones and clonal expansion in multiple sclerosis patients. *Frontiers in Immunology* **2023**, *14*, 1133967. https://doi.org/10.3389/fimmu.2023.1133967.
13. Lomakin, Y. A.; Zvyagin, I. V.; Ovchinnikova, L. A.; Kabilov, M. R.; *et al.* Deconvolution of B cell receptor repertoire in multiple sclerosis patients revealed a delay in tBreg maturation. *Frontiers in Immunology* **2022**, *13*, 803229. https://doi.org/10.3389/fimmu.2022.803229.
14. Sofou, E.; Vlachonikola, E.; Zaragoza-Infante, L.; Brüggemann, M.; *et al.* Clonotype definitions for immunogenetic studies: Proposals from the EuroClonality NGS working group. *Leukemia* **2023**, *37* (8), 1750–1752. https://doi.org/10.1038/s41375-023-01952-7.
15. Jiang, T.; Gradus, J. L.; Rosellini, A. J. Supervised machine learning: A brief primer. *Behavior Therapy* **2020**, *51* (5), 675–687. https://doi.org/10.1016/j.beth.2020.05.002.
16. Madeira, F.; Madhusoodanan, N.; Lee, J.; Eusebi, A.; *et al.* The EMBL-EBI job dispatcher sequence analysis tools framework in 2024. *Nucleic Acids Research* **2024**, *52* (W1), W521–W525. https://doi.org/10.1093/nar/gkae241.
17. Beltrán, E.; Obermeier, B.; Moser, M.; Coret, F.; *et al.* Intrathecal somatic hypermutation of IgM in multiple sclerosis and neuroinflammation. *Brain* **2014**, *137* (10), 2703–2714. https://doi.org/10.1093/brain/awu205.
18. Qin, Y.; Duquette, P.; Zhang, Y.; Talbot, P.; *et al.* Clonal expansion and somatic hypermutation of V(H) genes of B cells from cerebrospinal fluid in multiple sclerosis. *Journal of Clinical Investigation* **1998**, *102* (5), 1045–1050. https://doi.org/10.1172/jci3568.

## Authors

Ayaan Agarwal is an eleventh-grade high school student at the American School of Bombay, Mumbai, India. He has been interested in the biological sciences since eighth grade – his research projects have spanned topics such as neurodegenerative illness, autoimmune disease, and tuberculosis. He hopes to pursue microbiology in the future.

Aveer Singh is an eleventh-grade student at Vasant Valley School with a strong interest in environmental and data science. He uses data-driven research to explore sustainability, climate policy, and the relationship between science and society. He aims to develop practical solutions to environmental challenges and contribute to informed environmental decision-making.


DOI: 10.36838/IJHSR816.35