

TESIS DOCTORAL

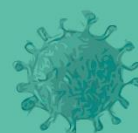
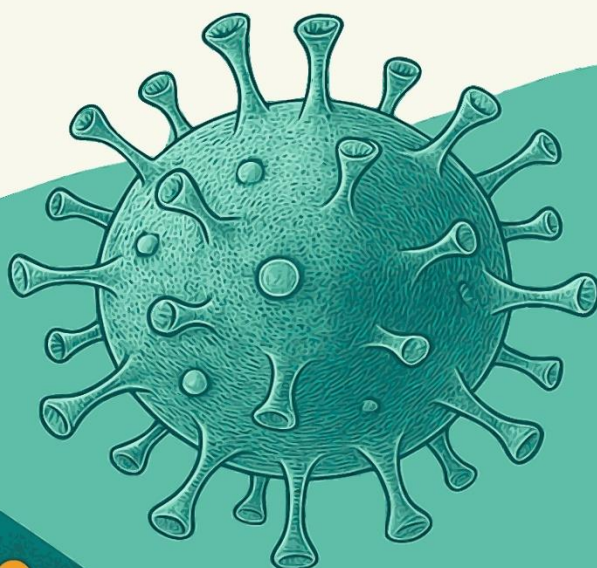
CARACTERIZACIÓN DEL REPERTORIO DE LINFOCITOS T EN COVID-19: UN ENFOQUE BIOINFORMÁTICO PARA LA ESTRATIFICACIÓN DE PACIENTES

Fernando Marín Benesiu

**Programa de doctorado en
Bioquímica y Biología Molecular**



**UNIVERSIDAD
DE GRANADA**



**Dirección de la tesis doctoral:
María Jesús Álvarez Cubero
Luis Javier Martínez González**



UNIVERSIDAD DE GRANADA

TESIS DOCTORAL

Programa de doctorado en bioquímica y biología molecular

(B.16.56.1)

CARACTERIZACIÓN DEL REPERTORIO DE LINFOCITOS T EN COVID-19: UN ENFOQUE BIOINFORMÁTICO PARA LA ESTRATIFICACIÓN DE PACIENTES

Doctorando

Fernando Marín Benesiu

Directores de Tesis

María Jesús Álvarez Cubero y Luis Javier Martínez González



CENTRO PFIZER-UNIVERSIDAD DE GRANADA-JUNTA DE ANDALUCÍA
DE **GENÓMICA E INVESTIGACIÓN ONCOLÓGICA**

Editor: Universidad de Granada. Tesis Doctorales
Autor: Fernando Marín Benesiu
ISBN: 978-84-1195-966-7
URI: <https://hdl.handle.net/10481/110569>

*“Como no estás experimentado en las cosas del mundo,
todas las cosas que tienen algo de dificultad te parecen imposibles.
Confía en el tiempo,
que suele dar dulces salidas
a muchas amargas dificultades.”*

Miguel de Cervantes

El doctorando / The *doctoral candidate* Fernando Marín Benesiu y los directores de la tesis / and the thesis supervisor/s: María Jesús Álvarez Cubero y Luis Javier Martínez González

Garantizamos, al firmar esta tesis doctoral, que el trabajo ha sido realizado por el doctorando bajo la dirección de los directores de la tesis y hasta donde nuestro conocimiento alcanza, en la realización del trabajo, se han respetado los derechos de otros autores a ser citados, cuando se han utilizado sus resultados o publicaciones.

/

Guarantee, by signing this doctoral thesis, that the work has been done by the doctoral candidate under the direction of the thesis supervisor/s and, as far as our knowledge reaches, in the performance of the work, the rights of other authors to be cited (when their results or publications have been used) have been respected.

Lugar y fecha / Place and date:

Granada a 10 de marzo de 2025

Director/es de la Tesis / *Thesis supervisor/s;*

Doctorando / *Doctoral candidate:*

María Jesús Álvarez Cubero

Fernando Marín Benesiu

Luis Javier Martínez González

Índice de contenido

1.Introducción	27
1.1 La COVID-19: Una pandemia sin precedentes.....	29
1.1.1 Breve introducción al virus SARS-Cov-2.....	31
1.1.2 Epidemiología de la COVID-19 en el mundo	34
1.1.3 Epidemiología de la COVID-19 en España	35
1.2 Linfocitos T: mediadores en la respuesta frente a la COVID-19 ..	37
1.2.1 Inmunidad adaptativa frente a la COVID-19: “Sensibilización celular sin seroconversión”.....	38
1.2.2 Linfocitos T y severidad de la COVID-19	42
1.2.3 Linfocitos T frente al COVID-19: influencia de la edad.	47
1.3 El receptor de linfocitos T: función, estructura y mecanismo de acción	48
1.3.1 Inmunorreceptores y el complejo TCR-pMHC	48
1.3.2 Estructura del TCR	49
1.3.4 La variabilidad de la región CDR3: la recombinación V(D)J. 53	
1.4 El repertorio de linfocitos T	55
1.4.1 ¿Qué son los clonotipos?	56
1.4.2 La diversidad estructural del TCR en números	56
1.4.3 Factores influyentes en la diversidad estructural.....	58
1.4.4 La importancia de los repertorios de TCR durante la enfermedad.	60
1.5 El uso de la inmunoinformática para estudiar el repertorio de linfocitos T	63
1.5.1 ¿Qué es la inmunoinformática?	63
1.5.2 El análisis del repertorio de linfocitos T (<i>TCRseq</i>)	64
1.5.2.1 <i>Modalidades</i>	65
1.5.2.2 <i>Principales técnicas de preparación de librerías de TCRseq</i>	67

1.5.2.3 Análisis primario de los repertorios de linfocitos T	68
1.5.2.4 Análisis secundario de los repertorios de linfocitos T	70
1.5.3 La combinación de <i>TCRseq</i> y <i>CyTOF</i> en favor de la medicina de precisión.....	73
2.Hipótesis y objetivos	77
2.1 Hipótesis.....	78
2.1 Objetivos.	79
3.Materiales y métodos	82
3.1 Metodología para el análisis del repertorio de linfocitos T.....	83
3.1.1 Población de estudio	83
3.1.2 Preparación de librerías y secuenciación de nueva generación (NGS).....	85
3.1.3 Preprocesamiento de datos	86
3.1.4 Análisis de diversidad y clonalidad	89
3.1.5 Análisis de solapamiento de repertorios	91
3.1.6 Análisis de uso de alelos V y J.....	94
3.1.7 Agrupamiento y selección de clonotipos mediante análisis de motivos	95
3.1.7 Análisis estadístico	99
3.1.8 Acceso a datos y código.....	100
3.2 Metodología para la integración de datos <i>CyTOF</i> , <i>TCRseq</i> , genotipado y sintomatología y posterior inferencia por <i>deep learning</i> de las secuencias CDR3b.	101
3.2.1 Población de estudio.....	101
3.2.2 Preprocesamiento de datos	101
3.2.2.1 Repertorios de células T	102
3.2.2.2 Recuento de células inmunitarias por <i>CyTOF</i>	102
3.2.2.3 Datos de genotipado.	103
3.2.2.3 Datos de sintomatología	104
3.2.3 Agrupamiento no supervisado de datos mixtos mediante modelo de clase latente (LCM)	104

3.2.4 Análisis factorial de datos mixtos (FAMD).....	107
3.2.5 Análisis de <i>deep learning</i> de Repertorios de TCR.....	108
3.2.5 Detección de Secuencias Específicas de SARS-CoV-2.....	110
3.2.6 Análisis estadísticos.....	110
3.2.7 Acceso a datos y código.....	111
4. Resultados (I). El análisis del repertorio de linfocitos T en pacientes COVID-19 revela cambios asociados con la severidad y la edad	114
4.1 Antecedentes.....	116
4.2. Asociaciones entre sintomatología y gravedad, edad y sexo del paciente.....	118
4.3 Cambios en la diversidad, clonalidad y solapamiento de repertorios entre grupos de pacientes.	119
4.3.1 Diversidad y clonalidad.	119
4.3.2 Solapamiento de repertorios	122
4.4. Uso de alelos V y J	124
4.4.1 Frecuencias alélicas ponderadas	124
4.4.2 Distribuciones generales.....	126
4.5 Inferencia sobre la secuencia CDR3b: agrupación de clonotipos mediante GLIPH2.	128
4.5.1 Distribución de los motivos estructurales GLIPH2 entre grupos de pacientes.....	128
4.5.2 Frecuencias de los motivos GLIPH2 con clonotipos reactivos al SARS-CoV-2.	128
4.5.3 Distribución de las secuencias CDR3b reactivas a SARS-CoV-2 entre repertorios.....	130
5. Resultados (II). La integración de datos de citometría, repertorio de TCR, genotipado y sintomatología revela nuevos subgrupos de pacientes COVID-19	134
5.1 Antecedentes	137
5.2 Resultados del análisis FAMD en los conjuntos pre-LCM y post-LCM.	139

5.2.1 Conjunto pre-LCM	139
5.2.2 Conjunto post-LCM.....	141
5.3 Distribución de las variables seleccionadas por LCM-BIC entre grupos de pacientes.	143
5.3.1 Descripción general	143
5.3.2 Distribución de la diversidad y la clonalidad	144
5.3.3 Distribución de los alelos V	146
5.3.4 Distribución de los motivos estructurales GLIPH2	146
5.3.5 Distribución de las subpoblaciones celulares por CyTOF	148
5.4 Detección de firmas en las secuencias CDR3b entre clústeres LCM.	148
5.4.1 <i>Deep learning</i> no supervisado.	149
5.4.2 <i>Deep learning</i> supervisado	149
5.4.3 Residuos de aminoácidos críticos en los CDR3b con mayor poder predictivo	150
5.5 Distribución de las secuencias en CDR3bs predictivas reactivas al SARS-CoV-2.	152
5.7 Creación de una base de datos interactiva.....	155
6. Discusión	159
6.1 Análisis del repertorio de linfocitos T	161
6.2 Integración de datos por medio de LCM.....	165
6.3 Limitaciones	169
6.4 Perspectivas de futuro	171
7. Conclusiones	174
8.Apéndice I. Figuras y Tablas suplementarias de Resultados (I)	179
9.Apéndice II. Figuras y tablas suplementarias de Resultados (II)	201
10.Apéndice III. Publicaciones científicas	212
Referencias.....	217

Índice de figuras

1. Introducción.

Figura 1.1. Esquema representando la estructura molecular y genética básica del SARS-CoV-2.....	29
Figura 1.2. Esquema con los principales grupos de linfocitos, citocinas secretadas y sus funciones durante la infección del SARS-CoV-2.	43
Figura 1.3. Estructura del TCR.....	49
Figura 1.4. Esquema básico ilustrativo de los cambios en el tamaño del repertorio de linfocitos T.....	58
Figura 1.5. Análisis posibles con <i>TCRseq</i>	68

4. Resultados (I). El análisis del repertorio de linfocitos T en pacientes COVID-19 revela cambios asociados con la severidad y la edad.

Figure 4.1 Resumen esquemático del primer estudio de esta tesis doctoral.	115
Figura 4.2. Resultados principales en los análisis de α -diversidad y clonalidad para el grupo combinado de edad y severidad	119
Figura 4.3. Resultados principales en los análisis de expansión clonal para el grupo combinado de edad y severidad.	121
Figura 4.4. Frecuencias significativas de TRBV entre el grupo combinado de edad y severidad	123
Figura 4.5. Heatmaps con agrupamiento jerárquico (distancia Euclídea) del número total de pacientes con el número total de alelos V (A) y alelos J (B) detectados.....	125
Figura 4.6 Principales resultados del análisis GLIPH2.	127
Figura 4.7 Boxplots y violinplots de los 4 clusters identificados por GLIPH2 después de seleccionar y filtrar las secuencias CDR3b COVID + con cambios significativos.....	129
Figura 4.8. Logos de secuencias de las CDR3b COVID+ de cada clúster GLIPH2 mostrado en 4.7.	130

5. Resultados (II). La integración de datos de citometría, repertorio de TCR, genotipado y sintomatología revela nuevos subgrupos de pacientes COVID-19.

Figure 5.1 Resumen esquemático del segundo estudio de esta tesis doctoral.	136
Figura 5.2. Resultados de FAMD para los conjuntos de datos pre y post-LCM.....	138
Figura 5.3. Resultados de FAMD para los conjuntos de datos post-LCM y agrupados	140
Figura 5.4. Resultados de FAMD para los conjuntos de datos post-LCM y agrupados.	143
Figura 5.5. Principales resultados de la tipología y distribución de variables seleccionadas por el modelo LCM-BIC	145
Figura 5.6 Principales resultados de los análisis de aprendizaje profundo supervisado utilizando DeepTCR	148
Figura 5.7. Principales resultados del estudio de residuos críticos en los CDR3b más predictivos	149
Figura 5.8. Secuencias del clúster 1 reactivas a SARS-CoV-2	152

8.Apéndice I. Figuras y Tablas suplementarias de Resultados (I).

Figura suplementaria 8.1. Recuento de clonotipos leídos (profundidad del repertorio)	180
Figura suplementaria 8.2. Principales resultados de los análisis de α -diversidad y clonalidad	181
Figura suplementaria 8.3. Distribución de frecuencias en % de clones raros, medianos, grandes e hiperexpandidos en el total de los repertorios	182
Figura suplementaria 8.4. Superposición del repertorio con el número de clonotipos.....	192
Figura suplementaria 8.5. Índices promedio de Jaccard, Morisita y recuento de repertorios en los diferentes grupos de pacientes	193
Figura suplementaria 8.6. Índices promedio de Jaccard, Morisita y recuento de repertorios en los diferentes grupos de pacientes	193

Figura suplementaria 8.7. Heatmaps con los valores de Jentsen-Shannon de los alelos V y J 194

Figura suplementaria 8.8. Distribución de frecuencias por repertorio/muestra de cada secuencia CDR3b COVID+ 195

9. Apéndice II. Figuras y tablas suplementarias de Resultados (II).

Figura suplementaria 9.1 Resultados de la FMAD en los conjuntos pre-LCM y post-LCM. 203

Figura suplementaria 9.2. Gráficos FAMAD..... 204

Figura suplementaria 9.3. Gráfico FAMAD para la reducción de la dimensionalidad de las observaciones etiquetadas por su genotipo en los tres SNP estudiados..... 204

Figura suplementaria 9.4. Gráfico de contribución de las variables cualitativas en el conjunto de datos pre-FAMAD..... 205

Figura suplementaria 9.5. Principales resultados del análisis no supervisado utilizando la Autocodificación Variacional (VAE) de DeepTCR 206

Figura suplementaria 9.6. Curva AUC-ROC de DeepTCR del modelo de *deep learning* supervisado para comparar muestras divididas por casos leves o graves..... 206

Figura suplementaria 9.7. Distribución del número de secuencias CDR3b seleccionadas..... 207

Figura suplementaria 9.8. Distribución de frecuencias en el uso de antígenos CDR3b seleccionados reactivos al SARS-Cov2 207

Figura suplementaria 9.9. Características de la secuencia de aminoácidos de las CDR3 reactivas y comunes del SARS-CoV2 208

Índice de abreviaturas

ACE2	Enzima Conversora de la Angiotensina 2 (<i>Angiotensin Converter Enzyme 2</i>)
ADN	Ácido Desoxirribonucleico
AIC	Criterio de Información de Akaike (<i>Akaike Information Criterion</i>)
AIRE	Regulador Autoimmune (<i>Autoimmune Regulator</i>)
ANOVA	Análisis de la Varianza (<i>Analysis of Variance</i>)
APCs	Células Presentadoras de Antígeno (<i>Antigen Presentation Cells</i>)
ARI	Índice Ajustado de Rand (<i>Adjusted Rand Index</i>)
ARN	Ácido Ribonucleico
AUC	Área Bajo la Curva (<i>Area Under de Curve</i>)
BCR	Receptor de Linfocitos B (<i>B cell Receptor</i>)
bulkTCRseq	Secuenciación de análisis del TCR (transcriptómica) a partir de una mezcla de células
CCR9	<i>C-C Motif Chemokine Receptor 9</i>
CD3	T-Cell Receptor T3 Glycoprotein
CD4	T-Cell Surface Glycoprotein CD4
CD8	T-Cell Surface Glycoprotein CD8
CyTOF	Citometría por tiempo de vuelo (<i>Time Of Flight mass cytometry</i>)
CXCR6	<i>C-X-C Motif Chemokine Receptor 6</i>
ECD	Dominios Extracelulares (<i>Extracellular Domains</i>)
FAMD	Análisis Factorial de Datos Mixtos (<i>Factorial Analysis of Mixed Data</i>)
FiO₂	Fracción inspirada de oxígeno.
GADS	<i>GRB2 related adaptor protein downstream of Shc</i>
GLIPH2	<i>Grouping of Lymphocyte Interactions by Paratope Hotspots 2</i>
GWAS	Estudio de asociación del genoma completo (<i>Genome-Wide Association Study</i>)

HLA	Antígeno Leucocitario Humano (<i>Human Leukocyte Antigen</i>)
HMGB1/2	<i>High Mobility Group Box 1/2</i>
IEDB	<i>The Immune Epitope Database</i>
IGV/D/J	Alelo V/D/J de la inmunoglobulina
INF-1	Interferón 1 (<i>Type 1 Interferon</i>)
ITAMS	Motivos de Activación Basados en Tirosina (<i>Immunoreceptor Tyrosine-based Activation Motifs</i>)
KNN	<i>K-Nearest Neighbourhood</i>
LCA	Análisis de Clases Latentes (<i>Latent Class Analysis</i>)
Lck	<i>Lymphocyte Cell-Specific Protein-Tyrosine Kinase</i>
MERS-CoV	Coronavirus del Síndrome Respiratorio del Medio Oriente (<i>Middle East Respiratory Syndrome Coronavirus</i>)
MHC	Complejo Mayor de Histocompatibilidad (<i>Major Histocompatibility Complex</i>)
MIRA	<i>Multiplex Identification of T cell Receptor Antigen Specificity</i>
MSA	Alineamiento Múltiple de Secuencias (<i>Multiple Sequence Alignment</i>)
Nck	<i>Non-catalytic region of tyrosine kinase adaptor protein 1</i>
NCM	Monocitos No Clásicos (<i>Non Classical Monocytes</i>)
NHEJ	Unión No Homóloga de Extremos
OAS1	<i>2'-5'-Oligoadenylate Synthetase 1</i>
OMS	Organización Mundial de la Salud (<i>World Health Organization</i>)
ORF	Pauta de Lectura Abierta (<i>Open Reading Frame</i>)
PaO₂	Presión parcial de oxígeno arterial.
PCR	Reacción en Cadena de la Polimerasa (<i>Polymerase Chain Reaction</i>)
PLCγ1	Fosfolipasa C gamma 1 (<i>Phospholipase C Gamma 1</i>)
pMHC	Complejo péptido-MHC
RACE	Amplificación Rápida de Extremos de cADN (<i>Rapid Amplification of cDNA Ends</i>)
RAG1/2	Recombinasa 1/2 (<i>Recombination Activating 1/2</i>)

