Integration of T Cell Repertoire, CyTOF, genotyping and symptomatology data reveals subphenotypic variability in COVID-19 Patients

Fernando Marín-Benesiu[1,2], Lucia Chica-Redecillas[1,2], Sergio Cuenca-López[3], Carmen Entrala-Bernal[3], Sara Martín-Esteban[4], Maria Jesús Alvarez-Cubero[1,2,5†*]. Luis Javier Martínez-González [1,2†],

[1]Department of Biochemistry, Molecular Biology III and Inmunology, Faculty of Medicine, University of Granada, Parque Tecnológico de la Salud, Granada 18016. Spain.

[2]GENYO, Centre for Genomics and Oncological Research: Pfizer, University of Granada, Andalusian Regional Government, Parque Tecnológico de la Salud, Granada, Spain 18016.

[3]LORGEN G.P., PT, Ciencias de la Salud - Business Innovation Centre (BIC), Armilla, 18100, Spain.

[4]Montefrío Health Center, Metropolitan District of Granada, Montefrío, 18009, Spain

[5]Ibs Granada, Biosanitary Research Institute of Granada, Granada, 18012, Spain.

*Correspondece: María Jesús Álvarez Cubero, PhD, Medicine Faculty, Av. de la Investigación nº 11, Tower C, 11th floor 18071. Granada (Spain). +34 958 243517 mjesusac@ugr.es.

† The authors have contributed equally to this work

Population of study

A total of 61 patients (33 mild and 28 severe) with a median age of 51 years who recovered from COVID-19 were included in this study (Table 1). All collected samples were confirmed in diagnosis of COVID-19 by RT-PCR and positive IgG serology. All of them follow-up inclusion criteria based on WHO classification. Inclusion criteria were revised periodically to update database trying to have balanced samples according to age, gender and severity. Those in mild disease were characterized by fever, malaise, cough, upper respiratory symptoms, and/or less common features of COVID-19 (headache, loss of taste or smell etc.). Moreover, patients in severe disease group fulfill the following features: (i) hypoxia: $SPO_2 \leq 93\%$ on atmospheric air or PaO_2:$FiO_2 < 300$ mmHg (SF ratio < 315); tachypnea: in respiratory distress or RR (respiratory rate) > 30 breaths/minutes; or more than 50% involvement seen on chest imaging (1). The study protocol was approved by the Granada Research Ethics Committee (CEI Granada) with internal code 1329-N-21. Written informed consent was obtained from all participants in accordance with the principles of the Helsinki Declaration

Data pre-processing.

The 61 patients included in this study were selected from larger cohorts where CyTOF (2), TCRseq (3), or genotyping of *TMPRSS2*, *ACE2*, and *MXI* SNPs (4) had been conducted. As a result, all 61 patients have these three main types of data. The dataset

consisted of 143 variables, including 129 quantitative metrics derived from TCRseq and CyTOF analyses, and 13 qualitative parameters related to symptomatology and genotyping. For clustering and multivariate analyses, quantitative variables were scaled using the z-score, while qualitative variables were factorized.

T cell repertories

T cell repertoire data were obtained from 61 patients in the TCRseq study and further analyzed in a larger cohort of 173 patients (3). RNA was extracted from whole blood and normalized to 1µg using the 'TempusTM Spin RNA Isolation Reagent' kit. High-throughput sequencing libraries were prepared using the AmpliSeq for Illumina Immune Repertoire Plus, TCR beta Panel, which efficiently converts RNA to cDNA. The NextSeq 500 platform (Illumina) with paired-end sequencing (150 bp x 2) was utilized to normalize and sequence the libraries, resulting in a final library depth of approximately 1 million reads per sample. MiXCR v4.0.0 (5) software was then employed to process the sequences, assemble clonotypes, determine V(D)J alleles, and clonotype frequency. To prevent excessive clonotype loss, the resulting repertoire files were normalized by downsampling to the depth of the smallest sample. Samples underwent analysis using VDJtools v1.2.1(6) and Immunarch v1.0.0 (7) to filter for quality and calculate diversity metrics (Chao1, Shannon-Wiener, Gini-Simpson), clonality (DE50, Gini coefficient), and proportions of clones based on the degree of clonal expansion (hyperexpanded, large, medium, small, and rare). The indices of diversity and clonality that are used are explained below:

Repertorie richness was calculated by Chao 1 estimator:

$$Chao1\ estimator = S_{obs}\frac{n_1^2}{2n_2}$$

Where S_obs is the observed number of species, n_1 is the number of singletons (clonotypes with count = 1), n_2 is the number of doubletons (clonotypes with count = 2) (8).

