Supplementary data

Immune repertoire profiling reveals its clinical application potential and triggers for Neuromyelitis Optica Spectrum Disorders

Yu Miao et al.

eMethods

TCR data Processing

Due to the presence of inter-sample contamination in the form of sequences from other samples observed in some samples, we removed inter-sample contamination using the filtering steps which have been described in the previous study [1]. To further eliminate the effect of sequencing errors, we removed low-frequency sequences with fewer than three repeats per million reads. In the subsequent analysis, we excluded non-productive sequences containing stop codons, pseudogenes, or CDR3 regions with nucleotide sequences that were not multiples of three. If the identified CDR3 sequences were less than three amino acids, the starting amino acid of CDR3 was not cysteine, or if the last amino acid was not phenylalanine, the corresponding sequences were dropped.

Classification of diseases using V, J genes and NMOSD-TCRs

We classified the NMOSD and HC using V, J genes, and VJ-pairing features or NMOSD-TCRs features (unique clone number and the proportion of specific NMOSD-TCRs in all samples), respectively, and performed a ten-fold cross-validation based on the lasso model. The results of the ten folds were combined and used to plot ROC curves (pROC package), and the areas under the ROC curve were calculated and evaluated the classifier's performance.

CDR3 and antigenic epitope affinity prediction

Affinity prediction of CDR3 amino acid sequences of antigen-specific clones with T-cell epitopes of antigens was performed using ERGO [2]. The average affinity of each CDR3 amino acid sequence to all epitopes was used as the overall affinity of the CDR3.

CDR3 Sequence Similarity Analysis

The CDR3 amino acid sequences of TCR were string clustered using the “clusteringr” package, and amino acid sequences with Levenshtein <= 2 were clustered together. The network graph was visualized using the “igraph” package.
Viral protein sequence alignment

The protein sequence database of AQP4 was built using “makeblastdb”, and the protein sequences of all currently known human viruses ([http://www.virusite.org](http://www.virusite.org)) were blasted using blastp. Viral proteins with five consecutive peptides with identical amino acid sequences to AQP4 were selected as possible antigenic epitopes for antigenic mimicry with AQP4.

Differential gene analysis and co-expression network analysis

Differential gene analysis was performed on the count matrix of bulk RNAseq using Deseq2 [3], using $|\log_{2}FC| > 1$ and p.adjust < 0.05 as thresholds to filter differential genes.

Using the FPKM (Fragments Per Kilobase of transcript per Million mapped reads) matrix of all genes as input, co-expression network analysis was performed on the expression data of bulk RNAseq using the WGCNA [4] package, and co-expression modules were identified.

Hub Genetic Analysis

Hub genes of specific modules in WGCNA were identified using GS>0.2 and MM>0.8, and the hub genes for the PPI network were then identified using “cytohubba”.

Enrichment Analysis

The clusterProfiler [5] and kobas3.0 packages were used for GSEA and KEGG enrichment analysis, respectively. And iRegulon was used for transcription factor prediction.

Single-cell transcriptome data processing

Single-cell expression matrices of one NMOSD and three HC samples were downloaded from the GEO database (GSE133028). DoubletFinder [6] (version 2.0.3) was applied to infer and remove doublets (the expected doublet ratio for 10x Genomics is about 0.03). Cells with >15% of mitochondrial genes or <200 or > 6000 expressed genes were considered as doublets and removed, and we removed all ribosomal genes.

Batch effects were removed using the harmony [7] package, and then dimensional reduction and clustering were performed using Seurat 4.0.1 (resolution = 1, PC = 1:30) for single-cell expression data. Cell types were defined for clustered subgroups using SingleR and the classical cell markers. After removing subgroups with high expression of PPBP gene (platelet-associated marker) and low-quality subgroups, we re-clustered the Seurat objects and defined subgroups. The Findallmarker function was used to identify highly expressed genes in each subpopulation with the following parameters: log2FC > 0.25 and p_adj < 0.05.

Single-cell BCR data processing

We downloaded the raw BCR data from the SRA database and used cellranger [8] VDJ (v6.1.2) to
analyze the VDJ composition and frequency for each of the four samples of single-cell RNA seq.

**Antibody modeling, molecular docking and Affinity prediction**

We extracted the full-length amino acid sequences of the top 5 BCRs in NMOSD and HC where clonal amplification occurred and used “Abbodybuilder” [9] for antibody modeling. After modeling was completed, the structures of all cloned antibodies to the AQP4 protein were globally docked to the AQP4 protein (pdb: 3gd8) using “cluspro” [10]. The affinity magnitude between the structures of the first-ranked complexes after docking was calculated using prodigy [11]. The smaller the Gibbs free energy, the higher the affinity.

Subsequently, the CDR3 amino acid sequences of the heavy and light chains of clontype1 were used to align with the SAbDab [12] antibody database and ranked according to the degree of sequence similarity. The protein structure was aligned and calculated structural similarity (Root Mean Square Deviation, RMSD) on pymol (v2.3.2).

**HLA analysis**

We performed whole exome sequencing on 20 AQP4-IgG+ NMOSD patients and downloaded the raw fastq data of 203 samples of healthy Han individuals from the 1000 Genomes Project. After using cutadapter to remove the adapter from the raw fastq data, we aligned the clean reads to the hg38 reference genome using bwa-mem (v0.7.15). Next, we used the xHLA [13] software to perform HLA typing for all samples and identified the risk HLA genes associated with the AQP4-IgG+ NMOSD phenotype based on the typing results using pyHLA [14].

**Synthesis of CMV peptide pools**

We used the sequence of the CMV antigenic epitope with four consecutive amino acid sequences that successfully aligned to AQP4 as the core peptide (VDMV), and extended 20 amino acids upstream and downstream in the CMV protein sequence as the peptide that might generate the T cell epitope, respectively. This peptide sequence and the HLA risk genes consistent with published literature in this study (HLA-DPB1*05:01 and DRB1*03:01) were used as input to predict possible T cell epitopes using NetMHCIIpan-4.1. All of the T-cell epitopes containing VDMV were screened for peptide synthesis and combined to form a peptide pool.

**Flow cytometry analysis**

After stimulating and washing, PBMCs were incubated with Human TruStain FcX™ (Biolegend) on ice for 30 min to avoid Fc binding. We stained cells at 4° for 30min in the presence of fluorescent antibodies: FITC anti-human CD3, APC anti-human CD4, Apc-Cy7 anti-human CD8 from Biolegend, and PE anti-human CD154 from BD. FACS Canto II flow cytometer was used to obtain the original data and Flowjo v10 was employed to perform further analysis.