RBD	Dominio de Unión al Receptor (<i>Receptor Binding Domain</i>)
RNA-seq	Secuenciación de análisis del ARN (Transcriptómica)
RSL	Logos de Sensibilidad de Residuos (<i>Residue Sensitivity Logo</i>)
RSS	Secuencias señal de la Recombinación (<i>Recombination Signal Sequences</i>)
RT-PCR	Reacción en Cadena de la Polimerasa con Transcriptasa Reversa (<i>Reverse Transcription Polymerase Chain Reaction</i>)
RUNX3	<i>Runt-Related Transcription Factor 3</i>
SARS-CoV	Coronavirus del Síndrome Respiratorio Agudo Severo (<i>Coronavirus del Síndrome Respiratorio Agudo Severo</i>)
scRNAseq	Secuenciación de análisis del ARN (transcriptómica) de célula única (<i>single cell</i>)
scTCRseq	Secuenciación de análisis del TCR y del repertorio de linfocitos T (transcriptómica) de célula única (<i>single cell</i>)
SLP76	<i>SH2-domain-containing leukocyte protein of 76 kDa</i>
SNPs	Polimorfismos de un Único Nucleótido (<i>Single Nucleotide Polymorphisms</i>)
SOP₂	Saturación de oxígeno periférica
T CD4/8	Glicoproteína de superficie CD4/CD8 (<i>T-Cell Surface Glycoprotein CD4/CD8</i>)
TCR	Receptor de Linfocitos T (<i>T Cell Receptor</i>)
TCRseq	Secuenciación y análisis de la secuencia del TCR y del repertorio de linfocitos T
TdT	Desoxinucleotidil Transferasa Terminal (<i>Terminal deoxynucleotidyl Transferase</i>)
Tfh	Linfocito T Cooperador Folicular (<i>Follicular Helper T cell</i>)
Th1/2/17	Linfocito Cooperador 1/2/17 (<i>Helper T cell type 1/2/17</i>)
ThPOK	<i>Zinc finger and BTB domain containing 7B</i>
TM	Monocitos Transicionales (<i>Transitional Monocytes</i>)
TMPRSS2	Proteasa Transmembrana de la Serina 2 (<i>Serine Transmembrane Protease 2</i>)
TRAV/J	Alelo V o J de la cadena alfa del TCR
TRBV/J	Alelo V o J de la cadena beta del TCR
UMIs	(Identificadores Moleculares Únicos (<i>Unique Molecular Identifiers</i>))

VAE	Autocodificadores Variacionales (<i>Variational Autoencoders</i>)
Vav1	guanine nucleotide exchange factor
VOC	Variante Preocupante (<i>Variants Of Concern</i>)
ZAP-70	<i>Zeta Chain Of T Cell Receptor Associated Protein Kinase 70</i>

Resumen

La COVID-19, enfermedad provocada por el SARS-CoV-2, ha causado un evento pandémico sin precedentes. Una de las características clínicas más relevantes de la COVID-19 es su amplio espectro sintomatológico, que abarca desde un desarrollo asintomático de la enfermedad hasta cuadros críticos y fatales. Por esta razón, la investigación clínica y biomédica ha centrado esfuerzos en el desarrollo de nuevos tratamientos farmacológicos, vacunas y en encontrar biomarcadores inmunológicos capaces de no solo estratificar a pacientes, sino de anticiparse al desarrollo grave de la COVID-19. En ese contexto la función de los linfocitos T es esencial.

Los linfocitos T actúan directamente sobre el virus y coordinando al resto de células del sistema inmunitario y se caracterizan por la persistencia en la respuesta durante la infección y la versatilidad en reaccionar adecuadamente frente a diferentes variantes del SARS-CoV-2. Estas propiedades se deben, entre otros factores al gran número de linfocitos T distintos si se toma como referencia la alta variabilidad de la secuencia de aminoácidos de la región determinante de la complementariedad 3(CDR3) del Receptor de Linfocitos T (TCR), responsable del reconocimiento de antígenos. El conjunto de TCRs distintos y funcionales presentes en un individuo se lo conoce como repertorio de TCR o de linfocitos T y cambios en sus dinámicas poblacionales están estrechamente asociadas con el pronóstico de la enfermedad. Gracias al auge de la inmunoinformática, la caracterización del repertorio de linfocitos T puede abordarse de forma rápida y precisa. Así, en esta tesis doctoral se ha estudiado el repertorio de linfocitos T en pacientes COVID-19 no vacunados, con el objetivo de lograr una estratificación robusta entre pacientes con distinto pronóstico. Para ello se ha realizado un abordaje bioinformático en diferentes modalidades (dinámicas poblacionales, agrupamiento de secuencias CDR3 e identificación de subfenotipos mediante la integración de datos), los cuales han sido divididos en dos estudios.

En el primer estudio se caracterizó al repertorio de linfocitos T mediante la secuenciación y posterior análisis bioinformático del TCR (*TCRseq*) sobre una cohorte de nueva creación de 173 pacientes, incluyendo 98 casos leves y 75 graves y con una edad media de 53 años. Específicamente, se amplificó y secuenció la región determinante de complementariedad 3 de la cadena β del TCR (CDR3b) y se realizaron análisis bioinformáticos para evaluar la diversidad del repertorio, la clonalidad y el uso de alelos V/J según edad, sexo y gravedad.

Los resultados mostraron una reducción significativa en la diversidad del repertorio TCR y un aumento en la expansión clonal en pacientes graves menores de 55 años, reflejando tendencias similares a las observadas en pacientes mayores de 55 años, tanto leves como graves. Por otro lado, los análisis de frecuencia de alelos V y J identificaron una reducción en el uso de alelos V clave (TRBV14, TRBV19, TRBV15 y TRBV6-4) asociados con la gravedad de la enfermedad, y se observó que los pacientes graves menores de 55 años presentaban patrones alélicos similares a los de los mayores de 55, junto con una frecuencia sesgada de motivos estructurales reactivos a antígenos del SARS-CoV-2. Por tanto, los resultados obtenidos del primer estudio sugieren que los pacientes graves menores de 55 años podrían tener un repertorio de linfocitos T comprometido, propio al de una persona mayor, lo que contribuiría a un peor pronóstico.

El segundo estudio integró la información obtenida del análisis *TCRseq* con información sobre la respuesta inmunológica innata (conteo de subpoblaciones de monocitos y macrófagos mediante citometría de masas), genotipos de Polimorfismos de Nucleótido Unico (SNPs) asociados a la infección por SARS-CoV-2 y datos sintomatológicos. Se analizó una subcohorte de 61 pacientes (33 leves y 28 graves) que contaban con estos datos junto con los obtenidos mediante *TCRseq*, con el objetivo de identificar subgrupos de pacientes (subfenotipos) no identificados por el criterio clínico inicial, y así lograr una estratificación robusta de los pacientes. Para ello, utilizamos el algoritmo Latent Class Model con Bayesian Information Criterion (LCM-BIC). Además, se aplicaron técnicas de

aprendizaje profundo (*deep learning*) para examinar las firmas de aminoácidos con potencial valor pronóstico en la región CDR3b de la cadena β del TCR.

Los resultados identificaron tres subgrupos de pacientes (Clusters 1, 2 y 3) a partir de 70 variables relevantes seleccionadas por BIC, con diferencias significativas en la clasificación de casos leves y graves. El Clúster 1 mostró un mayor predominio de casos graves, mientras que los clusters 2 y 3 presentaban distribuciones diferenciadas en variables clave como subpoblaciones de monocitos, diversidad del TCR, uso de alelos V y motivos de secuencia. El análisis de *deep learning* permitió distinguir con precisión el Clúster 1 de los otros dos, destacando la presencia de un gran número de secuencias CDR3b con alto valor predictivo, reactivas a SARS-CoV-2 y asociadas a distintos niveles de gravedad de la enfermedad. Esta capacidad de distinción no se observó cuando las muestras fueron divididas entre leves y graves, siguiendo la clasificación inicial.

Con todo ello, en esta tesis doctoral se ha conseguido por un lado dar una descripción de la arquitectura del repertorio de linfocitos T en una cohorte representativa, teniendo en cuenta la complejidad y la alta variabilidad de repertorios entre muestras. Por otro lado, también se ha conseguido abordar una estratificación de pacientes COVID-19 robusta, capaz detectar secuencias hipervariables del TCR reactivas al SARS-CoV-2 con potencial valor en la reclasificación de pacientes. Aun cuando es necesario realizar futuros estudios que confirmen los hallazgos a nivel bioinformático y laboratorio, este trabajo constituye un punto de partida a nuevas aproximaciones que consigan un mejor entendimiento de la respuesta inmunológica de la COVID-19 y de patologías futuras.

1.Introducción

1.1 La COVID-19: Una pandemia sin precedentes.

El 31 de diciembre de 2019, la Organización Mundial de la Salud (OMS) es informada de la irrupción de una neumonía que afectaba a una proporción preocupante de habitantes de la ciudad china de Wuhan. Los posteriores análisis metagenómicos de fluidos broncoalveolares en pacientes graves de esta neumonía, dan como resultado que el patógeno responsable es un tipo de coronavirus, similar al Coronavirus del Síndrome Respiratorio Agudo Grave (SARS-CoV). Este virus, junto con el Coronavirus del Síndrome respiratorio de Oriente Medio (MERS-CoV) ya habían causado brotes de infección graves a inicios de este siglo. Ante la expansión de este nuevo patógeno entre los habitantes de la ciudad, el 30 de enero la OMS declara este nuevo brote como una emergencia sanitaria. El 11 de febrero, con Wuhan y prácticamente toda China sometida a estrictas medidas de control sanitario, el Comité Internacional de Taxonomía de Virus nombra por primera vez dos términos que ya forman parte de la historia reciente de la humanidad en todos sus aspectos. Al nuevo virus se le denominó SARS-CoV-2 dada la alta similitud con el anterior, y a la enfermedad que provocaba: COVID-19.

La irrupción del SARS-CoV-2 en la ciudad de Wuhan por causas zoonóticas y su rápido contagio entre personas por medio de aerosoles, expande la COVID-19 por el mundo. El 11 de marzo de 2020, la OMS declara oficialmente a esta enfermedad como pandemia. Cinco años después, la COVID-19 ha dejado tras de sí centenares de millones de contagios que han causado más de 7 millones de vidas perdidas a nivel global, de las cuales aproximadamente más de 150000 fueron en España. Consecuencia de ello, la pandemia provocó sistemas sanitarios colapsados, medidas de cuarentena, crisis económicas y el desarrollo de uno de los procesos de creación y posterior campaña de vacunación más rápido y amplio de la historia. El SARS-CoV-2 sigue a día de hoy infectando y, pese a que muchos de nosotros la sufrimos de forma leve, sigue habiendo muertes y sigue habiendo personas que desarrollan síntomas que perduran durante meses y que les condiciona su vida cotidiana (COVID-19 persistente). Pese a que la investigación biomédica en este campo tomó un impulso como nunca se había visto con otra patología, muchas preguntas quedan por responder aún. Si todos reconocemos esta complejidad subyacente de la COVID-19, ¿Qué hay de la aún más compleja respuesta inmunitaria que hay detrás?

1.1.1 Breve introducción al virus SARS-Cov-2

El virus del SARS-CoV-2 se clasifica en la familia Coronaviridae, dentro del género Betacoronavirus y en el subgénero Sarbecovirus [1–4]. Los Coronaviridae son virus cuyo genoma es de ARN de sentido positivo y envueltos por una membrana lipídica [1,4] ([Figura 1.1 A](#)). En dicha membrana se sitúan las proteínas de espiga o espícula (S), las cuales dan al virus la característica forma de corona [5]. Concretamente, el genoma del SARS-CoV-2 posee un tamaño cercano a las 30 kilobases [3,6], con una similitud de aproximadamente el 80% al genoma del SARS-CoV [4,7,8].

El genoma del SARS-CoV-2 posee seis pautas de lectura abiertas funcionales: replicasa (Pautas de lectura abiertas: ORF1a/ORF1b), y las proteínas estructurales de espiga (S), membrana (M) nucleocápside (N) y envoltura (E) [3,4,8] ([Figura 1.1 B](#)). Además, cuenta con otras siete pautas de lectura abiertas codificantes de proteínas accesorias. La replicasa es una proteína compleja cuyo ORF cubre gran parte del genoma del virus. Codifica para 16 proteínas no estructurales implicadas en las funciones de transcripción y replicación vírica [4,8,9]. Por su parte, la proteína S es una de las más importantes tanto en la biología del virus como en el cambio de patogenicidad, surgimientos de nuevas variantes y el desarrollo de vacunas [4,5]. Se trata de un homotrímero, y cada monómero está compuesto por

las subunidades S1 y S2 [3,4]. Para entrar a las células endoteliales del tracto respiratorio, la proteína S se acopla al receptor de la Enzima Conversora de la Angiotensina 2 (ACE2) a través del Dominio de Unión al Receptor (RBD) de S. Para completar su internalización dentro de la célula hospedadora, es necesaria la acción de la Proteasa Transmembrana de la Serina 2 (TMPRSS2) sobre S2 para promover la fusión de membranas y la liberación de genoma vírico [9,10].

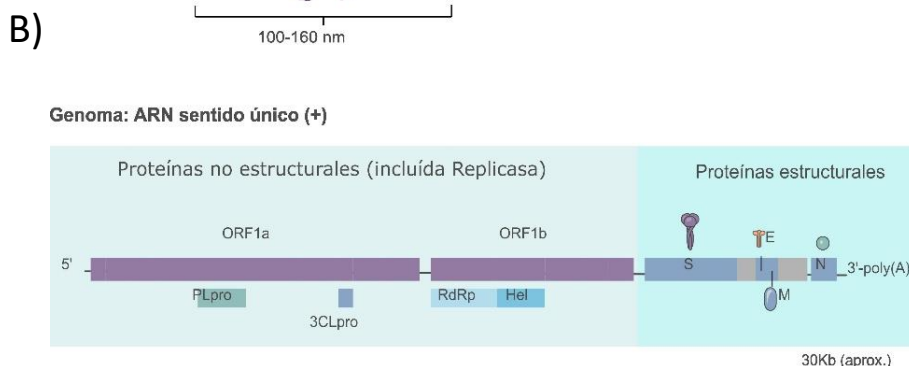
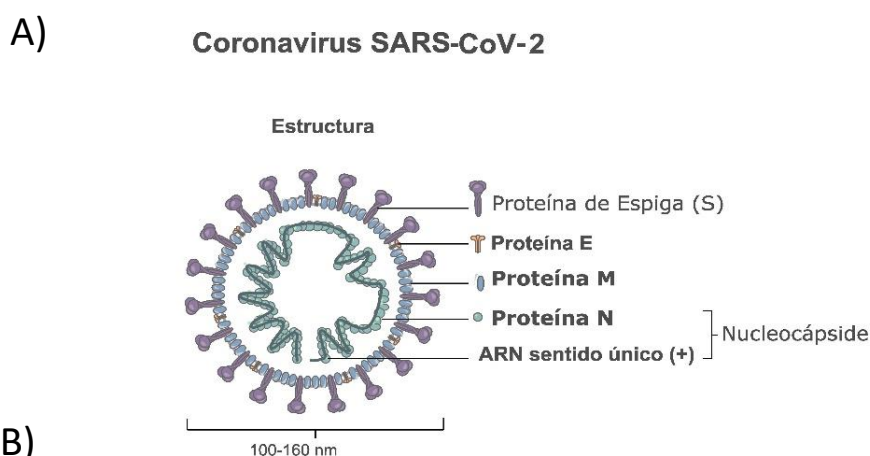


Figura 1.1. Esquema representando la estructura molecular y genética básica del SARS-CoV-2. (A). Diagrama de la estructura viral del SARS-CoV-2 y sus principales proteínas. **(B).** Esquema de los principales componentes del genoma del SARS-CoV-2, distinguiendo entre proteínas estructurales y no estructurales. PLpro (Papain-Like Protease); 3CLpro – (3C-Like Protease); RdRp (RNA-dependent RNA polymerase); Hel – Helicase). Figura adaptada de Vega Asensio, CC BY-SA 4.0 <<https://creativecommons.org/licenses/by-sa/4.0/>>, via Wikimedia Commons.

Otro mecanismo de entrada se produce por endocitosis, mediado esta vez por ACE2 y L-catepsina [11]. El SARS-CoV-2, tras este proceso el virus se replica y libera para infectar nuevas células hasta llegar a las correspondientes del epitelio alveolar. Durante la infección en esta fase, los casos severos producen una fuerte y desregulada respuesta inmunitaria, caracterizada por el síndrome de la tormenta de citoquinas [12–14]. A esta alteración se le atribuye el posterior desarrollo del síndrome de distrés respiratorio agudo e insuficiencia respiratoria, principales complicaciones críticas de la COVID-19 [12,15,16].

Las manifestaciones clínicas de la COVID-19 cubren un amplio espectro de sintomatologías. En pacientes no vacunados, tras el contagio la COVID-19 desarrolla un periodo de incubación que dura aproximadamente 5-7 días [16–18]. En este período se da el cuadro asintomático de la enfermedad, en la que algunos pacientes, en especial los más jóvenes, desarrollan la enfermedad en este estado hasta completar la descarga viral. Un 81% de los pacientes desarrolla tras el periodo de incubación sintomatología leve, con una duración aproximada de 8 días desde el momento de la infección [15,16]. El cuadro leve se caracteriza por el desarrollo de fiebre, fatiga, tos seca o neumonía leve en casos moderados. Un 14% de los pacientes desarrollan síntomas severos, como es el caso de la disnea, y en los casos críticos se requiere el ingreso en la

Unidad de Cuidados Intensivos (UCI) [15,16,19]. De no haber una reducción de la carga viral y con ello una mejoría de síntomas, los pacientes pasan a un estado crítico a los 14-20 días tras la aparición de síntomas [20] aunque hay estudios que indican un desarrollo crítico más temprano [15]. Esta nueva fase se caracteriza por el desarrollo del síndrome de distrés respiratorio agudo, insuficiencia respiratoria y fallo multiorgánico. El 5% de los pacientes desarrollan este cuadro crítico de diversa intensidad y duración, causando en algunos de ellos la muerte por COVID-19 [15,16,19].

Es conocida que la edad, así como la existencia de otras complicaciones como el asma influyen fuertemente en la progresión de la COVID-19 de un estado leve a uno grave-crítico. En términos generales, las personas mayores de 60 años son susceptibles a desarrollar sintomatología grave, mientras que en los mayores de 70 las probabilidades de desarrollar un cuadro crítico se incrementan [21,22]. Como se verá más adelante, una de las principales causas se debe al estado del sistema inmunitario adaptativo.

1.1.2 Epidemiología de la COVID-19 en el mundo

A fecha en la que este proyecto de tesis está siendo acabada (marzo de 2025) el número acumulado de casos de COVID-19 a nivel mundial es

de 777 millones, y el de muertes por esta patología asciende a 7 millones. La propia OMS tiene en cuenta la infraestimación de estas cifras, en especial las relacionadas con el diagnóstico y la positividad, con grandes diferencias entre países. De acuerdo con estimaciones obtenidas a partir de mediciones en aguas residuales, el número de casos confirmados clínicamente sufre una infraestimación de 2 a 19 veces inferior. Mientras que en regiones de la OMS como Europa, Américas o Sureste asiático se está produciendo una disminución de nuevos casos, en regiones como África o el Mediterráneo Este (Norte de África y Oriente Medio), se está produciendo un nuevo incremento. Con todo ello, el porcentaje global de positividad se sitúa actualmente en un 10.86%, con 92 países reportando casos activos.