Repertoire diversity were measured using the normalized Shannon-Wiener index (SW) and the Gini-Simpson index:

$$Normalized\ Shannon - Wiener\ index = \frac{-\sum_{i=1}^{N} p_i\ log_b\ (p_i)}{\log\ (N)}$$

Where N is the number of unique clonotypes, and p_i is the frequency of clonotype i within a repertoire. Values range from 0 (minimal diversity with dominance of hyperexpanded clonotypes) to 1 (maximum diversity) (8, 9).

The Gini-Simpson index is derived from the Simpson index, representing the probability that two randomly selected clonotypes with replacement belong to the same clonotype. It is subtracted from 1 to orient high values toward high diversity. Values close to 0 indicate minimal diversity with high oligoclonality, while values close to 1 represent maximum diversity and polyclonality (8).

$$Gini - Simpson\ index = 1 - Simpson\ index = 1 - \sum_{i=1}^{N} p_i^2$$

The DE50 index is an indicator of clonality, representing the minimum number of clonotypes required to cover 50% of counts (read abundance). Low DE50 values indicate oligoclonality and dominance of hyperexpanded clonotypes, while high values suggest a more polyclonal clonotype distribution (9).

The Gini coefficient is also used to study repertoire heterogeneity based on clonality, quantifying the balance of a system based on equity in the frequency distribution of its

elements. Values range from 0 (all clonotypes are in equal proportions) to 1 (oligoclonal repertoires, with the proportion concentrated in one or very few sequences) (8)

$$Gini_c = \frac{\sum_{i=1}^{N}\sum_{j=1}^{N}|p_i - p_j|}{2N^2\bar{p}}$$

Where p_i and p_j are the frequencies of respective clonotypes i and j in the repertoire, $\bar{p}$ is the mean frequency of clones, and N is the total number of unique clonotypes.

Chao1, Gini-Simpson, DE50, and Gini metrics were calculated using the repDiversity() function of Immunarch, with the CDR3b amino acid sequence as the clonotype (argument .col=''aa''). The calculation of the normalized Shannon-Wiener index was performed using the CalcDiversityStats() function of VDJtools.

Additionally, weighted frequencies of V and J alleles were calculated for each sample. The Grouping of Lymphocyte Interactions by Paratope Hotspots 2 (GLIPH2) algorithm (10) was then utilized to cluster selected top 100 expanded clones per repertoire using complementarity-determining region 3 (CDR3b) amino acid sequences. The GLIPH2 clusters were filtered to identify COVID19-positive clonotypes by antigen binding validation available in the Multiplex Identification of Antigen-Specific T-Cell Receptors Assay (MIRA) dataset (11). CyTOF immune cell counts. In total, there are 10 diversity and clonality metrics, 48 V alleles, 13 J alleles, and 21 MIRA-filtered GLIPH2 motifs available for analysis.

CyTOF immune cell counts.

CyTOF metrics were obtained from the study (2). Blood samples collected for flow cytometry analysis were processed by fixing blood cells in Proteomic Stabilizer PROT1, frozen at -80°C until staining. The samples were then diluted, lysed, stained and fixed

before being pooled in a single tube. Surface antigens were stained using a pre-thawed antibody cocktail, followed by fixation and overnight DNA staining. After phenotyping, samples were frozen and thawed for CyTOF analysis using a 15-parameter monocyte and macrophage CyTOF panel on a Helios® mass cytometer. Samples were filtered through 35µm cell sieve cap tubes and resuspended in Maxpar Cell Acquisition Buffer with EQ beads. Acquisition was performed at a rate of 250–300 events per second using CyTOF software version 6.7.1016. For this study, 37 analyzed populations have been included, comprising 32 monocytic subpopulations, and characterization of CD3+ lymphocytes, B lymphocytes, NK cells, and CD45- cells.