1.1.3 Epidemiología de la COVID-19 en España

En el caso de España, la Organización Mundial de la Salud reportó que hasta junio de 2024 un número acumulado de casos de casi 14 millones de personas con 154.490 muertes desde el inicio del registro de casos. La positividad se encuentra actualmente en el 17.07 %, con un incremento del 36.55 respecto a la semana anterior. No existen datos epidemiológicos disponibles actualizados a marzo de 2025

Desde el inicio de la contabilización de casos a inicios de 2020, en nuestro país podemos distinguir 6 olas de infección, con diferentes impactos en función de la mortalidad y saturación del sistema sanitario. La primera ola, que comenzó en marzo del 2020, se caracterizó por ser la que más mortalidad presentó y la causante de la activación del Estado de Alarma, con una saturación grave del sistema sanitario. Con el desarrollo de las primeras vacunas y su administración preferente a la población vulnerable (mayores de 60 o personas de riesgo por comorbilidades) y a personal sanitario, las siguientes olas se caracterizaron por una menor mortalidad y saturación, aunque todavía presente. Una excepción en cuanto a la saturación de ingresos hospitalarios fue la cuarta ola, ubicada entre enero y marzo de 2022, que se caracterizó por ser la de mayor incidencia acumulada. Cabe señalar que la menor incidencia en olas posteriores puede estar fuertemente sesgadas por la ausencia de una prueba específica de detección del SARS-CoV-2 o la dificultad de acceso a la misma, además de la creación por ese momento del sistema de seguimiento de la pandemia.

Si bien hemos pasado un periodo de baja incidencia en todos los indicadores (casos confirmados, ingresos hospitalarios y/o mortalidad), en el momento en el que esta tesis se estaba comenzando a escribir (verano de 2024) nuestro territorio experimentó un nuevo incremento de casos,

debido a la irrupción de FlIRT. Se trata de una subvariante de la ya contagiosa y menos severa Variante de Preocupación (VOC) Ómicron, la cual es dominante en nuestro país desde inicios de 2022 y es responsable de las últimas olas de contagio. El impacto de Ómicron a partir de ese periodo ha traído consigo la detección y rastreo de diversas subvariantes como BA.1*, BA.5* o BQ.1*.

1.2 Linfocitos T: mediadores en la respuesta frente a la COVID-19

El sistema inmunitario adaptativo consiste en una colección de células, factores y mecanismos que responden a elementos reconocidos como extraños. Estos elementos pueden ser intrínsecos, como tumores o tejidos en enfermedades autoinmunes, o externos, como alérgenos y virus como el SARS-CoV-2. El proceso central de este sistema es la identificación por medio de receptores celulares específicos de pequeños fragmentos de dichos elementos extraños, conocidos como antígenos. Las principales células responsables de tal reconocimiento se las conoce como linfocitos, y existen dos grandes tipos, los linfocitos B y los linfocitos T. Los linfocitos B actúan mediante la secreción de moléculas conocidas como anticuerpos, mientras que los linfocitos T poseen funciones de modulación, eliminación de células afectadas y adquisición de una capacidad de respuesta mayor y más rápida en caso de una nueva presencia

de ese mismo elemento extraño en episodios futuros. Son estas propiedades las que hacen a los linfocitos T uno de los componentes más importantes en la respuesta inmunitaria frente al SARS-CoV-2. En esta sección, se presenta el potencial de los linfocitos T en el desarrollo de la respuesta inmunitaria a partir del fenómeno conocido como “Sensibilización celular sin seroconversión”, y se resumen las principales alteraciones de esta población de células inmunitarias asociadas a la gravedad de la COVID-19 y la influencia de otros factores como la edad y el sexo de la persona en dicha respuesta.

1.2.1 Inmunidad adaptativa frente a la COVID-19:

“Sensibilización celular sin seroconversión”.

El control efectivo de la infección por SARS-CoV-2 requiere una respuesta coordinada entre la inmunidad innata y adaptativa [23,24]. La inmunidad innata consiste en una respuesta tanto celular como humoral que es rápida y específica, actuando contra cualquier elemento reconocido como extraño. Una de las principales características es que su respuesta no es dependiente de que haya existido un evento de exposición previa a dicho agente extraño. Dentro de las células que componen el sistema inmunitario innato, destacamos en la lucha contra la COVID-19 las células de la línea monocito-macrófago, y en especial subtipos como los monocitos CD14+

o CD33- [25,26]. De hecho, la caracterización de subpoblaciones específicas de la línea monocito-macrófago supone uno de los principales tipos de análisis englobado dentro de un amplio conjunto conocido como análisis de inmunofenotipado. Para ello, el desarrollo de técnicas de citometría como es el caso de la citometría de masas (CyTOF) permite caracterizar de forma muy precisa diferentes subpoblaciones de células inmunitarias y sus diferentes estadios de activación, en especial para la línea monocito-macrófago [27]. Esto es debido al amplio conjunto de proteínas y otras biomoléculas que pueden ser marcadas para su posterior detección. Fruto de ello, son numerosos los estudios donde se ha aplicado CyTOF para inmunofenotipado de COVID-19 [28–31].

En el caso de la inmunidad adaptativa, dicha respuesta ha de ser específica, diversa y con capacidad de memoria (es decir, la creación de subtipos celulares capaces de responder de nuevo a un determinado antígeno), a la vez de evitar la reactividad frente a lo propio. Los anticuerpos desempeñan un papel crítico en la neutralización del virus, mientras que la respuesta celular es clave para la eliminación del virus una vez establecida la infección. Los linfocitos B son células especializadas en la secreción de moléculas conocidos como anticuerpos, los cuales se unen y neutralizan al elemento extraño a partir del reconocimiento a uno o varios antígenos específicos [32,33]. Este proceso constituye el núcleo de

la inmunidad humoral, y ha sido el principal foco en la investigación y desarrollo de vacunas frente al SARS-CoV-2 [32,34] . Por otro lado, los linfocitos T son los principales mediadores de la inmunidad celular y también participan como mediadores de la respuesta humoral [35] (Figura 1.2). Esta población está a su vez dividida en diversos grupos, en donde destacan las grandes poblaciones de linfocitos T CD4 (*helper*) y T CD8 (citotóxicos) [35–37] . Ambos tipos son inicialmente activados mediante el reconocimiento del receptor de linfocitos T (TCR) a antígenos presentados por células específicas en esta función. Los linfocitos T CD4 activados poseen varias funciones efectoras encargadas de modular la respuesta celular, mientras los linfocitos T CD8 matan aquellas células que expresan en su superficie antígenos de elementos extraños [24,35–37]. En el caso de la COVID-19 serán eliminadas aquellas células infectadas por el virus.

Aunque tradicionalmente se ha considerado que la producción de anticuerpos es la mejor medida para evaluar la protección contra la infección, la evidencia acumulada sugiere que los linfocitos T también desempeñan un papel esencial en la prevención de infecciones [35–37]. Se ha observado que algunas personas desarrollan respuestas celulares específicas sin generar anticuerpos detectables, lo que se conoce como "sensibilización celular sin seroconversión" [37] . La infección por SARS-

CoV-2 generalmente conduce a respuestas frente antígenos específicos, aclaramiento viral y creación de memoria inmunológica. Se ha observado que los linfocitos T CD4 y CD8, así como los linfocitos B, persisten en el pulmón, médula ósea, bazo y múltiples ganglios linfáticos hasta al menos seis meses después de la infección [38]. Otros estudios han demostrado correlaciones significativas entre linfocitos T de memoria circulantes y residentes en los tejidos [39]. Además, personas no expuestas previamente al SARS-CoV-2 presentan reactividad de linfocitos T de memoria hacia los epítomos del virus [40]. Este descubrimiento sugiere una existencia de linfocitos T inducidos por patógenos anteriores como SARS-CoV o MERS, y que podrían desempeñar un papel relevante en la respuesta inmunitaria frente a la COVID-19 [39].

La inmunidad adaptativa mediada por linfocitos T muestra diferentes patrones de estabilidad. Mientras que los anticuerpos se degradan rápidamente, las respuestas celulares tienden a permanecer estables [39]. Incluso con VOCs del SARS-CoV-2 con alta tasa de mutación como Ómicron, se ha detectado que del 70% al 80% de la respuesta de células T CD4 y CD8 frente a la proteína S se mantiene [41]. Además, la infección por SARS-CoV-2 induce respuestas vigorosas de células T CD4 y CD8 que están dirigidas no solo contra el antígeno de la proteína S, ampliamente utilizado en la mayoría de las estrategias de

vacunación, sino también contra otros antígenos, como los provenientes de la nucleocápside, proteína de la membrana y proteínas no estructurales [42]. Además, la efectividad de la respuesta queda influenciada también por la acción de las vacunas actuales contra la COVID-19. En esta línea, los análisis bioinformáticos han revelado que la inmensa mayoría de los epítomos de SARS-CoV-2 reconocidos por los linfocitos T se conservan en el resto de las VOCs virus, incluso en aquellas variantes con un mayor número de mutaciones acumuladas como es el caso de Ómicron [43].

1.2.2 Linfocitos T y severidad de la COVID-19

Los linfocitos T desempeñan un papel fundamental tanto en la prevención de infecciones por SARS-CoV-2 como en la limitación de su gravedad. Su función cobra especial importancia dada la alta divergencia en los diferentes cuadros pronósticos observados en individuos no vacunados, en donde existen diferencias clave asociadas a un desarrollo de la enfermedad leve o grave/crítico. En el siguiente apartado se describen las principales alteraciones de la respuesta de los linfocitos T asociado al pronóstico de la COVID-19.

Una de las principales características en los casos graves de COVID-19 en relación con los linfocitos T es el desarrollo de linfopenia profunda, es decir, una reducción del conteo total de linfocitos T, en

especial a las poblaciones CD4 y CD8 [37,44–47]. Se ha observado linfopenia profunda en poblaciones CD8 en tejido pulmonar en casos graves y fatales. Sin embargo, en casos leves la linfopenia es mucho menos acusada [23,24,37,46]. Incluso en pacientes asintomáticos se ha detectado una respuesta linfocitaria T frente a la infección. Una de las principales causas de la linfopenia podría ser un incremento en la apoptosis de los linfocitos T debido a una desregulación en la acción del interferón 1 (INF-1) [48]. Se ha demostrado que la modulación por INF-1 juega un papel crítico en la respuesta inmunitaria y en la progresión de la COVID-19 [42].

Otros de los factores observados es el retraso en la respuesta linfocitaria. Las respuestas tempranas de las células T CD4 específicas para el SARS-CoV-2 se encontraron que tenían la asociación más fuerte con la reducción de la gravedad de la enfermedad [42]. Las células T CD4 específicas del SARS-CoV-2 también pueden detectarse en los días 2-4 después de la aparición de los síntomas y son más prominentes que las respuestas de células T CD8+ [46]. Por otro lado, el desarrollo precoz de una respuesta citotóxica de células T CD8, que suele observarse a los 7 días de los síntomas y alcanza su punto máximo a los 14 días, se correlaciona con un aclaramiento vírico eficaz y una enfermedad leve, y coincide con una cinética similar de las respuestas humorales [46,49]. Esta respuesta rápida viene de nuevo coordinada por INF-1 [50].

Por el contrario, una ausencia de respuesta temprana y efectiva de los linfocitos T está asociada a una progresión grave de la enfermedad, provocando una descoordinación entre la respuesta celular y humoral. Los pacientes que no desarrollan una respuesta celular específica temprana e inician una respuesta inmunitaria humoral con la consiguiente producción de altos niveles de anticuerpos desarrollan síntomas graves [39]. Específicamente, en pacientes hospitalizados se ha reportado una respuesta tardía y persistente de los linfocitos CD8+ [39].

La persistencia en la respuesta también es otro factor determinante. En esta línea, respuesta linfocitaria T queda preservada enormemente, detectándose presencia de células T de memoria específicas para el SARS-CoV-2 en la sangre hasta 22 meses después de la infección [46]. Además, se ha detectado que la ausencia de respuesta linfocitaria a los 22 días post aparición de síntomas está asociada a un desarrollo crítico o mortal de la enfermedad [51].

No solo influye la cantidad (presencia/ausencia de linfopenia) o la velocidad y persistencia de la respuesta, también influye la efectividad de la misma y su coordinación con otros componentes del sistema inmunitario. Esta desregulación del sistema inmunitario es un elemento común en la patogenicidad de otros coronavirus, como el SARS-CoV1 y el MERS [24,37,42,46]. En la COVID-19 se ha observado que la

heterogeneidad en la funcionalidad de las células T CD8 está relacionada con la evolución de la enfermedad [42]. En esta línea, se ha detectado que una mayor expresión de moléculas efectoras de la respuesta CD8⁺ en fases agudas de la infección está asociada a un cuadro leve [46,52,53]. Otra observación sobre la efectividad de la respuesta linfocitaria T tiene que ver con la reactividad a determinados antígenos del SARS-CoV-2. Las respuestas de las células T CD8 contra el epítipo NP105–113 restringido por B*07:02 demuestran una fuerte actividad antiviral y se correlacionan con la protección contra enfermedades graves [54].

En relación con la descoordinación y la falta de efectividad, existen poblaciones específicas de linfocitos T cuyos cambios son provocados y a su vez consecuentes de un desarrollo grave de la enfermedad. Las células T CD4 específicas del SARS-CoV-2 que responden a la infección son predominantemente de los tipos T cooperadores 1 (Th1) y T cooperadores foliculares (Tfh) [23,42,46]. Un desequilibrio de la ratio de linfocitos Th1/Th2 así como de la ratio Th1/Th17 y consecuente alteración en la secreción de citoquinas estaría asociado a una mayor gravedad de la enfermedad [55,56] (**Figura 1.2**). Por otro lado, se ha encontrado una asociación entre un desarrollo grave de la enfermedad y cambios en la frecuencia de linfocitos Tfh circulantes y específicos de SARS-CoV-2 [57].

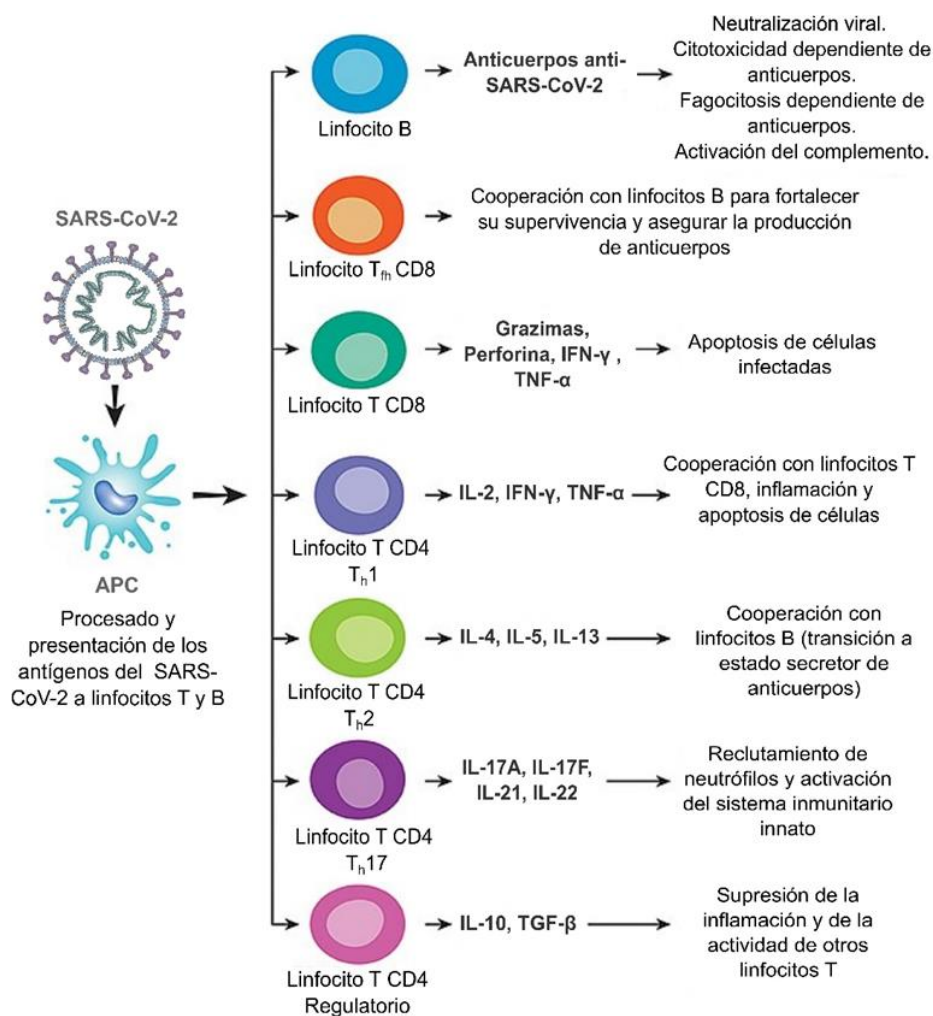


Figura 1.2. Esquema con los principales grupos de linfocitos, citocinas secretadas y sus funciones durante la infección del SARS-CoV-2. Se hace énfasis en las distintas subpoblaciones de linfocitos T. Adaptación de la Figura 3 de Flanagan y colaboradores [47].

1.2.3 Linfocitos T frente al COVID-19: influencia de la edad.

La edad influye significativamente en el desarrollo de la respuesta adaptativa de los linfocitos T al SARS-CoV-2. Se ha observado un incremento significativo en el desarrollo de COVID-19 severo y de fallecimientos en individuos mayores de 65 años [46,58]. Esto es debido a que, entre otros factores, las personas mayores tienden a tener una respuesta más débil debido a una reducción en el número de células T naïve, así como a cambios en la inmunidad innata con la edad [22,23,52]. Además, los cambios en el sistema inmunológico relacionados con el envejecimiento pueden reducir la eficacia de las vacunas en los ancianos [59]. Las personas mayores también pueden tener una mayor predisposición a desarrollar formas graves de COVID-19 debido a la inmunosenescencia, que se asocia con una disminución de la inmunidad mediada por células T y un aumento de la inflamación sistémica [46]. Finalmente, con el aumento de la edad existe un desajuste del mecanismo de acción del INF-1 con las consecuentes alteraciones en la respuesta efectiva de los linfocitos T [59,60]

1.3 El receptor de linfocitos T: función, estructura y mecanismo de acción

La importancia de la respuesta inmunitaria mediada por linfocitos T no sería la misma en ausencia del receptor de linfocitos T (TCR). Este receptor de membrana es, junto con otros componentes que se mencionan a continuación, el principal mediador en la respuesta adaptativa por linfocitos T. Por lo tanto, es necesario introducir sus principales características estructurales, mecanismo de acción y de síntesis, que nos servirán posteriormente para definir uno de los conceptos centrales de esta tesis doctoral: el repertorio de linfocitos T o repertorio de TCR.

1.3.1 Inmunorreceptores y el complejo TCR-pMHC

Los TCR, junto con los receptores de células B (BCR), forman una amplia familia de proteínas de membrana llamadas inmunorreceptores [61,62]. Sus principales características son la alta complejidad de sus dominios extracelulares, adaptados a la interacción con ligandos, y una marcada ausencia de actividad catalítica en los dominios intracelulares [61]. Una vez unido el ligando (antígenos en el TCR), se desencadena una serie de cambios conformacionales que, a su vez, activan una cadena de fosforilación de residuos en el propio receptor y en proteínas adicionales, dando como resultado un complejo mecanismo de segundos mensajeros.