Genotyping data

SNP data were obtained from the study and subsequent analysis of a larger cohort of 330 patients (4). Genomic DNA was extracted from whole blood samples using the RealPure 'SSS' kit from Durviz. DNA quantification was performed using the Qubit™ 3.0 fluorometer from Invitrogen™ by Thermo Scientific and the nanodrop 2000 system from Thermo Scientific, USA. The 260/280 ratio was checked as a quality control. The extracted DNA was stored at -20°C until genotyping. To select SNPs, we consulted data from NCBI's website on published articles related to COVID-19 and TMPRSS2 up to 2022. We selected SNPs with a minor allele frequency (MAF) higher than 10% in the Caucasian population from the Ensembl database. Specifically, we selected ACE2 (rs2285666), MX1 (rs469390), and TMPRSS2 (rs2070788) for the present analysis. We performed DNA genotyping using the TaqMan® Genotyping Master Mix from Applied Biosystems. The 7900HT Fast real-time PCR system from Applied Biosystems was used to perform allelic discrimination tests.

Symptomatology data

Clinical symptoms were recorded by their presence or absence, including dermatological involvement, anosmia, ageusia, myalgia, headache, fever, dyspnea, asthenia, systemic inflammatory response, and cough.

Unsupervised Mixed Data Clustering by Latent Class Model (LCM)

After scaling, the dataset was subjected to unsupervised clustering using the VarSelLCM package (12). VarSelLCM is an R package that enables complete model selection in model-based clustering, including the identification of relevant features and the selection of the number of clusters. The LCM protocol has a useful functionality for selecting relevant variables for cluster division, which is particularly beneficial given the large number of initial variables. The probability density function of the mixture distribution from LCM is:

$$f(x_i|m,\theta) = \prod_{j\in\Omega^c} h_j(x_{ij}|\alpha_{1j}) \sum_{g=1}^{G} \tau_g \prod_{j\in\Omega} h_j\left(x_{ij}|\alpha_{gj}\right)$$

Where $x_i = (x_{i1}, \ldots, x_{id})$ represents the observation value for variable *d*. *G* is the number of clusters and g is an individual cluster. Ω is the set of variables classified as relevant and $\Omega^c$is the set of variables that are not relevant. $\tau_g \epsilon$ (0,1] , is the proportion of cluster g. h_j denotes the Gaussian or multinomial distribution of quantitative or qualitative variables, respectively, with parameters α_{gj} (12).

In addition, LCM applies feature selection using the Bayesian Information Criterion (BIC), which penalises models with more parameters, encouraging the selection of simpler models that are less prone to overfitting (12).

To delimit an optimal range of clusters to test, we used an R script that iteratively adjusts variable selection models (VarSelCluster) for different numbers of clusters (gvals = 1 to 9). At each iteration, the BIC was calculated for the models, repeating the process 50 times for each number of clusters. The analysis identified the optimal number of clusters (from 1 to 4) by comparing the average BIC values for each model, visualised by box plots and statistical summary plots. To achieve a more comprehensive severity stratification, we constructed 100 LCM-BIC models with pre-defined clusters selected previously. The final model with highest adjusted Rand index (ARI) from the initial division by severity and the combined severity and age group was selected for further analysis. The ARI computes a similarity score by considering all pairs of samples and counting pairs that are assigned to the same or different clusters in both the predicted and true clustering. Its formula is as follows:

$$ARI = \frac{RI - expectedRI}{\max RI - expectedRI}$$

Where RI is the Rand Index:

$$RI = \frac{(a+b)}{(nC_2)}$$

In RI formula, a is the number of times a pair of elements belongs to the same cluster across two clustering methods. For b, the number of times a pair of elements belong to different clsuters across two clustering methods and nC_2 is the number of unordered pairs in a set of n elements. ARI index takes values between 0 (the two clustering methods do

not agree) to 1 (the two clustering methods perfectly agree regarding the clustering od every pair of elements). The selected LCM-BIC model has been saved as an .RData object for future use.

Factorial Analysis of Mixed Data (FAMD)

Two datasets were defined: the 'pre-LCM dataset' contained all variables and was not subjected to clustering by LCM-BIC, while the 'post-LCM dataset' resulted from applying LCM-BIC and selecting the top 70 variables. Both sets underwent multiple correspondence analysis (FAMD) using the FactoMineR and factoextra packages. Observations were plotted on a factorial plane with two principal dimensions to identify sample groupings in a lower-dimensional space. Eigenvalues and the percentage of explained variance were calculated for these dimensions. The contribution of the main variables to each of the two principal dimensions was evaluated by calculating the squared cosine (cos2) for the top 20 variables with the highest contribution percentage. The cos2 value represents the proportion of variance in the original variable that is explained by the factorial plane. A value close to 1 indicates a high representation and significant contribution of the variable to the factorial plane. Biplots were also generated for the quantitative variables to effectively study their variance and degree of correlation.