Un antígeno es una molécula o fragmento de molécula que el sistema inmunológico reconoce como extraño o potencialmente dañino, provocando una respuesta inmunitaria. Pueden ser proteínas, polisacáridos, lípidos o ácidos nucleicos presentes en patógenos como bacterias, virus, hongos y parásitos, así como en células dañadas o anormales del propio organismo. Dichos antígenos son, en el caso del TCR, expuestos por células alteradas por algún tipo de patología. En el COVID-19, los antígenos son fragmentos de proteínas del propio SARS-CoV-2 [63]. El antígeno es expuesto en la superficie celular de la célula infectada mediante su unión al Complejo Principal de Histocompatibilidad (MHC) [35,62,64], formando lo que se conoce como el complejo péptido-MHC (pMHC) [62,65]. En humanos, el MHC también se lo conoce como Antígeno Leucocitario Humano (HLA) [66].

1.3.2 Estructura del TCR

La estructura del TCR es un complejo proteico formado por las subunidades alfa y beta ($\text{TCR}\alpha$ y $\text{TCR}\beta$) y el complejo CD3[61,62,67,68] (Figura 1.3 A-B). Las subunidades $\text{TCR}\alpha$ y $\text{TCR}\beta$ forman parte de la región extracelular del TCR, conociéndose como $\text{TCR}\alpha\beta$ [61,62,65,68] . Aunque la mayoría de los linfocitos T poseen el receptor $\text{TCR}\alpha\beta$, entre un 0.5% y un 5% de los linfocitos T circulantes en sangre poseen el receptor

gamma-delta ($\text{TCR}\gamma\delta$), el cual se caracteriza por reconocer antígenos sin necesidad del complejo pMHC [67,69–71] .

Podemos distinguir en $\text{TCR}\alpha$ y $\text{TCR}\beta$ regiones variables ($V\alpha$ y $V\beta$) y constantes ($C\alpha$ y $C\beta$) [62,65,68,72]. Las regiones variables son las encargadas de establecer contacto con el pMHC [62]. Para que se realice esta interacción, cada región variable posee tres secuencias de aminoácidos hipervariables conocidas como Regiones Determinantes de la Complementariedad (CDR1, CDR2 y CDR3, respectivamente) [62,65,68]. Mientras que los CDR1 y CDR2 establecen contacto con regiones variables del MHC, el CDR3 interactúa directamente con el antígeno [62]. Por otro lado, las regiones constantes proporcionan una estructura estable al TCR, asegurando su integridad mediante su anclaje a la membrana plasmática y su interacción con proteínas del complejo CD3 [61,62,73]. Ambas subunidades, $\text{TCR}\alpha$ y $\text{TCR}\beta$, se encuentran unidas mediante un enlace disulfuro entre las regiones $C\alpha$ y $C\beta$ [61,62]. Además, dicha unión de subunidades se refuerza mediante interacciones no covalentes entre otros residuos de las regiones tanto variables como constantes [62] .

El complejo CD3 forma parte de la región transmembrana y citosólica, y su principal función es la de transductor de señal mediante la regulación de procesos de fosforilación y desfosforilación. Está compuesto

por las subunidades épsilon ($CD3\epsilon$), gamma ($CD3\gamma$), delta ($CD3\delta$) y ζ ($CD3\zeta$), que forman heterodímeros ($CD3\epsilon\gamma$ y $CD3\epsilon\delta$) y un homodímero ($CD3\zeta\zeta$) [61,62,68,72]. Estas subunidades poseen dominios extracelulares (ECDs) del tipo inmunoglobulina-*like*, los cuales establecen los puntos de unión a las regiones constantes de $TCR\alpha\beta$ [61].

Concretamente, el dominio constante de $TCR\alpha$ ($C\alpha$) interactúa con el ECD de $CD3\delta$, mientras que $C\beta$ se conecta con los ECDs de $CD3\delta$ y $CD3\gamma\epsilon$ [61]. Además, el ECD de $CD3\epsilon$ del heterodímero $\delta\epsilon$ interactúa con el ECD de $CD3\gamma$ [61] (Figura 1.3 A-B). Los avances técnicos en la modelización de la estructura tridimensional de proteínas han revelado que el $TCR\alpha\beta$ se asocia con los tres módulos de señalización dimerizados $CD3\epsilon\gamma$, $CD3\epsilon\delta$ y $CD3\zeta\zeta$ en una estequiometría 1:1:1:1, formando un ensamblaje octamérico [61,74].

Además de los ECDs, las subunidades del complejo CD3 contienen dominios transmembrana que estabilizan el complejo y dominios citoplasmáticos con Motivos de Activación Basados en Tirosina (ITAMs) (Figura 1.3 A) [61,62,68]. El gran número de ITAMs contribuye a la alta sensibilidad del complejo TCR [61]. Las colas citosólicas de las subunidades CD3 que contienen ITAMs están cargadas positivamente y podrían estar inicialmente protegidas de la fosforilación al interactuar con lípidos cargados negativamente en la cara interna de la membrana

plasmática [61,62]. La estructura del TCR guarda por tanto un complejo sistema de transducción de señal que comienza con la unión del TCR $\alpha\beta$ al pMHC y se traslada a la región citosólica mediante la exposición y fosforilización de ITAMs.

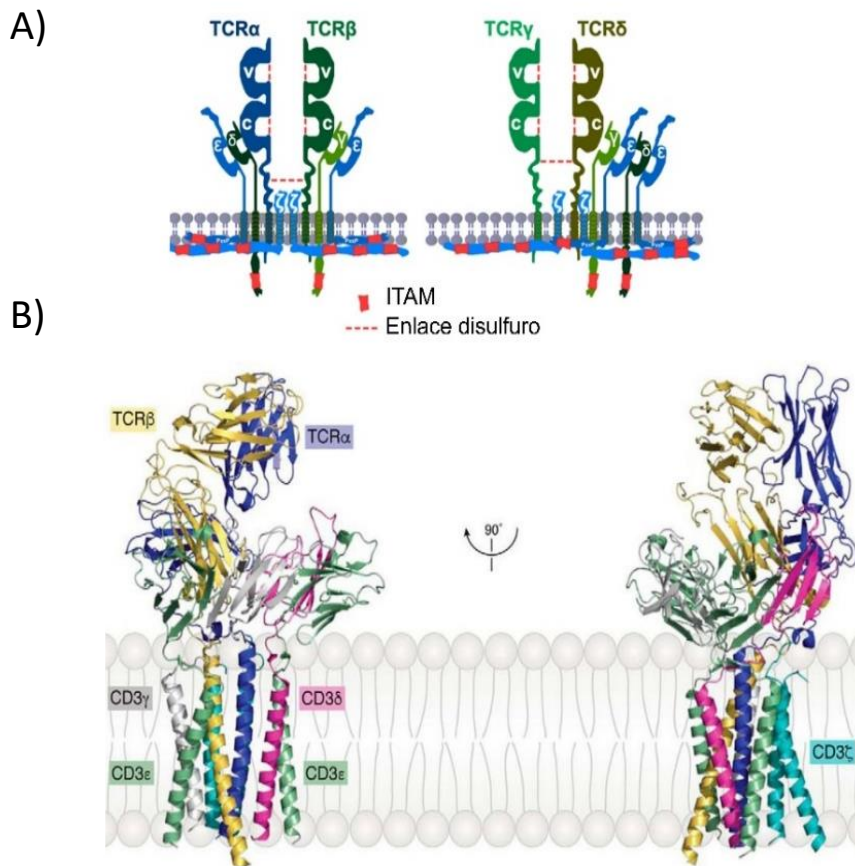


Figura 1.3. Estructura del TCR. (A) Esquema de los principales componentes del TCR $\alpha\beta$ y TCR $\gamma\delta$ con la distribución de sus correspondientes CD3. Extraído de la Figura 1A de Shah y colaboradores [61]. (B). Estructura tridimensional del complejo TCR $\alpha\beta$ -CD3 (ID de Protein Data Bank 6JXR). Extraído de la Figura 2 A de Mariuzza y colaboradores [72]

1.3.4 La variabilidad de la región CDR3: la recombinación V(D)J.

El TCR, al igual que BCR y los anticuerpos, está codificado por una serie de segmentos de ADN que forman el vasto locus del TCR, abarcando varios centenares de kilobases [75]. En estos *loci* se distinguen segmentos que codifican para la región constante del TCR y aquellos que forman la región variable, la cual se divide en tres tipos de segmentos: V (Variable), J (Joining) y D (Diversity) [76–78]. En el TCR α hay aproximadamente 42 segmentos V y 61 segmentos J, mientras que en el TCR β hay 47 segmentos V, 12 segmentos J y 2 segmentos D [79], aunque otras revisiones apuntan a otros valores sumando o restando una o pocas unidades a al conteo final de segmentos [65]. La región CDR3, responsable de la unión del TCR con el antígeno, se codifica en la región comprendida entre los segmentos V y J para el TCR α , y entre los segmentos V, D y J para el TCR β [76–78].

La recombinación V(D)J es posible mediante la acción coordinada de elementos en cis y trans. Los elementos en cis están formados principalmente por las secuencias de señal de recombinación (RSS), las cuales flanquean cada alelo V, D y J [77,78,80]. Consisten en un heptámero y un nonámero de secuencias conservadas, separadas por una región espaciadora de 12 a 23 nucleótidos (12RSS y 23RSS, respectivamente)

[77,80]. En los elementos en trans, las enzimas recombinasa RAG1 y RAG2 son esenciales, junto con las proteínas B1 del Grupo de Alta Movilidad (HMGB1) y HMGB2. Este proceso se divide en dos fases: escisión y unión [77,78,80].

En la fase de escisión, el complejo RAG-HMGB1 se une a un 12RSS y otro complejo similar a un 23RSS en la cadena complementaria, formando un complejo pareado o sinapsis [76,77]. Esta preferencia a la hora de combinar 12RSS con 23RSS se lo conoce como regla 12/23 [76,78]. El complejo RAG-HMGB1 corta el ADN entre el heptámero del RSS y el alelo, generando horquillas de cierre y separando los alelos V(D)J [76]. El resultado son segmentos de ADN bicatenario-unidos a sí mismos por una horquilla y un complejo de fin de señal compuesto por RAG-HMGB1 unido a los fragmentos de ADN de los RSS [77,80].

La fase de unión, las horquillas formadas por los alelos escindidos son procesadas para ser cortados y unidos a la secuencia del otro alelo, quedando finalmente unidos en una misma secuencia un alelo V con otro J con un paso extra incluyendo a D en el caso del TCR β [76]. El proceso de unión de ambos ADN bicatenarios se realiza mediante la Unión No Homóloga de Extremos (NHEJ en sus siglas en inglés), realizado por la acción coordinada de enzimas de reparación y unión al ADN [76–78]. Durante el NHEJ, la enzima Desoxinucleotidil Transferasa Terminal (TdT)

añade nuevos nucleótidos en la unión de alelos, sumando aún más variabilidad a la añadida por NHEJ [76–78].

1.4 El repertorio de linfocitos T

El número de alelos V(D)J presentes y el sofisticado mecanismo de recombinación con adición de nucleótidos al azar descrito en el apartado anterior genera una ingente cantidad de CDR3s y por ende de TCRs distintos. Al conjunto de TCRs que encontramos en un organismo lo catalogamos como repertorio de linfocitos T o de TCR [81] y su diversificación viene a responder lo definido en la teoría de la selección clonal, en la que una respuesta vigorosa ante patógenos junto con la ausencia de reactividad a lo propio viene determinada por un repertorio de TCR diverso, clonalmente distribuido y con especificidad a antígeno [82]. Por tanto, este proceso de generación de diferentes TCRs es la forma en el que el sistema inmunitario trata de garantizar que el organismo sea capaz de reaccionar en un evento futuro frente a la presencia de un patógeno de forma precisa y efectiva. En este apartado se describirá cuáles son las principales características del repertorio TCR en humanos, desde el final de la recombinación V(D)J hasta llegar a la población de linfocitos funcionales y específicos de antígenos, qué son los clonotipos y su potencial como biomarcadores de una enfermedad concreta.

1.4.1 ¿Qué son los clonotipos?

La unidad básica del repertorio de TCR es el clonotipo. Atendiendo a su definición estricta, un clonotipo corresponde a cada secuencia de nucleótidos V(D)J única tras el proceso de síntesis del CDR3 al que pertenece [83–85]. Sin embargo, también hay estudios en donde se refieren a clonotipo atendiendo a la secuencia final de aminoácidos y no de nucleótidos [86,87]. Teniendo en cuenta que el TCR está formado por una cadena α y otra β con sus respectivos CDR3s (también aplicable a los TCR $\gamma\delta$), también se define como clonotipo a aquel linfocito T idéntico a los demás de la población, pero con un TCR distinto al de los demás a partir de la secuencia CDR3 de una o de las dos cadenas. Un término nuevo fruto del avance en el estudio de los repertorios de TCR es el concepto de meta-clonotipo: conjuntos de clonotipos con estereotipia, es decir, distintos pero que poseen combinaciones V(D)J casi idénticas [85]. Estos clonotipos con estereotipia comparten características comunes como la afinidad por un mismo conjunto de HLAs [84] .

1.4.2 La diversidad estructural del TCR en números

El resultado del proceso de síntesis de las regiones CDR3, se ha estimado que gracias a este proceso se puede obtener $1-5.2 \times 10^{15}$ posibles TCRs, a partir de numerosas recombinaciones V(D)J y adición de

nucleótidos y la unión de las cadenas α y β resultantes, en lo que se conoce como diversidad teórica [64,66]. Otro estudio ha llegado a calcular una estimación de hasta 10^{61} TCRs distintos tras esta fase de síntesis [88,89], un número tremendamente más grande que la estimación clásica anteriormente mencionada, e inconmensurablemente mayor que la cantidad total de linfocitos T que un humano podría generar a lo largo de su vida [88].

Sin embargo, y atendiendo únicamente a la estimación clásica, los 5.2×10^{15} TCRs posibles difieren en gran cantidad respecto a lo observado en la realidad. En la diferenciación y selección de timocito a linfocito T en el timo, se ha estimado con una estimación inicial de 10×10^{13} linfocitos T inmaduros o naïve [64], aunque estimaciones más conservadoras dan un repertorio de 1×10^9 o 1×10^8 [66,83,88,89]. Una estimación más reciente del 2023 propone un valor del tamaño del repertorio (contando TCRs $\alpha\beta$) de 3.8×10^8 [81]. No obstante, se han detectado experimentalmente en órganos periféricos 2×10^7 clonotipos de linfocitos T completamente funcionales [64]. Esta observación se ve apoyada por la aparición de clonotipos públicos, es decir, TCRs compartidos entre dos o varios individuos, en ocasiones entre grandes grupos [58,66,88–90]. Por lo tanto, varios factores y procesos contribuyen a la reducción de la diversidad

estructural, afectando la recombinación V(D)J, la selección tímica y la especificidad antigénica.

1.4.3 Factores influyentes en la diversidad estructural

La diversidad estructural del repertorio de linfocitos T funcionales se genera mediante procesos deterministas, que influyen tanto en la recombinación V(D)J como en la selección tímica y en la ya población activa de linfocitos T.

La recombinación V(D)J vendría regulada por los mecanismos agrupados dentro de la hipótesis de accesibilidad [90]. En este proceso, los diferentes segmentos V, D y J deben de estar previamente accesibles al complejo RAG [80,88,91]. Esta accesibilidad espaciotemporal viene dada por cambios epigenéticos como la organización de la cromatina, el estado de metilación del ADN o la acción de potenciadores de la transcripción [80,90,92]. En consecuencia, se observa un sesgo durante la recombinación V(D)J, lo que se conoce como recombinación convergente [66,90,93]. Este proceso favorece un uso preferente de ciertos segmentos V, D y J frente a otros.

El mecanismo de reducción de la diversidad estructural del repertorio desde las estimaciones iniciales hasta los valores observados más influyente es el producido durante la selección tímica [82]. En este

proceso, los timocitos recién formados pasan un exhaustivo control con el objetivo de garantizar una reactividad efectiva a lo ajeno y prevención de lo propio. La selección tímica consta de dos fases: selección positiva y selección negativa [82,94]. En la selección positiva, las células epiteliales del timo presentan complejos pMHC I y pMHCII, los cuales tratarán de unirse a los TCR recién formados de los timocitos [94]. Aquellos timocitos cuyo TCR concreto no llegue a unirse a ninguno de los MHC serán eliminados mediante la activación de rutas por apoptosis. En este paso también se produce la diferenciación a linfocito CD4 o CD8 en un proceso mediado por thPOK y RUNX3 [82,95,96]. En la selección negativa, los linfocitos inmaduros se someten a la presentación de antígenos por parte de Células Presentadoras de Antígeno (APCs). Un grupo especializado de APCs expresan el Regulador Autoinmune (AIRE), el cual a su vez permite la expresión y consecuente exposición como antígeno de numerosas proteínas propias de diferentes tejidos del organismo [82,97,98]. En esta selección, aquellos linfocitos T que se unen con una alta afinidad con el pMHC de las APCs son eliminados, pues presentan un riesgo para el desarrollo futuro de enfermedades autoinmunes [66].

Tras el proceso de selección positiva y negativa, los linfocitos T naïve resultantes sufren un proceso denominado competencia de proliferación, que consiste en un cambio en los procesos metabólicos que

permite a las células poder migrar y activarse a partir de una fuerte cascada de señalización proveniente del TCR [82,94,99]. Aquellos linfocitos T naïve que no desarrollan este proceso de forma adecuada, son inducidas a apoptosis. Finalmente, los linfocitos T maduros pasan a la periferia hasta los órganos linfoides secundarios, en donde comienzan con las funciones como células del sistema adaptativo [82]. En la [Figura 1.4](#) se resume de forma esquemática los diferentes niveles del repertorio de linfocitos T y los principales procesos de moldeado del repertorio.

1.4.4 La importancia de los repertorios de TCR durante la enfermedad.

Los linfocitos T maduros reconocen el antígeno presentado por las APCs e inician las correspondientes funciones como la secreción de citocinas, memoria, citotoxicidad y proliferación. El mecanismo de

proliferación resulta esencial para una correcta respuesta inmunitaria adaptativa celular al ser uno de sus primeros pasos. Durante la proliferación, aquellos clonotipos (linfocitos T con una misma secuencia de CDR3 y por extensión de TCR) que reconocen con buena afinidad al complejo pMHC en cuestión, se dividen para aumentar en primer lugar la población de clonotipos reactivos para ese antígeno y en segundo lugar para adquirir las correspondientes funciones mencionadas. A este proceso

se le denomina selección o expansión clonal y supone un cambio en el repertorio de linfocitos T [82,100,101] . Ante una infección, como es el caso del SARS-CoV-2 el repertorio disminuye su diversidad hacia un estado oligoclonal, en donde se produce una expansión clonal de aquellos clonotipos reactivos a antígenos del virus, cuya abundancia será mayor, en detrimento de los no reactivos [102–104] (Figura 1.4).