Deep Learning Analysis of TCR Repertoires

For a more comprehensive analysis of CDR3b sequences, we utilized the DeepTCR Python library (13). This package provides specialized tools for Deep Learning analysis of repertoires, which take CDR3 sequences, their corresponding V and J alleles, and HLA antigen type data as input. We applied Deep Learning analysis, both supervised and unsupervised, to the top 1000 expanded clonotypes from each sample in the post-LCM

set. For unsupervised analysis, we utilized the Variational Autoencoder (VAE) algorithm with the Train_VAE() function. The number of latent features that explain 99% of the total variance was selected to simplify the VAE model training process, and a compact size was chosen for the 3 convolutional networks.

To evaluate the VAE model's predictive ability, we utilized the K-Nearest Neighbor Repertoire Classifier (KNN_Repertoire_Classifier()) with Euclidean distance. We conducted 50 folds with a 75% training and 25% testing split. The area under the curve (AUC) values for each class were visually depicted using violin plots. For supervised analysis, Monte_Carlo_CrossVal() was used to model class pair comparisons. One hundred Monte Carlo simulations were performed with a 75% training and 25% testing split for cross-validation. The model architecture parameters consist of a reduced network size, with 12, 32, and 64 neurons for the respective layers. The model was trained using information from CDR3 amino acid sequences, V alleles, and J alleles with default parameters. To further improve the performance of the neural network and to speed up the training process, multi-sample dropout was enabled. The num_of_concepts hyperparameter was adjusted to 64 to effectively manage the heterogeneity between repertoires, To improve training speed and promote greater regularization, we confidently subsampled 100 sequences per repertoire in each fold. Furthermore, in our study, we adjusted the model using the weight_by_class argument to avoid bias due to differences in the number of samples per class.

To identify critical residues and distinctive structural motifs for each class, we selected those CDR3bs with higher predictive power (AUC $\geq$0.99) for each class in the generated supervised models. We then used the Residue Sensitivity Logos (RSLs) function to create logos for the top 25 predictors with the highest predictive power. Residue Sensitivity

Logos created logos for top 25 of those selected sequences. To detect significantly enriched structural motifs among the most relevant CDR3bs, we applied the Motif_Identification() function on the generated supervised model.

Detection of SARS-CoV-2 specific sequences

The CDR3 top sequences defined in the previous section were processed to find those catalogued as reactive to SARS-CoV2 and to study their specificity to virus antigens. The TCRmatch software was used for this purpose, using the default parameters (threshold=0.97). The Immune Epitope Database and Tools (IEDB) database was used as a reference. The distributions of total shared and SARS-CoV2-reactive sequences among the resulting clusters were represented via pie charts and upset-plots. The selected SARS-CoV2 reactive sequences were subsequently studied via the IEDB database to study the distribution of the specific antigens to which they bind. . Multiple Sequence Alignment (MSA) was performed using Clustal Omega alignment tool (14). After MSA, sequences were analyzed in Jalview v2.11.3.2 (15) for tree analysis and MSA results visualization. The trees were buily using de BLOSUM 62 matrix calculated from the MSA sequences. The resulting trees were finally processed with iTOL software tool (16)

Statistics

In the FAMD analyses, the PERMANOVA test was performed using the adonis2() function of the vegan package. 1000 permutations with Euclidean distance were performed. Post-hoc comparisons were conducted using the pairwise.perm.manova() function of the RVAideMemorie package, allowing 1000 permutations with Euclidean distance and p-value correction with Benjamini-Hochberg method. To compare qualitative variables of interest between groups, we used the chi-squared test, allowing

for 1000 permutations to calculate the p-value.To compare two groups with continuous variables, we used the Mann Whitney U test. For comparisons of two or more groups, we used the Kruskal-Wallis test. For post-hoc comparisons, we applied the Dunn's test with p-value corrected by the Benjamini-Hochberg method. Effect size coefficients were calculated for each test, including Crammer's V, rank-biserial (rrb) and rank-epsilon squared (E_R^2) for Chi-squared, Mann-Whitney and Kruskal-Wallis respectively. Two-tailed P-value less than 0.05 was considered significant. Analyses involving R packages were performed with R v 4.1.3.

References

1. Clinical management of severe acute respiratory infection (SARI) when COVID-19 disease is suspected: interim guidance, 13 March 2020.