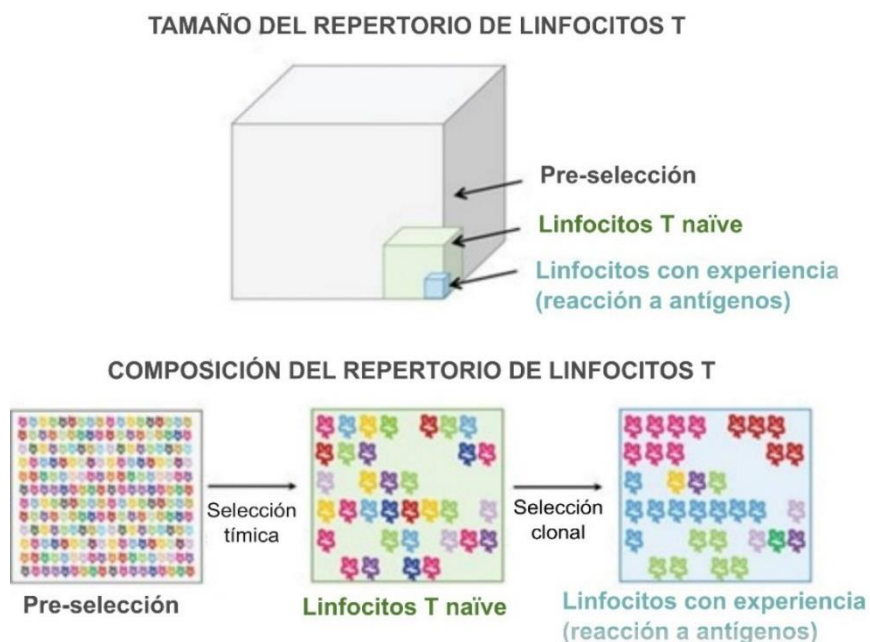


Figura 1.4. Esquema básico ilustrativo de los cambios en el tamaño del repertorio de linfocitos T. Se representa el cambio proporcional del tamaño del repertorio tras la recombinación V(D)J (preselección), tras el proceso de selección tímica (Linfocitos T naïve) y una vez madurados, liberados a la periferia y tras interactuar con un antígeno durante una infección u otra patología (Linfocitos con experiencia). Al proceso de selección de linfocitos maduros mediado por el antígeno se lo conoce por selección clonal y viene determinado por la afinidad del TCR por el complejo pMHC. Figura modificada a partir de la original publicada por Attaf, Huseby y Sewell [57].

A lo largo de la vida de un individuo, la exposición a diversos patógenos modula la arquitectura del repertorio. Estos cambios son extremadamente diversos y complejos, e incluso con la presencia de clonotipos públicos y el mismo trasfondo genético. Se ha observado que hasta en gemelos monocigóticos el repertorio es distinto pese a poseer el mismo conjunto de genotipos también en el HLA, un factor altamente influyente en la arquitectura del repertorio junto con la edad [66]. Tales observaciones apuntan a que el repertorio de TCR representa un complejo sistema sujeto a cambios que determinan las diferentes evoluciones clínicas de la enfermedad [58]. Esta característica va en concordancia con el papel clave y las propiedades observadas de la respuesta de los linfocitos T frente a la COVID-19 mencionadas en la sección 1.2.

Por tanto, el estudio de la respuesta inmunitaria de los linfocitos T a partir del repertorio de TCR supone una aproximación prometedora para poder entender las consecuencias patogénicas de la COVID-19, una enfermedad diversa y compleja en el sentido clínico y epidemiológico. Dicho estudio debe estar orientado hacia las bases de la Medicina Personalizada y de Precisión, lo que requiere el desarrollo de técnicas de análisis de laboratorio y bioinformáticas que permitan un abordaje profundo, detallado y que minimice el coste económico y las molestias al

paciente. Para ello, ha sido clave el desarrollo de una nueva subdisciplina de la bioinformática: la inmunoinformática.

1.5 El uso de la inmunoinformática para estudiar el repertorio de linfocitos T

Para lograr los objetivos marcados en la Medicina Personalizada y de Precisión, el auge de la inmunoinformática ha logrado desarrollar sofisticadas técnicas analíticas enfocadas a comprender algo tan complejo como es nuestro sistema inmunitario. La inmunoinformática a su vez ha favorecido la puesta a punto del *TCRseq*, un conjunto de análisis bioinformáticos cada vez más popular que permite un estudio detallado del repertorio de linfocitos T. En este último apartado, presentaremos sus principales características y potenciales aplicaciones, junto con una breve introducción a la importancia de la integración de datos inmunológicos en el contexto de la COVID-19.

1.5.1 ¿Qué es la inmunoinformática?

La inmunoinformática es la rama de la bioinformática especializada en unir la inmunología y las ciencias computacionales, mediante el uso de recursos, material y métodos computacionales sobre datos inmunológicos [105,106]. La aplicación de la inmunoinformática va

desde el análisis de datos obtenidos en laboratorio hasta el análisis *in-silico*, mediante la simulación o reciclaje de datos para generar nuevo conocimiento. Se trata de una disciplina emergente con gran potencial en aplicaciones como la vacunómica [107,108], interacciones huésped-patógeno [109,110], inmunoterapia [111–113] o identificación y seguimiento de la virulencia [114,115]. Estas aplicaciones se han conseguido mediante la implementación de técnicas ya consolidadas en bioinformática, como el RNA-seq, Genome-Wide Association studies (GWAS) [116] , reposición de fármacos [117] o implementaciones específicas de *machine learning* y *deep learning* [118–120]. Sin embargo, en los últimos años han aparecido nuevas técnicas analíticas específicas de la inmunoinformática, siendo una de las más populares el análisis del repertorio de linfocitos T.

1.5.2 El análisis del repertorio de linfocitos T (*TCRseq*)

El análisis de repertorios de linfocitos T o *TCRseq* es un conjunto de técnicas de laboratorio y principalmente bioinformáticas que se basan en la amplificación, secuenciación y posterior análisis computacional para poder extraer numerosa información en relación con el repertorio de linfocitos T. A continuación se describen las principales características de las distintas modalidades y métodos analíticos.

1.5.2.1 Modalidades

Un uso más popular ha llevado al desarrollo de numerosas variantes del *TCRseq* orientado al estudio de una o las dos cadenas del TCR (*βTCRseq* o *αβTCRseq*, respectivamente) y combinado desde un enfoque de células únicas (*scTCRseq*) o a partir de poblaciones celulares agrupadas (*bulk TCRseq*) [121,122]. Cada modalidad presenta diferentes ventajas y desventajas. El *scTCRseq* permite una distinción de repertorios en alta resolución al discernir subtipos celulares de linfocitos T y una rápida integración con los datos de *scRNAseq* [123–125], pero a su vez implica un mayor coste económico, computacional y el empleo de una menor cohorte de muestras, con el riesgo que supone la pérdida de información por problemas de viabilidad celular. Por su parte, el *bulk TCRseq* permite un estudio de cohortes más grandes a un menor coste económico y en especial con una mayor sencillez metodológica y analítica, a cambio de la pérdida de la resolución a nivel celular [126,127].

Dentro del *bulk TCRseq*, muchos estudios se han orientado hacia una cadena (*bulk βTCRseq*) [128], como es el caso de esta tesis doctoral. El tamaño relativamente pequeño de la cadena β y una mayor variabilidad en la secuencia CDR3 en comparación a la cadena α debido al uso del segmento D lo convierte en una variante más rápida, económica y manejable que el *bulk αβTCRseq*, a la vez que se conserva la mayor parte

de la diversidad estructural [113,127]. Debido a estas características, la mayoría de los estudios que conllevan análisis de TCRseq en COVID-19 aplican la modalidad β TCRseq). Tal y como se recopila en la revisión de Kockelbergh y colaboradores 9 de los 15 análisis de *bulk* TCRseq en COVID-19 recopilados fueron en la modalidad β TCRseq, y los 6 restantes mediante $\alpha\beta$ TCRseq [129]. Además, el coste más reducido de preparación de librerías y secuenciación únicamente de la cadena β supone un aumento considerable del tamaño de las cohortes estudiadas. De esta forma, el tamaño promedio redondeado de las cohortes de las modalidades *bulk* $\alpha\beta$ TCRseq es de 28 muestras frente a las 128 del conjunto de estudios de β TCRseq recopilados en [129]. Por otro lado, herramientas de clustering de TCRs popularmente usadas como GLIPH2 (descritas más adelante), dan importancia a la cadena β del CDR3 y los correspondientes alelos V J, dejando a la cadena α del CDR3 como información adicional. La base biológica para que exista esta preferencia en el análisis no solo se basa en la mayor diversidad de la cadena β como ya se ha comentado. Diversos estudios han demostrado que la frecuencia de cadenas β específicas de epítomos es concordante con la de sus respectivas cadenas α , lo cual respalda su uso aislado en análisis del repertorio de linfocitos T [130,131].

1.5.2.2 Principales técnicas de preparación de librerías de TCRseq

Podemos distinguir 3 tipos principales de captura, amplificación y posterior secuenciación para ensayo de *bulk TCRseq*: Amplificación Rápida de los Extremos 5' complementarios de ADN (5' RACE en sus siglas en inglés), la captura por hibridación, y la reacción en cadena de la polimerasa múltiple (Multiplex PCR).

La técnica 5' RACE se orienta al ARN e incluye la transcripción reversa a ADN complementario (cADN), seguida de elongación con desoxicitidina trifosfato (dCTPs) y dos ciclos de amplificación con adaptadores 5' y 3', los cuales hibridan en el extremo dCTPs elongado y la región constante del TCR, respectivamente [126,132,133]. Este método flanquea la región V(D)J, reduciendo sesgos en PCR y permitiendo la adición de UMIs, aunque presenta cierta complejidad técnica.

La captura por hibridación se aplica tanto a ADN como a ARN. Involucra la preparación de librerías de secuenciación (fragmentación, reparación de extremos, poliadenilación, unión de adaptadores e indexado). Tras ello se procede a la hibridación con oligonucleótidos biotinilados específicos para *loci* de TCR, y su consecuente captura mediante microesferas magnéticas [132].

La PCR múltiple puede emplear ADN o ARN como sustrato inicial, requiriendo transcripción reversa para ARN. Utiliza múltiples cebadores para unirse a regiones V conocidas y, en el caso del ARN, también a regiones C para el extremo 3', mientras que, para ADN, estos últimos cebadores hibridan con alelos J debido a la separación intrónica de V(D)J respecto a C [128,132,133]. Este método es sencillo y popular, ampliamente implementado en kits comerciales, aunque presenta limitaciones con cebadores dirigidos a alelos conocidos y dificultades para añadir UMIs. En cuanto a la secuenciación las plataformas Illumina MiSeq y HiSeq son las más comúnmente usadas.

1.5.2.3 Análisis primario de los repertorios de linfocitos T

El análisis primario consiste en el primer bloque de procesos bioinformáticos, los cuales están orientados al preprocesamiento de muestras. Concretamente, este pre-procesamiento trata de convertir los archivos. fastq resultantes de la secuenciación de las librerías en archivos de repertorios (uno por muestra en el caso del *bulkTCRseq*). Los archivos de repertorios son archivos tabulados en donde se guarda la información de los clonotipos detectados, su abundancia y frecuencia, secuencia de nucleótidos y aminoácidos de su CDR3 y de los alelos V(D)J que codifica, entre otras variables de interés.

Para llegar a este archivo de repertorios, las secuencias de ADN almacenadas en los archivos. fastq son inicialmente alineadas y anotadas para extraer qué alelos V(D)J codifican estas lecturas. Para ello, se emplean como referencia archivos con secuencias y coordenadas de alelos V(D)J, en este caso [121,134,135] . Tras ello, las lecturas mapeadas a diferentes alelos son reorganizadas para construir la secuencia completa del CDR3, pudiendo de esta manera asociar un alelo V con su correspondiente J y, dependiendo del protocolo, C. Los clonotipos reconstruidos son procesados para corregir errores durante la preparación de librerías y secuenciación. Al resultado de esta nueva fase se le denomina ensamblaje de clonotipos, y tiene como resultado la creación el archivo tabular de repertorios inicial [136].

El archivo de repertorios inicial es sometido ahora a un proceso de filtrado en varios aspectos. Primero, se eliminan clonotipos con mutaciones fuera de marco de lectura. Luego, se eliminan alelos no específicos de TCR, como los de inmunoglobulinas, especialmente en procesos que alinean conjuntamente TCR y BCR. El archivo final puede ser normalizado para ajustar las abundancias de clonotipos, asegurando que las diferencias en abundancias reflejan motivos biológicos y no variaciones en la profundidad de secuenciación. Una técnica común de

normalización es el *downsampling*, detallada en la sección de materiales y métodos.

1.5.2.4 Análisis secundario de los repertorios de linfocitos T

El análisis secundario se realiza sobre los archivos de TCR ya generados, filtrados y normalizados. El objetivo principal es el de, mediante técnicas estadísticas, obtener información acerca de las arquitectura, dinámicas y firmas del repertorio de TCR (Figura 1.5).

Orientado a la arquitectura y la dinámica general del repertorio, encontramos los análisis de diversidad, clonalidad y solapamiento de repertorios [126,137]. La estadística empleada se basa en el cálculo de índices ampliamente usados en el campo de la ecología, como es el estimador Chao 1 para la riqueza de especies, índice de Gini-Simpson para la diversidad α o el índice de Morisita para el solapamiento de repertorios (diversidad β) [113,126,137,138]. Las bases de estos índices y otros empleados en el desarrollo de esta tesis doctoral se describen con detalle en el capítulo de Materiales y Métodos. Otras métricas más propias del análisis del repertorio de linfocitos T es el uso de genes: concretamente el

cálculo de frecuencias alélicas de los diferentes alelos V, D o J y su inferencia estadística entre repertorios [121].

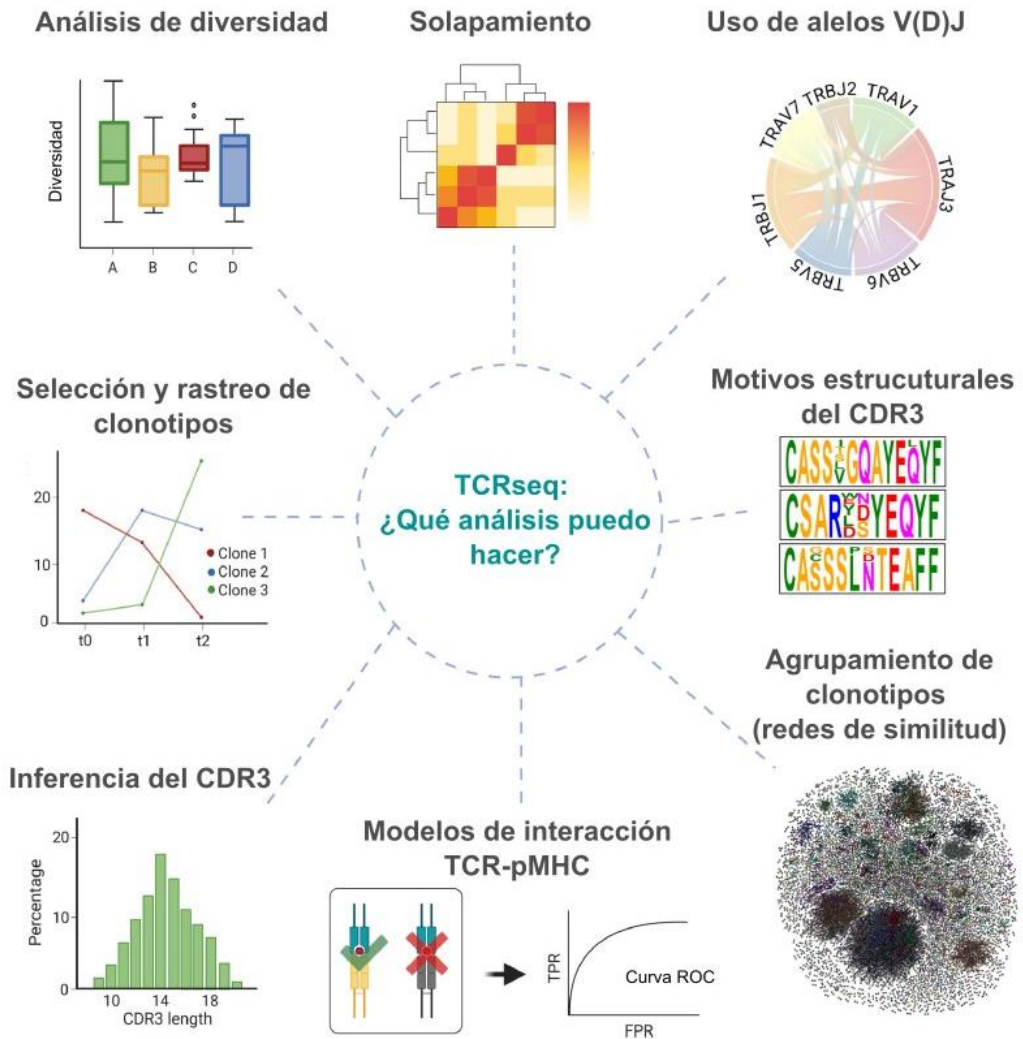


Figure 1.5. Análisis posibles con TCR-seq. La secuenciación de receptores de células T (TCR-seq) permite una variedad de análisis, incluyendo la evaluación de la diversidad clonal, el solapamiento entre repertorios, el uso de alelos V(D)J, la identificación de motivos estructurales del CDR3, el agrupamiento de clonotipos basado en redes de similitud, la inferencia del CDR3, la modelización de interacciones TCR-pMHC y el seguimiento dinámico de clonotipos a lo largo del tiempo. Figura modificada a partir de la original publicada por Valkiers y colaboradores [78].

Un elemento que tienen en común el cálculo de las métricas mencionadas anteriormente es que son catalogadas como ‘*sequence agnostic*’, es decir, no realizan una inferencia directa sobre la secuencia de nucleótidos o aminoácidos del CDR3 de cada clonotipo [128]. De forma más reciente, se han desarrollado nuevas herramientas basadas en el agrupamiento (*clustering*) de secuencias CDR3 en base a similitudes o diferencias en su secuencia de aminoácidos y/o propiedades fisicoquímicas de los mismos [121,139–141] (Figura 1.5). El principal objetivo de estas técnicas es la de poder extraer firmas (por ejemplo, motivos estructurales) distintivos en cada conjunto de repertorios analizados. Otro valor añadido es el de agrupar distintas secuencias CDR3 en base a la afinidad a uno o varios antígenos a partir de la consulta de información de fuentes de datos externas [128,142]. Algunos de los algoritmos y/o programas más populares son GLPIH2 [143]. TCRdist [144], GIANA [145] o iSMART [146] y recientemente se han creado nuevas herramientas con funcionalidades como el modelaje 3D y el *docking* del complejo con pMHC [147].

Recientemente, la inferencia en la secuencia del CDR3 ha cobrado mayor importancia con el auge de la inteligencia artificial, se han desarrollado nuevos y sofisticados métodos de análisis de *machine learning* o *deep learning* [128]. Estas nuevas técnicas incluyen nuevas y

diversas funcionalidades que permiten un estudio más profundo y amplio. Algunos de estos programas son DeepTCR [148], MixTCRpred [149] o immuneML [150].

1.5.3 La combinación de *TCRseq* y *CytoF* en favor de la medicina de precisión.

La ventaja y el potencial que supone la aplicación del *TCRseq* en sus diferentes modalidades y métodos analíticos es aún mayor cuando es incorporado dentro de una estrategia multiómica de inmunofenotipado. Dentro de este marco de posibilidades una de las estrategias que más valor añadido aporta al avance de la medicina de precisión es la combinación del *TCRseq* con CyTOF. Son varios los estudios donde se ha hecho una caracterización del sistema inmunitario de los pacientes aplicando estas dos estrategias junto con otras más. Por ejemplo, en el estudio de Napolitani y colaboradores aplicaron inicialmente CyTOF para detectar subpoblaciones muy específicas de linfocitos T CD4⁺ efectores relevantes en la respuesta a la infección por *Salmonella* [151]. Posteriormente, se aplicó un análisis de *TCRseq* sobre esta población específica para detectar clonotipos asociados a antígenos de serovariantes de este patógeno.

Otro ejemplo es el publicado recientemente por Wang y colaboradores [152]. Se realizó un enfoque multiómico, combinando estas

dos técnicas con análisis de *scRNAseq* y *bulkRNAseq* en un estudio de inmunofenotipado a gran escala para estudiar los cambios del sistema inmunitario asociado a la edad. Concretamente, en este enfoque holístico se empleó el CyTOF como técnica de validación experimental de los resultados de interés extraídos de los análisis de *scRNAseq*, a la hora de identificar supopulaciones inmunitarias asociadas a la edad. En los grupos de linfocitos T validados a partir de esta estrategia (principalmente los linfocitos T efectores de memoria $GNLY^{+}CD8^{+}$) se evaluó su grado de expansión clonal por *TCRseq* y el rol de los genes responsables de tal proceso por *RNAseq*. Finalmente, a partir de los datos de inmunifonotipado de las diferentes estrategias empleadas se entrenó un algoritmo de predicción de la edad inmunitaria [152].

En el contexto de la COVID-19 también se ha empleado de forma conjunta TCRseq y CyTOF, aunque de forma puntual. Uno de los mejores ejemplos es el aplicado en el Chu y colaboradores, en donde caracterizan el rol de linfocitos T citotóxicos específicos de SARS-CoV-2. Concretamente se aplicó *bulk β TCRseq* sobre linfocitos T reactivos a COVID-19 y activados por interferón gamma. Junto con la caracterización de estas poblaciones por CyTOF se observó que existían cambios en la diversidad y clonalidad de linfocitos T de memoria efectora, es decir, linfocitos T ya activados previamente [153].

En definitiva, la combinación de *TCR-seq* y CyTOF se orienta fundamentalmente a la caracterización detallada de subpoblaciones específicas de linfocitos T, ofreciendo una visión complementaria y más profunda de las dinámicas de la inmunidad adaptativa. No obstante, hasta donde alcanza nuestro conocimiento, aún no se han desarrollado estudios sobre COVID-19 que integren de manera simultánea ambos conjuntos de datos mediante herramientas de inteligencia artificial. Este enfoque integrador resulta especialmente prometedor cuando se busca establecer vínculos entre las características moleculares de la inmunidad adaptativa (a través de *TCR-seq*) y los perfiles funcionales de la inmunidad innata (analizados mediante CyTOF).

2.Hipótesis y objetivos

2.1 Hipótesis.

La acción de los linfocitos T frente a la infección por SARS-CoV-2 ha demostrado ser crítica para la progresión de la COVID-19. Caracterizar el repertorio de TCR nos permite dar un enfoque muy profundo a las dinámicas de la respuesta inmunitaria de linfocitos T frente al desarrollo del COVID-19.

El estudio por *TCRseq* permitiría comprender la respuesta de linfocitos T en el COVID-19 y su importancia en la progresión de la enfermedad. En España los estudios de *TCRseq* frente a la COVID-19 de grandes cohortes son relativamente escasos. Sumado a la alta variabilidad no sólo entre poblaciones sino también entre individuos, dificulta enormemente la detección de cambios y/o clonotipos específicos y que permitan una estratificación adecuada a un pronóstico más preciso de la COVID-19.

Por otro lado, en el contexto de la COVID-19 todavía son escasos los estudios donde se hace una integración de datos de inmunidad innata (datos CyTOF) e inmunidad adaptativa (*TCRseq*), teniendo en cuenta además otras variables informativas y cualitativas, como son genotipos o sintomatología. Unido al alto nivel de variación en el repertorio de

linfocitos T, urge el poder detectar subfenotipos de pacientes a partir de la integración de datos.

Así, la **hipótesis principal** de esta tesis doctoral es que la población española de pacientes COVID-19 no vacunados presenta un repertorio de TCR con características específicas, vinculadas a un desarrollo grave de la enfermedad y cuya integración con datos de inmunología innata, genotipado y sintomatología permitiría detectar subfenotipos previamente no identificados que nos permitirían estratificar a los pacientes según la gravedad de sus síntomas.

2.1 Objetivos.

El *objetivo principal* de la tesis es investigar los cambios en el repertorio de linfocitos T mediante *bulk-TCR β seq* en una cohorte de 173 pacientes no vacunados con COVID-19 que nos permitan identificar marcadores de predisposición a la patología.

Se han fijado los *objetivos específicos* de este proyecto de tesis doctoral:

- Estudiar los cambios en la dinámica poblacional del repertorio de linfocitos T en función de la gravedad, edad y sexo del paciente.

- Detectar alelos V(D)J asociados a la gravedad de la COVID-19 a partir del cálculo de frecuencias alélicas ponderadas en dichos repertorios.
- Detectar subgrupos de clonotipos reactivos a SARS-CoV-2 a partir de la inferencia sobre la secuencia de aminoácidos del CDR3b, aplicando técnicas de agrupamiento a partir del cálculo de afinidades a antígenos.
- Integrar datos de *bulk-TCR β seq*, CyTOF, genotipado y sintomatología en un modelo de clases latentes (LCM) como técnica de *machine learning* no supervisado para detectar subgrupos de pacientes.
- Estudiar la distribución de observaciones y grupos creados por LCM mediante análisis univariados y multivariados.
- Detectar los clonotipos y residuos críticos en la secuencia de CDR3b entre los grupos creados por LCM, a partir de la inferencia por *deep learning*.
- Filtrar y estudiar residuos críticos, afinidad por antígenos y detección de motivos estructurales de los clonotipos más predictivos, mediante herramientas basadas en alineamiento de secuencias.

3.Materiales y métodos

En el siguiente capítulo describiremos la metodología empleada para llevar a cabo esta tesis doctoral. El contenido está dividido en dos grandes apartados, de acuerdo con los dos principales estudios definidos en el objetivo principal y objetivos específicos presentados anteriormente. En primer lugar, describiremos la metodología empleada para el análisis de *TCRseq* (apartado 3.1). Tras ello, se continuará con la metodología empleada para la integración de datos por modelos de clases latentes y posterior inferencia por *deep learning* (apartado 3.2). Para ambas secciones, se describe de forma detallada la cohorte de estudio, obtención de datos, preprocesado de datos, bases estadísticas de los principales análisis y software empleado.

3.1 Metodología para el análisis del repertorio de linfocitos T

3.1.1 Población de estudio

Se reclutó una cohorte inicial de 174 pacientes. Tras los análisis de preprocesamiento del repertorio de linfocitos T (descrito a continuación), se eliminó una muestra, siendo la cohorte final de 173 pacientes (98 leves y 75 graves). Las muestras fueron tomadas entre 2020 y 2021 desde centros de atención primaria y hospitales de Granada. Cabe señalar que, durante este periodo la situación de saturación de infraestructuras y

consecuentes medidas restrictivas añadieron una mayor dificultad a la hora de poder reunir una cohorte de tales características, en especial en la situación dada en 2020. Para cada muestra se extrajeron muestras de sangre y se almacenaron en tubos “Tempus™ Blood ARN” para la posterior extracción de ARN total. Los síntomas clínicos registrados incluyeron afectación dermatológica, anosmia, ageusia, mialgia, cefalea, fiebre, disnea, astenia y tos. A continuación, se definieron los criterios de clasificación de pacientes leves y graves. Dichos criterios de inclusión se basan en las pautas de clasificación de la OMS [154] y se revisaron periódicamente, con el objetivo de actualizar la base de datos y tratar de tener muestras poblacionales balanceadas según edad, sexo y gravedad. Los pacientes catalogados como “enfermedad leve” se caracterizan por presentar fiebre, malestar, tos, síntomas respiratorios superiores y/o características comunes de COVID-19 (cefalea, pérdida del gusto o del olfato, etc.). Por otro lado, pacientes en el grupo de “enfermedad grave” cumplían con los criterios de admisión hospitalaria y las siguientes características: (i) hipoxia: $SPO_2 \leq 93\%$ al aire atmosférico o $PaO_2:FiO_2 < 300$ mmHg (relación SF < 315); taquipnea: en dificultad respiratoria o FR (frecuencia respiratoria) > 30 respiraciones por minuto; o más del 50% de afectación observada en imágenes del tórax. La infección por SARS-CoV-2 se confirmó mediante prueba positiva de Reacción en Cadena de la

Polimerasa con Transcriptasa Inversa (RT-PCR) o mediante prueba de Inmunoglobulina M (IgM) positiva.

Las diferencias en los síntomas fueron estudiadas según la severidad, el género y el grupo de edad. El protocolo para este estudio fue aprobado por el Comité Ético de Investigación de Granada (CEI Granada) con el código interno 1329-N-21. Se obtuvo el consentimiento informado por escrito de todos los participantes de acuerdo con los principios de la Declaración de Helsinki.

3.1.2 Preparación de librerías y secuenciación de nueva generación (NGS)

La extracción de ARN total de muestras de sangre completa se realizó usando el kit "Tempus™ Spin ARN Isolation Reagent" (Applied Biosystems, EE. UU.). Para comprobar su rendimiento se realizó un control de calidad calculando el valor de integridad del ARN (RIN) y se normalizó a 1 µg total. Las librerías de secuenciación se prepararon utilizando el kit "AmpliSeq for Illumina Immune Repertoire Plus TCR beta Panel" (Illumina, San Diego, CA, EE. UU), un kit de secuenciación dirigido altamente multiplexado para medir la diversidad de células T y la expansión clonal a través de la secuenciación del TCRβ.

La preparación de la librería comienza con la conversión del ARN total a cADN, seguido por un protocolo basado en la reacción en cadena de la polimerasa (PCR) para amplificar las regiones objetivo de las muestras de cADN y la digestión parcial del amplicón. En el siguiente paso, los adaptadores Índice 1 (i7) e Índice 2 (i5) se ligan a cada muestra, preparando las librerías para la hibridación en una celda de flujo. Posteriormente, las librerías se enriquecen mediante PCR y finalmente se normalizan, se agrupan y se secuencian utilizando secuenciación de nueva generación (NGS).

La secuenciación masiva de librerías de ARN multiplexadas se realizó en la plataforma NextSeq 500 (Illumina, San Diego, CA, EE. UU.), utilizando una celda de flujo de alto rendimiento con secuenciación tipo *paired-end* (150 bp x 2) y una profundidad final de aproximadamente 1 millón de lecturas por librería.

3.1.3 Preprocesamiento de datos

Después de la secuenciación de las cadenas TCR β , los archivos FASTQ con lecturas fueron procesados en la misma plataforma Illumina usando el software MiXCR v4.0.0 [134]. En este marco y usando la configuración predeterminada, las lecturas se filtraron y luego se alinearon a los segmentos variable (V), diversidad (D), unión (J) y constante (C) del

CDR3b para ensamblar clonotipos.. Para cada muestra, se genera un archivo de repertorio que contiene los alelos V(D)J anotados para cada clonotipo, su abundancia basada en lecturas anotadas (recuentos, también llamado clones), frecuencia y secuencias de aminoácidos y nucleótidos del correspondiente clonotipo. Posteriormente, se empleó VDJtools v1.2.1 para un segundo paso de filtrado del repertorio [155]. Este proceso involucró la eliminación de clonotipos no funcionales (aquellos con codones de parada o cambios en el marco de lectura) usando la función *FilterNonFunctional negative*. Posteriormente, se eliminaron clonotipos con alelos diferentes a los presentes en TRB, ya que había clonotipos con alelos del CDR3 de la cadena α del TCR (TRAV/J) e inmunoglobulinas (IGV/D/J). Para ello se utilizó la función *FilterBySegment -negative*, especificando los alelos detectados.

Para abordar problemas relacionados con diferencias en el tamaño de la librería y facilitar comparaciones entre muestras, empleamos el submuestreo (también llamado *downsampling*) para normalizar los repertorios. El *downsampling* ajusta el tamaño del repertorio al valor de la profundidad de secuenciación más pequeña aceptable. Para determinar este valor aceptable, calculamos la profundidad de secuenciación de cada repertorio basado en la suma total del recuento, identificando muestras clasificadas como valores atípicos inferiores utilizando el criterio de

Tukey. De no eliminar estas muestras atípicas, el *downsampling* a estos valores conllevaría una pérdida excesiva de clonotipos raros y una reducción drástica en la frecuencia de los que quedan. Detectamos únicamente una muestra clasificada como un valor atípico inferior, la cual fue descartada. De esta forma, el número de final de repertorios (uno por muestra) que pasaron a los análisis secundarios pasaron de 174 a 173. Tras ello, se eligió el valor de profundidad de secuenciación de la segunda muestra más pequeña (334699 clones), la cual presenta el mínimo sin ser valor atípico. Si bien el *downsampling* reduce los recuentos proporcionalmente en clonotipos con recuentos más altos y elimina clonotipos únicos (1 recuento), también facilita comparaciones métricas directas, que estarían sesgadas al comparar repertorios con profundidades de secuenciación diferentes. Aunque los clonotipos de baja frecuencia pueden verse afectados, su pérdida es menos crítica al analizar métricas de diversidad y clonalidad, ya que éstas se encuentran dominadas por clonotipos expandidos [156].

Esta técnica de normalización se aplicó a los 173 repertorios finales usando la función *DownSample size*. Finalmente, los repertorios procesados fueron importados a R v4.1.3 para análisis posteriores usando *immunarch* v.0.6.0 [157].

3.1.4 Análisis de diversidad y clonalidad

Para cada muestra, se calcularon índices de riqueza, α -diversidad y clonalidad ampliamente utilizados en ecología y análisis de repertorio de TCR [158,159]. La riqueza de clonotipos por repertorio se calculó utilizando el estimador Chao1. Para la α -diversidad, se utilizaron el índice de Shannon-Wiener normalizado (índice tipo 1) y el índice de Gini-Simpson (índice tipo 2). La diferencia entre los dos radica en la mayor sensibilidad del índice de Shannon-Wiener a la presencia de clonotipos raros, mientras que el índice de Gini-Simpson da mayor peso a los clonotipos más abundantes [158]. Para evaluar la clonalidad, se aplicaron el índice de Diversidad 50 (DE50) y el coeficiente de desigualdad de Gini. Los cálculos para cada índice se describen a continuación:

El estimador Chao1 viene dado por la siguiente fórmula:

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

Donde S_{obs} es el número observado de especies, n_1 es el número de *singletons* (clonotipos con recuento = 1), n_2 es el número de *doubletons* (clonotipos con recuento = 2) [159].

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

Donde N es el número de clonotipos únicos y p_i es la frecuencia del clonotipo i dentro de un repertorio. Los valores van de 0 (diversidad mínima con dominio de clonotipos hiperexpandidos) a 1 (diversidad máxima) [158].

El índice de Gini-Simpson se deriva del índice de Simpson, que representa la probabilidad de que dos clones seleccionados al azar con reemplazo pertenezcan al mismo clonotipo. Se resta 1 para orientar valores altos hacia alta diversidad. Valores cercanos a 0 indican diversidad mínima con alta oligoclonalidad, mientras que valores cercanos a 1 representan diversidad máxima y policlonalidad [158].

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

El índice DE50 es un indicador de clonalidad que representa el número mínimo de clonotipos requeridos para cubrir el 50% de los recuentos. Valores bajos de DE50 indican oligoclonalidad y dominio de clonotipos hiperexpandidos, mientras que valores altos sugieren una distribución de clonotipos más policlonal [158,159].

El coeficiente de Gini también se usa para estudiar la heterogeneidad del repertorio basada en la clonalidad. Cuantifica el equilibrio de un sistema basado en la equidad en la distribución de

frecuencias de sus elementos. Los valores van de 0 (todos los clonotipos están en proporciones iguales) a 1 (repertorios oligoclonales con la proporción concentrada en una o muy pocas secuencias) [158].

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

Donde p_i y p_j son las frecuencias de los clonotipos respectivos i y j en el repertorio, $\bar{p}$ es la frecuencia media de clones y N es el número total de clonotipos únicos [158].

Chao1, Gini-Simpson, DE50 y las métricas de Gini se calcularon utilizando la función `repDiversity()` de Immunarch con la secuencia de aminoácidos CDR3b como clonotipo (argumento `.col='aa'`). El cálculo del índice de Shannon-Wiener normalizado se realizó utilizando la función `CalcDiversityStats` de VDJtools.

3.1.5 Análisis de solapamiento de repertorios

Para el análisis de solapamiento de repertorios, se utilizó la función `repOverlap` de Immunarch. Se siguieron dos enfoques basados en la selección de clonotipos: primero, se estudió la superposición considerando la secuencia de aminoácidos CDR3b de cada clonotipo, el alelo V y el alelo J (argumento `.col="aa+v+j"`). En segundo lugar, se realizó el mismo análisis basándose únicamente en la secuencia de aminoácidos CDR3b. En

ambos enfoques, se calcularon tres parámetros ampliamente utilizados en análisis de repertorio de TCR para los clonotipos totales en todas las muestras y por grupos.

Clonotipos públicos: El número de clonotipos compartidos entre dos repertorios.

$$\text{Clonotipos unicos} = A \cap B$$

Siendo A y B los clonotipos identificados en los repertorios A y B , respectivamente [157].

Índice de Jaccard: Definido como el número de clonotipos superpuestos dividido por el tamaño de la unión de los dos repertorios comparados .

$$\text{Indice Jaccard} = J(i, j) = \frac{c_{ij}}{N_i + N_j - c_{ij}}$$

Donde C_{ij} es el número de clonotipos superpuestos y N_i y N_j son el número total de clonotipos en los repertorios i y j respectivamente. El índice de Jaccard varía de 0 (sin similitud) a 1 (similitud completa) [158,159] .

Índice de superposición de Morisita: Este índice compara la distribución actual de clonotipos con una distribución aleatoria esperada [158].

$$\text{Índice de solapamiento de Morisita} = \frac{2 \sum_{i=1}^S x_i y_i}{(D_x + D_y)XY}$$

Para dos repertorios x e y con un total de X e Y clonotipos, x_i y y_i representan el número de veces (recuentos) que el clonotipo i aparece en las muestras x e y respectivamente. D_x y D_y son los índices de diversidad de Simpson para los repertorios x e y . S es el número total de clonotipos únicos en ambos repertorios. Similar al índice de Jaccard, el índice de Morisita varía de 0 (sin superposición) a 1 (superposición completa), con la ventaja adicional de que considera la frecuencia relativa de cada clonotipo, añadiendo robustez al resultado [160].

Además, la superposición de repertorios se estudió seleccionando los N clonotipos más abundantes basándose en su abundancia (top 100, 500, 1000, 5000, 10000, 50000), usando el argumento. `step` en `repOverlap`. Para todas las selecciones, se calculó el grado promedio de superposición a partir de los tres tipos de cálculos descritos para cada agrupación de pacientes en función de severidad, grupo de edad y sexo.