2. Martínez-Diz,S., Marín-Benesiu,F., López-Torres,G., Santiago,O., Díaz-Cuéllar,J.F., Martín-Esteban,S., Cortés-Valverde,A.I., Arenas-Rodríguez,V., Cuenca-López,S., Porras-Quesada,P., *et al.* (2022) Relevance of TMPRSS2, CD163/CD206, and CD33 in clinical severity stratification of COVID-19. *Front Immunol*, **13**.

3. Marín-Benesiu,F., Chica-Redecillas,L., Arenas-Rodríguez,V., de Santiago,E., Martínez-Diz,S., López-Torres,G., Cortés-Valverde,A.I., Romero-Cachinero,C., Entrala-Bernal,C., Fernandez-Rosado,F.J., *et al.* (2024) The T-cell repertoire of Spanish patients with COVID-19 as a strategy to link T-cell characteristics to the severity of the disease. *Hum Genomics*, **18**.

4. Martinez-Diz,S., Morales-Álvarez,C.M., Garcia-Iglesias,Y., Guerrero-González,J.M., Romero-Cachinero,C., González-Cabezuelo,J.M., Fernandez-Rosado,F.J., Arenas-Rodríguez,V., Lopez-Cintas,R., Alvarez-Cubero,M.J., *et al.* (2023) Analyzing the role of ACE2, AR, MX1 and TMPRSS2 genetic markers for COVID-19 severity. *Hum Genomics*, **17**.

5. Bolotin,D.A., Poslavsky,S., Mitrophanov,I., Shugay,M., Mamedov,I.Z., Putintseva,E. V and Chudakov,D.M. (2015) MiXCR: software for comprehensive adaptive immunity profiling. *Nat Methods*, **12**, 380–381.

6. Shugay,M., Bagaev,D. V., Turchaninova,M.A., Bolotin,D.A., Britanova,O. V., Putintseva,E. V., Pogorelyy,M. V., Nazarov,V.I., Zvyagin,I. V., Kirgizova,V.I., *et al.* (2015) VDJtools: Unifying Post-analysis of T Cell Receptor Repertoires. *PLoS Comput Biol*, **11**.

7. Bioinformatics Analysis of T-Cell and B-Cell Immune Repertoires • immunarch.

8. Chiffelle,J., Genolet,R., Perez,M.A., Coukos,G., Zoete,V. and Harari,A. (2020) T-cell repertoire analysis and metrics of diversity and clonality. *Curr Opin Biotechnol*, **65**, 284–295.

9. Chaudhary,N. and Wesemann,D.R. (2018) Analyzing immunoglobulin repertoires. *Front Immunol*, **9**.

10. Huang,H., Wang,C., Rubelt,F., Scriba,T.J. and Davis,M.M. (2020) Analyzing the Mycobacterium tuberculosis immune response by T-cell receptor clustering with GLIPH2 and genome-wide antigen screening. *Nat Biotechnol*, **38**, 1194–1202.

11. Nolan,S., Vignali,M., Klinger,M., Dines,J.N., Kaplan,I.M., Svejnoha,E., Craft,T., Boland,K., Pesesky,M., Gittelman,R.M., *et al.* (2020) A large-scale database of T-cell receptor beta (TCRβ) sequences and binding associations from natural and synthetic exposure to SARS-CoV-2. *Res Sq*, 10.21203/rs.3.rs-51964/v1.

12. Marbac,M. and Sedki,M. (2019) VarSelLCM: an R/C++ package for variable selection in model-based clustering of mixed-data with missing values. *Bioinformatics*, **35**, 1255–1257.

13. Sidhom,J.W., Larman,H.B., Pardoll,D.M. and Baras,A.S. (2021) DeepTCR is a deep learning framework for revealing sequence concepts within T-cell repertoires. *Nat Commun*, **12**.

14. Sievers,F. and Higgins,D.G. (2018) Clustal Omega for making accurate alignments of many protein sequences. *Protein Science*, **27**, 135–145.

15. Waterhouse,A.M., Procter,J.B., Martin,D.M.A., Clamp,M. and Barton,G.J. (2009) Jalview Version 2-A multiple sequence alignment editor and analysis workbench. *Bioinformatics*, **25**, 1189–1191.

16. Letunic,I. and Bork,P. (2024) Interactive Tree of Life (iTOL) v6: recent updates to the phylogenetic tree display and annotation tool. *Nucleic Acids Res*, 10.1093/nar/gkae268.