3.1.6 Análisis de uso de alelos V y J.

Para el cálculo del uso de los alelos V y J por separado, se empleó la función `geneUsage` de `Immunarch`, excluyendo clonotipos con anotación ambigua (argumento. `ambig = "exc"`), es decir, diferentes alelos V y J para la misma secuencia de aminoácidos CDR3b. Primero, el uso de genes se calculó basado en la frecuencia relativa de cada alelo, ponderada por la abundancia de recuentos de clonotipos donde se encuentra (argumento. `quant="count"`). La matriz resultante con la frecuencia relativa de cada alelo (columnas) en cada repertorio (filas) se utilizó para pruebas de comparación de grupos. En segundo lugar, se calculó la misma matriz, pero esta vez se emplearon los recuentos de cada alelo ponderados en lugar de su frecuencia. La matriz resultante de este segundo enfoque se estandarizó utilizando el método de z-score, y se empleó para el agrupamiento jerárquico y la representación en heatmaps; uno para los alelos V y otro para los alelos J. El agrupamiento jerárquico se realizó a partir del cálculo de distancias euclídeas.

Para las matrices de recuentos de alelos V y J, también se calculó el índice de Divergencia de Jensen-Shannon utilizando la función `geneUsageAnalysis`. Este parámetro proporciona una medida de similitud entre dos distribuciones de probabilidad, en este caso, las distribuciones de los diferentes alelos V o J entre dos muestras [159]. Para dos

distribuciones del mismo alelo en repertorios X e Y , el índice JSD se calcula de la siguiente manera:

$$JSD(A||B) = 0.5 * (KLD(A||M) + KLD(B||M))$$

Donde $M = 0.5 * (A + B)$ - el punto medio entre las dos distribuciones (A y B) de los alelos V o J y $KLD(A||M)$ y $KLD(B||M)$ son la Divergencia de Kullback-Leibler de A y M y B y M respectivamente. El índice JSD varía de 0 a 1, siendo 0 la máxima igualdad de distribución [7].

La matriz simétrica resultante de divergencias que compara todos los pares de repertorios (173 x 173) se representó como un mapa de calor junto con el agrupamiento jerárquico.

3.1.7 Agrupamiento y selección de clonotipos mediante análisis de motivos

Los análisis anteriores permiten el estudio de la arquitectura y dinámica del repertorio de TCR pero no la inferencia directa sobre la secuencia CDR3b en sí. El estudio directo de la secuencia de aminoácidos CDR3b ayudaría a identificar grupos de clonotipos con motivos/firmas estructurales compartidos o similares entre diferentes grupos de pacientes, los cuales influyen en la especificidad de la unión a un cierto número de antígenos. Para agrupar clonotipos susceptibles de tener firmas con

especificidades de unión a antígenos similares, se seleccionaron los 100 clonotipos más abundantes en cada uno de los 173 repertorios utilizando el algoritmo GLIPH2 (Grouping of Lymphocyte Interaction by Paratope Hotspots versión 2) a través del paquete turboglyph v0.99.2 [143,161]. Para el agrupamiento, el algoritmo toma como referencia una base de datos preconstruida de secuencias CDR3b presentes en linfocitos T CD8 y CD4. Para nuestro análisis, se utilizó la referencia combinada de linfocitos CD4 y CD8 naïve proporcionada en <http://50.255.35.37:8080/>. Concretamente, se utilizó la función `glyph2()` para el agrupamiento, seleccionándose aquellos motivos presentes en al menos 5 secuencias CDR3b distintas (argumento `cluster_min_size=5`) y con una longitud de 3 a 5 aminoácidos (argumento `motif_length=c(3,4,5)`). Los grupos resultantes, significativamente representados por alineación global o local (`fisher.score<0.05`), fueron seleccionados para representación gráfica utilizando la función `plot_network`.

Las secuencias de aminoácidos CDR3b que forman parte de los grupos seleccionados por GLIPH2 se sometieron a un segundo proceso de filtrado. Se seleccionaron aquellas secuencias CDR3b cuya unión a antígenos SARS-CoV-2 ha sido validada experimentalmente a través del ensayo MIRA (Multiplex Identification of T cell Receptor Antigen Specificity) de ImmuneCODE [162]. El conjunto de datos MIRA incluye

más de 135,000 TCRs específicos de SARS-CoV-2 altamente confiables [162]. Los clonotipos con sus respectivas secuencias CDR3b encontradas en MIRA se clasificaron como positivos para COVID-19 (COVID-19+).

Para estudiar la distribución de motivos entre grupos de pacientes, se seleccionaron clonotipos COVID-19+ de los 173 repertorios finalmente estudiados, y se reagruparon según su grupo GLIPH2. Se generó una matriz de recuentos para cada motivo por repertorio. Posteriormente, se estudió la presencia/ausencia de grupos entre muestras (recodificando la matriz a valores de 0=no recuentos y 1=al menos 1 recuento) y su frecuencia (dividiendo el valor de recuento de cada grupo por el total de recuentos de todos los grupos en cada repertorio). Para visualizar las firmas estructurales de secuencias CDR3b pertenecientes a grupos con diferencias significativas entre grupos de edad y gravedad, se utilizó la herramienta WebLogo v2.8.2 [163].

Finalmente, se utilizaron datos filtrados por motivos para construir modelos de regresión logística para seleccionar los motivos más relevantes y así distinguir grupos de pacientes basados en divisiones de gravedad y edad establecidas en este estudio. Para este propósito, se realizó una simulación de Monte Carlo con 1000 iteraciones utilizando scripts R propios. En cada iteración, los datos de frecuencia filtrados por motivos se dividieron en entrenamiento del 75% y una división de prueba del 25% y

se construyó un modelo logístico binomial. Luego, el modelo se sometió a una selección de características paso a paso empleando el Criterio de Información de Akaike (AIC) para optimizar el ajuste del modelo. El AIC sigue la fórmula:

$$AIC = -2 \times \text{loglikelihood} + 2 \times \text{number of parameters}$$

Donde loglikelihood es el logaritmo de la función de verosimilitud del modelo y número de parámetros es el recuento de parámetros estimados en el modelo, incluyendo el intercepto y los coeficientes de las variables predictoras [164].

Después de esto, se hicieron predicciones con el modelo seleccionado en cada iteración y se calculó la curva ROC con su correspondiente Área Bajo la Curva (AUC) utilizando los datos originales. Tras la finalización de la validación cruzada de Monte Carlo de 1000 veces, se calculó el AUC promedio y se registró el número de veces que cada motivo filtrado fue considerado como una variable relevante en los modelos de regresión logística. Una mayor frecuencia de consideración indicó una mayor importancia de la característica para distinguir grupos de pacientes.

3.1.7 Análisis estadístico

Para las variables de sintomatología, se aplicó la prueba de independencia Chi-cuadrado (χ^2) y se calculó el tamaño del efecto utilizando V de Crammer. Para las variables de suero, métricas de diversidad, clonalidad y frecuencias, se verificó la normalidad utilizando la prueba de Lilliefors-Kolmogorov-Smirnov. Para distribuciones no normales, se utilizó la prueba U de Mann-Whitney para comparaciones entre dos grupos y la prueba de Kruskal-Wallis para comparaciones entre tres o más grupos, con la prueba de Dunn para comparaciones múltiples. Para comparaciones entre dos distribuciones normales, se utilizó la prueba t. Para tres o más grupos, se aplicó ANOVA con la prueba t o la prueba de Games-Howell para varianzas iguales o desiguales, respectivamente. El P valor resultante de comparaciones múltiples se ajustó utilizando el método Bonferroni-Holm. Se calcularon coeficientes de tamaño de efecto para cada prueba, incluyendo biserial-rangado (rrb), d de Cohen (d), épsilon cuadrado (E_R^2) y omega cuadrada (ω_p^2) para Mann-Whitney, prueba t, Kruskal-Wallis y ANOVA, respectivamente. Todas las pruebas fueron de dos colas y se consideró significativa una $P < 0.05$. Todos los análisis y la visualización de datos se realizaron utilizando R v4.1.3. El paquete nortest v1.0.4 se utilizó para pruebas de normalidad, el paquete rstatix v0.7.2 para comparaciones de grupos y el paquete effsize v0.8.3 de easystats v0.6.0

para cálculos de tamaño de efecto [165]. Se generaron gráficos de caja y violín utilizando ggpubr v0.6.0 y los mapas de calor y el agrupamiento jerárquico se realizaron utilizando pheatmap 1.0.12. Los modelos de regresión logística binomial se construyeron utilizando la función glm. La función stepAIC del paquete MASS v.7.3 se especificó con el argumento direction establecido en "both", permitiendo la selección de características a través de eliminaciones y adiciones de variables al modelo [166].

3.1.8 Acceso a datos y código

Los scripts de código utilizados en este estudio están disponibles en GitHub:<https://github.com/fmarinb/bulk-TCR-beta-chain-repertoire-analysis-COVID-19>. Los conjuntos de datos de secuenciación de TCR β utilizados en este proyecto están depositados en el *European Genome-Phenome Archive* (EGA) con el número de acceso EGAD50000000477.

3.2 Metodología para la integración de datos CyTOF, *TCRseq*, genotipado y sintomatología y posterior inferencia por *deep learning* de las secuencias CDR3b.

3.2.1 Población de estudio

Se incluyeron un total de 61 pacientes (33 leves y 28 graves) con una edad media de 51 años que se recuperaron de COVID-19, constituyendo una subcohorte proveniente del análisis del repertorio del primer estudio. Por tanto, todos ellos se les realizaron análisis de TCR-seq, junto con análisis CyTOF y genotipado. Todas las muestras recolectadas fueron confirmadas en el diagnóstico de COVID-19 por RT-PCR y serología IgG positiva. Los criterios de inclusión y exclusión son los mismos que los descritos en la sección 3.1.1

3.2.2 Preprocesamiento de datos

Para los 61 pacientes, se generaron y analizaron datos de TCR-seq, CyTOF y genotipado en estudios separados como parte de cohortes más grandes. En total hay 143 variables, siendo 129 cuantitativas (métricas de *TCRseq* y CyTOF) y 14 cualitativas (sintomatología y genotipado). Para el análisis de agrupamiento y multivariante, las variables cuantitativas se

escalaron usando el z-score mientras que las variables cualitativas fueron factorizadas.

3.2.2.1 Repertorios de células T

Los datos del repertorio de células T se obtuvieron de 61 pacientes en el estudio de *TCRseq* y se analizaron en una cohorte más grande de 173 pacientes. Los detalles de este análisis se describieron en la sección anterior.

3.2.2.2 Recuento de células inmunitarias por CyTOF

Se obtuvieron métricas CyTOF de un estudio llevado a cabo por nuestro grupo, en el cual se caracterizaron las diferentes poblaciones de células de sistema inmunitario, con especial interés en la línea monocito-macrófago [167]. Dicho estudio formó parte proyecto de tesis doctoral de la doctora Silvia Martínez Diz, investigadora adscrita a nuestro grupo de investigación. El artículo queda enmarcado dentro del mismo proyecto de investigación en el que queda englobada esta tesis (PECOVID-0006-2020).

El protocolo queda resumido a continuación. En primer lugar, las muestras de sangre recolectadas para análisis de citometría de flujo fueron procesadas fijando las células sanguíneas en estabilizador proteómico PROT1, congeladas a -80°C hasta su tinción. En segundo lugar, las

muestras se diluyeron, lisaron, tiñeron y fijaron antes de ser agrupadas en un solo tubo. Los antígenos de superficie se tiñeron usando un cóctel de anticuerpos seguido por fijación y tinción de ADN durante la noche. Después del fenotipado, las muestras se congelaron y descongelaron para el análisis CyTOF usando un panel CyTOF de 15 marcadores orientados a la detección de subpoblaciones de monocitos y macrófagos en un citómetro de masas Helios®. Las muestras se filtraron a través de tubos capilares de tamiz celular de 35µm y se resuspendieron en Buffer de Adquisición de Células Maxpar. La adquisición se realizó a una tasa de 250–300 eventos por segundo usando el software CyTOF versión 6.7.1016. Tras aplicar el protocolo descrito, se obtuvo una matriz de proporciones de expresión fenotípica (en tantos por ciento) de un conjunto de 37 subpoblaciones celulares ([Tabla Suplementaria 9.1](#))

3.2.2.3 Datos de genotipado.

Los datos de genotipado de los 61 pacientes fueron extraídos de un estudio previo de nuestro grupo en el cual se analizaron un total de 330 muestras [168]. Se extrajo ADN genómico de sangre total usando el kit 'RealPure SSS' de Durviz. La cuantificación del ADN se realizó usando el fluorómetro Qubit™ 3.0 de Invitrogen™ por Thermo Scientific y el sistema nanodrop 2000 de Thermo Scientific, EE.UU. Se verificó la relación 260/280 como control de calidad. El ADN extraído se almacenó

a -20°C hasta el genotipado. Para seleccionar los SNPs, consultamos datos del sitio web de NCBI sobre artículos publicados relacionados con COVID-19 y *TMPRSS2* hasta 2022. Seleccionamos aquellos SNPs con una frecuencia alélica menor superior al 10% en la población caucásica de la base de datos Ensembl. Específicamente seleccionamos *ACE2* (rs2285666), *MXI* (rs469390) y *TMPRSS2* (rs2070788) para el análisis presente. Realizamos el genotipado de ADN utilizando la TaqMan® Genotyping Master Mix de Applied Biosystems. El sistema PCR a tiempo real 7900HT de Applied Biosystems se utilizó para realizar pruebas de discriminación alélica.

3.2.2.3 Datos de sintomatología

Se registraron síntomas clínicos por su presencia o ausencia, incluyendo afectación dermatológica, anosmia, ageusia, mialgia, dolor de cabeza, fiebre, disnea, astenia, respuesta inflamatoria sistémica y tos.

3.2.3 Agrupamiento no supervisado de datos mixtos mediante modelo de clase latente (LCM)

Tras escalar el conjunto de datos, este fue sometido a agrupamiento no supervisado utilizando el paquete VarSelLCM [136]. VarSelLCM es un paquete de R que permite la selección completa de modelos en agrupamiento basado en modelos, incluyendo la identificación de

características relevantes y la selección del número de clústeres. El protocolo LCM tiene una funcionalidad útil para seleccionar variables relevantes para la división de clústeres, lo que es particularmente beneficioso dada la gran cantidad de variables iniciales. La función de densidad de probabilidad de la distribución de mezclas de LCM es:

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

En donde cada observación x_i proviene de una mezcla de G componentes y la relevancia de una variable se define por el vector binario ω : si $\omega_j = 0$, la variable es irrelevante e independiente de la mezcla, mientras que si $\omega_j = 1$, es informativa y depende de las clases latentes. La función de probabilidad incluye tres términos clave: la probabilidad de las variables irrelevantes ($f(x_i|m, \theta) = \prod_{j \in \Omega^c} h_j(x_{ij}|\alpha_{1j})$), la suma ponderada de probabilidades de cada clase latente ($\sum_{g=1}^G \tau_g$), y la probabilidad de las variables relevantes dentro de cada clase ($\prod_{j \in \Omega} h_j(x_{ij}|\alpha_{gj})$) [136,169]. Esto permite identificar automáticamente las variables útiles para el clustering y estimar el número óptimo de clusters. Además, LCM aplica selección de características utilizando el Criterio de Información Bayesiano (BIC), que penaliza modelos con más parámetros, fomentando la selección de modelos más simples que son menos propensos a

sobreajustar [136]. Para lograr una estratificación de severidad más integral, construimos 100 modelos LCM-BIC y seleccionamos el modelo con el índice Rand ajustado (ARI) más alto de la división inicial por severidad y el grupo combinado de severidad y edad. El ARI calcula una puntuación de similitud considerando todos los pares de muestras y contando pares que están asignados al mismo o diferentes clústeres en ambos el agrupamiento predicho y real. Su fórmula es la siguiente [170]:

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

Donde RI es el índice de Rand, el cual se expresa como:

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

En la fórmula de RI, a es el número de veces que un par de elementos pertenece al mismo clúster a través de dos métodos de agrupamiento. Para b el número de veces que un par de elementos pertenecen a clústeres diferentes a través de dos métodos de agrupamiento y n es el número de pares no ordenados en un conjunto de n elementos. El índice ARI toma valores entre 0 (los dos métodos de agrupamiento no están de acuerdo) hasta 1 (los dos métodos de agrupamiento están perfectamente de acuerdo respecto al agrupamiento de cada par de

elementos). El modelo LCM-BIC seleccionado ha sido guardado como un objeto. RData para uso futuro.

3.2.4 Análisis factorial de datos mixtos (FAMD)

Se definieron dos conjuntos de datos: el 'conjunto de datos pre-LCM' contenía todas las variables y no fue sometido a agrupamiento por LCM-BIC, mientras que el 'conjunto de datos post-LCM' resultó de aplicar LCM-BIC y seleccionar las 70 variables principales. Ambos conjuntos se sometieron a análisis de correspondencias múltiples (FAMD) utilizando los paquetes FactoMineR y factoextra [171,172]. Las observaciones fueron trazadas en un plano factorial con dos dimensiones principales para identificar agrupaciones de muestras en un espacio de menor dimensión. Se calcularon los valores propios y el porcentaje de varianza explicada para estas dimensiones. La contribución de las variables principales a cada una de las dos dimensiones principales fue evaluada calculando el coseno cuadrado ($\cos^2$) para las 20 variables principales con el mayor porcentaje de contribución. El valor de $\cos^2$ representa la proporción de varianza en la variable original que es explicada por el plano factorial [172]. Un valor cercano a 1 indica una alta representación y contribución significativa de la variable al plano factorial. También se generaron biplots para las variables cuantitativas para estudiar eficazmente su varianza y grado de correlación.

3.2.5 Análisis de *deep learning* de Repertorios de TCR

Para un análisis más integral de las secuencias CDR3b, utilizamos el paquete de Python DeepTCR [148]. Este paquete proporciona herramientas especializadas para el análisis de *deep learning* de repertorios, tomando como entrada las secuencias CDR3, sus correspondientes alelos V y J, y datos de tipo de antígeno HLA. Aplicamos el análisis de *deep learning* tanto supervisado como no supervisado a los 1000 clonotipos expandidos principales de cada muestra en el conjunto post-LCM. Para el análisis no supervisado, utilizamos el algoritmo de *Variational Autoencoders* (VAE) con la función Train_VAE. Se seleccionó el número de características latentes que explican el 99% de la varianza total para simplificar el proceso de entrenamiento del modelo VAE y se eligió un tamaño compacto para las 3 redes convolucionales.

Para evaluar la capacidad predictiva del modelo VAE, utilizamos el Clasificador de Repertorio de Vecino Más Cercano (KNN_Repertoire_Classifier) con distancia euclidiana. Realizamos 50 divisiones con una división de entrenamiento del 75% y prueba del 25%. Los valores del área bajo la curva (AUC) para cada clase se representaron visualmente usando gráficos de violín. Para el análisis supervisado, se utilizó Monte_Carlo_CrossVal para modelar comparaciones de pares de clases. Se realizaron cien simulaciones de Monte Carlo con una división

de entrenamiento del 75% y prueba del 25% para la validación cruzada. Los parámetros de la arquitectura del modelo consisten en un tamaño de red reducido con 12, 32 y 64 neuronas para las capas respectivas. El modelo se entrenó utilizando información de secuencias de aminoácidos CDR3, alelos V y J con parámetros predeterminados. Para mejorar aún más el rendimiento de la red neuronal y acelerar el proceso de entrenamiento, se habilitó el abandono de múltiples muestras. El hiperparámetro `num_of_concepts` se ajustó a 64 para gestionar eficazmente la heterogeneidad entre repertorios. Para mejorar la velocidad de entrenamiento y promover una mayor regularización, submuestreamos con confianza 100 secuencias por repertorio en cada pliegue. Además, en nuestro estudio, ajustamos el modelo utilizando el argumento `weight_by_class` para evitar sesgos debido a diferencias en el número de muestras por clase.

Para identificar residuos críticos y motivos estructurales distintivos para cada clase, seleccionamos aquellos CDR3b con mayor poder predictivo ($AUC \geq 0.99$) para cada clase en los modelos supervisados generados. Luego utilizamos la función Logos de Sensibilidad de Residuos (RSLs) para crear logotipos para los 25 predictores principales con el mayor poder predictivo. Para detectar motivos estructurales significativamente enriquecidos entre los CDR3b más relevantes,

aplicamos la función `Motif_Identification` en el modelo supervisado generado.

3.2.5 Detección de Secuencias Específicas de SARS-CoV-2

Las secuencias CDR3 top definidas en la sección anterior fueron procesadas para encontrar aquellas catalogadas como reactivas a SARS-CoV-2 y estudiar su especificidad para antígenos virales. Para este propósito se utilizó el software `TCRmatch` utilizando los parámetros predeterminados ($\text{umbral}=0.97$) [173]. Se utilizó la base de datos de Epítomos Inmunes y Herramientas (IEDB) como referencia [174]. Las distribuciones de secuencias totales compartidas y reactivas a SARS-CoV-2 entre los clústeres resultantes se representaron mediante gráficos de pastel y diagramas de intersección. Las secuencias frente al SARS-CoV-2 seleccionadas se estudiaron posteriormente a través de la base de datos IEDB para estudiar la distribución de los antígenos específicos a los que se unen.

3.2.6 Análisis estadísticos

En los análisis FAMD, se realizó la prueba PERMANOVA utilizando la función `adonis2` del paquete `vegan` [175]. Se realizaron 1000 permutaciones con distancia euclidiana. Las comparaciones post-hoc se llevaron a cabo utilizando la función `pairwise.perm.manova` del paquete

RVAideMemorie [176], permitiendo 1000 permutaciones con distancia euclidiana y corrección del P valor con el método Bonferroni-Holm. Para comparar variables cualitativas de interés entre grupos, utilizamos la prueba de chi-cuadrado permitiendo 1000 permutaciones para calcular el P valor. Para comparar dos grupos con variables continuas, utilizamos la prueba de Mann Whitney U. Para comparaciones de dos o más grupos, utilizamos la prueba de Kruskal-Wallis. Para comparaciones post-hoc, aplicamos la prueba de Dunn con el P valor corregido por el método Bonferroni-Holm. Se calcularon coeficientes de tamaño de efecto para cada prueba, incluyendo biserial-rangado (r_{rb}), Cohen's d (d), epsilon-rangado cuadrado (E_R^2) y omega cuadrada (ω_p^2) para Mann-Whitney, prueba t, Kruskal-Wallis y ANOVA, respectivamente. Un p-valor de dos colas menor que 0.05 se consideró significativo. Los análisis que involucran paquetes de R se realizaron con R v 4.1.3.

3.2.7 Acceso a datos y código

El código fuente de los scripts de R y Python (*Jupyter notebooks*) se puede encontrar en https://github.com/fmarinb/COVID19_LCM_integration y en Figshare con el DOI: 10.6084/m9.figshare.26994868. Los conjuntos de datos están disponibles en el EGA con los identificadores EGAD50000000477) (<https://ega-archive.org/datasets/EGAD50000000477>) para el conjunto de repertorios

de TCR (mismo conjunto descrito en 3.1.7) y EGAD50000000840 (<https://ega-archive.org/datasets/EGAD50000000840>) para el dataset conjunto de variables *TCRseq*, CyTOF, genotipado y sintomatología. Hemos desarrollado un sitio web para garantizar que nuestros resultados sean fácilmente accesibles y reproducibles, permitiendo a los usuarios explorar los hallazgos clave, ver gráficos interactivos y acceder a bases de datos curadas: [LCM-BIC immunophenotyping clustering in COVID-19 patients Interactive Results and Databases.](#)

4. Resultados (I). El análisis del repertorio de linfocitos T en pacientes COVID-19 revela cambios asociados con la severidad y la edad

4.1 Antecedentes

Este capítulo constituye el primero de la serie de resultados expuestos en este proyecto de tesis doctoral. Se describirán los principales resultados obtenidos del análisis del repertorio de linfocitos T en la cohorte principal de 173 pacientes COVID-19 recuperados. Se ha abordado el análisis desde el punto de vista de la arquitectura del repertorio y la inferencia directa de la secuencia del CDR3b (consultar sección 1.5.2.4 y la sección 3.1). Entre 2020 y 2021, se reclutaron a 173 pacientes (98 con síntomas leves y 75 graves) en distintos centros de salud y hospitales en Granada. La infección por SARS-CoV-2 se confirmó mediante RT-PCR o pruebas de IgM positivas, y se registraron síntomas de acuerdo a los criterios establecidos por la OMS. Para facilitar la comparación, dividimos a los pacientes en dos grupos de edad: mayores y menores de 55 años.

El repertorio de TCR se analizó mediante AmpliSeq y secuenciación paired-end en NextSeq 500 (~1 millón de lecturas por muestra). Los datos se procesaron con MiXCR para ensamblar clonotipos e identificar alelos V(D)J y regiones CDR3. Se filtraron clonotipos no funcionales con VDJtools y se normalizaron las librerías. Con Immunarch se calcularon métricas de diversidad y clonalidad, y se evaluó el solapamiento entre repertorios. También se analizó el uso de alelos V/J. Los clones más expandidos se agruparon con GLIPH2 y se identificaron

clonotipos específicos de SARS-CoV-2 usando MIRA. Finalmente, se construyeron modelos de regresión para clasificar pacientes por severidad y edad. En el siguiente graphical abstract se presenta un resumen esquemático del protocolo bioinformático realizado (Figura 4.1).

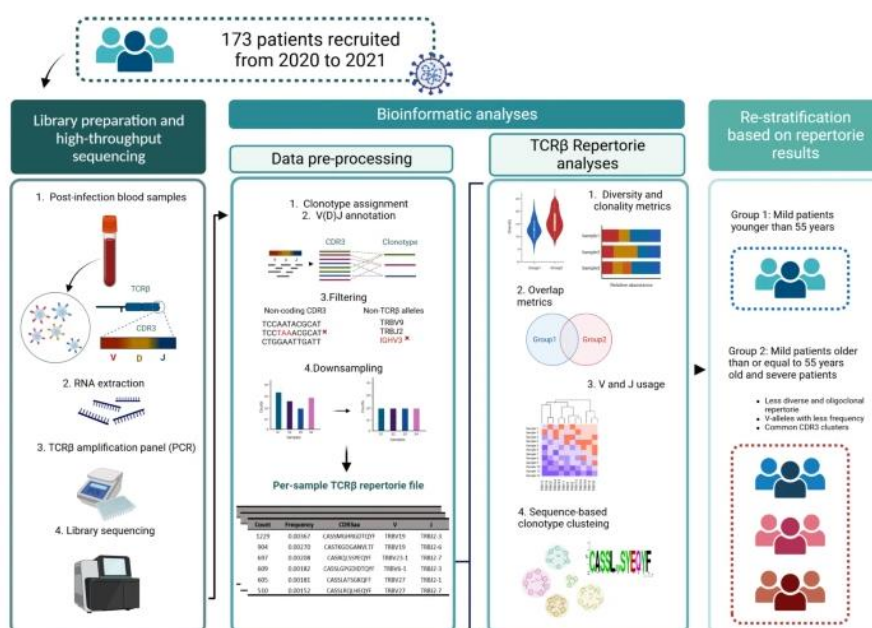


Figure 4.1 Resumen esquemático del primer estudio de esta tesis doctoral.

4.2. Asociaciones entre sintomatología y gravedad, edad y sexo del paciente.

Se puede encontrar una descripción resumida de la cohorte en la [Tabla 1.1](#). Las figuras y tablas suplementarias se encuentran en el capítulo 7: Apéndice I y en el [Material Suplementario Online](#) (MSO). Antes de nuestro análisis del repertorio de linfocitos T, realizamos estudios para investigar los cambios en la distribución de síntomas según la gravedad del paciente, el sexo y el grupo de edad. Los resultados se pueden encontrar en la [Tabla suplementaria 8.1](#). El análisis chi-cuadrado no reveló asociaciones significativas entre la gravedad y las variables de sexo, grupo de edad o el grupo combinado de ambas variables ([Tabla suplementaria 8.2](#)).

Tabla 4.1. Principales variables descriptivas de la cohorte de estudio empleada para el análisis TCRseq.

	Leve (n=98)	Grave (n=75)	Total (n=173)
Severity, n (%)			
Asintomático	37 (37.8%)		37 (21.4%)
Leve	61 (62.2%)		61 (35.3%)
Grave		65 (86.7%)	65 (37.6%)
Grave/Crítico		10 (13.30%)	10 (5.80%)
Sexo, n (%)			
Hombre	45 (45.9%)	40 (53.3%)	85 (49.1%)
Mujer	53 (54.1%)	35 (46.7%)	88 (50.9%)
Edad (yr), mediana [IQR]	52 [14,75]	55 [13,5]	53 [14]
Mayores de 55 años, n (%)			
<55	58 (59.2%)	35 (46.7%)	93 (53.8%)
≥55	40 (40.8%)	40 (53.3%)	80 (46.2%)

Se encontró una asociación significativa entre la gravedad del paciente y la presencia de varios síntomas. Los síntomas más fuertemente asociados fueron fiebre ($p < 0,001$, V de Cramer 0.66), la disnea ($p < 0.001$, V de Cramer 0.59) y la astenia (0.001 y V de Cramer 0.508). Al categorizar según grupos de edad (< 55 y ≥ 55), no se encontraron asociaciones significativas. Sin embargo, hubo una débil asociación entre la fiebre y el género, siendo la fiebre más frecuente en mujeres que en hombres ($p = 0.011$, V de Cramer = 0.197).

4.3 Cambios en la diversidad, clonalidad y solapamiento de repertorios entre grupos de pacientes.

4.3.1 Diversidad y clonalidad.

Para el análisis comparativo de los repertorios de linfocitos T, realizamos una normalización de muestras mediante *downsampling*. Después del *downsampling*, nuestras bibliotecas *TCRseq* produjeron un total de 334,599 clones (Figura Suplementaria 8.1).

Al comparar los estimadores de riqueza, los índices de diversidad y los índices de clonalidad, encontramos diferencias significativas entre pacientes menores de 55 años y aquellos de 55 años o más. (Figura Suplementaria 8.2). Específicamente, para los pacientes de 55 años o más,

observamos un repertorio con una riqueza de especies reducida (Chao1: $p=0.034$, $rrb=0.188$), así como una diversidad general disminuida (Shannon-Wiener normalizado: $p=0.034$, $rrb=0.188$) y una diversidad disminuida hacia clonotipos más expandidos (Gini-Simpson: $p=0.006$, $rrb=0.258$). Esta reducción en la riqueza y diversidad de clonotipos condujo a un repertorio más pobre, evidenciado por una disminución en el índice de Gini (Gini: $p=0.004$, $d=0.450$) y una mayor dominancia clonal (DE50: $p=0.006$, $rrb=0.244$).

No se observaron diferencias significativas al dividir a los pacientes en categorías de leve y severo. Sin embargo, cuando se combinan gravedad y grupos de edad, los pacientes graves menores de 55 años tienen un repertorio menos diverso y más oligoclonal que los pacientes leves del mismo grupo de edad. Además, los resultados indican que la diversidad y clonalidad de este grupo de pacientes es similar a la de los pacientes mayores, ya sean leves o severos. Estas diferencias mostraron un tamaño del efecto moderado para todas las métricas (Chao1: $p<0.001$, $E2R^2=0.071$, Shannon-Wiener normalizado: $p<0.001$, $E2R=0.096$, Gini-Simpson: $p<0.001$, $E2R=0.100$, Gini: $p<0.001$, $E2R=0.107$, DE50: $p<0.001$, $E2R=0.095$) (Figura 4.2). No se encontraron diferencias significativas en las métricas de diversidad basadas en el sexo del paciente, ni en el grupo combinado de sexo y gravedad. Sin embargo, se observaron

diferencias significativas en el grupo combinado de sexo y edad, indicando un repertorio diverso en pacientes femeninas menores de 55 años y un aumento moderado en pacientes masculinos menores de 55 años, sugiriendo una influencia sustancial de la edad al comparar estas métricas (Tabla Suplementaria 8.3)

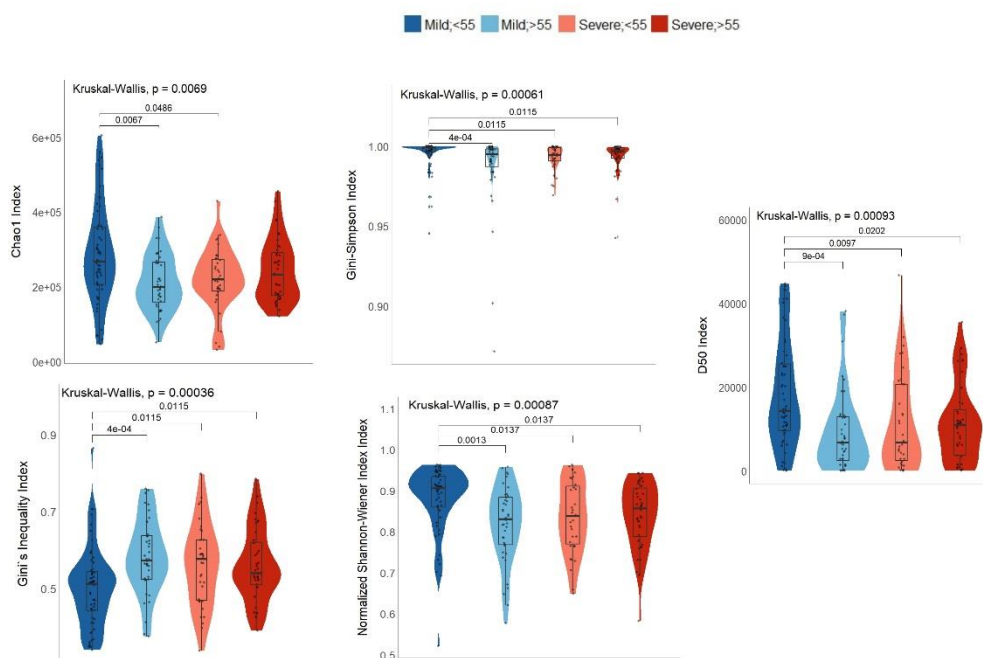


Figura 4.2. Resultados principales en los análisis de α -diversidad y clonalidad para el grupo combinado de edad y severidad. Estimador de riqueza CHAO1, índices de α -diversidad Shannon-Wiener normalizados y el índice Gini-Simpson, índice de clonalidad DE50 y coeficiente de desigualdad Gini para los grupos de pacientes. Para comparaciones de más de dos grupos, se utilizó la prueba de Kruskal-Wallis con la prueba a posteriori de Dunn. Para múltiples comparaciones post-hoc, el P valor se corrigió con el método de Bonferroni-Holm. Nivel de significancia de 0.05.

Los cálculos de las diferentes categorías clonalidad del repertorio de cada paciente muestran patrones de distribución observables en los clonotipos hiperexpandidos y raros ([Figura Suplementaria 8.3](#)). La proporción de clones hiperexpandidos es significativamente mayor en el grupo de ≥ 55 años ($p=0.003$, $rrb=0.261$). También se observó un aumento no significativo en la proporción de clones hiperexpandidos en pacientes severos en comparación con pacientes leves ($p=0.052$, $rrb=0.172$). Como en métricas anteriores, la clasificación combinada de pacientes por gravedad y grupo de edad indica que los pacientes severos menores de 55 años tienen una distribución de clones hiperexpandidos y raros más similar a la de los mayores de 55 años que a los pacientes leves del mismo grupo de edad ([Figura 4.3](#), [Tabla Suplementaria 8.4](#)). Según las comparaciones de sexo, los cambios sólo se observaron en el grupo combinado con edad ([Tablas Suplementarias 8.5 y 8.6](#))

4.3.2 Solapamiento de repertorios

En cuanto a las métricas de solapamiento (clonotipos públicos, índice de Jaccard e índice de Morisita), encontramos un alto nivel de unicidad, considerando un clonotipo la combinación de su secuencia de aminoácidos CDR3 y además al V y J que codifica ([Figura Suplementaria 8.5](#)). Esta unicidad también se reporta al comparar el solapamiento en

diferentes profundidades de repertorio (primeros 50,000, 10,000, 5,000, 1,000, 500 y 100 clonotipos expandidos) y también los hiperexpandidos (Figuras suplementarias 8.6 y 8.7). Cuanto mayor es el grado de expansión clonal, menor es el grado de solapamiento.

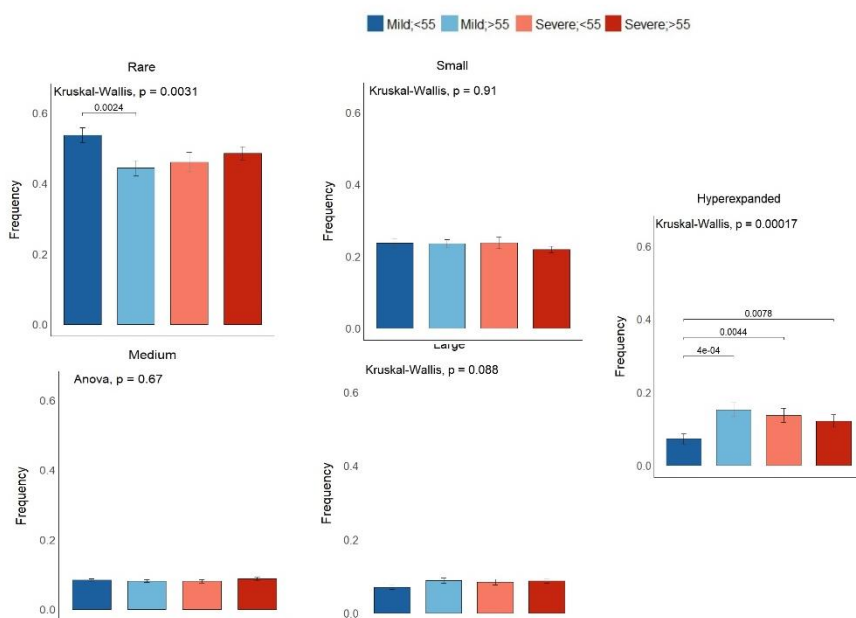


Figura 4.3. Resultados principales en los análisis de expansión clonal para el grupo combinado de edad y severidad. Diferencias en la proporción de clones hiperexpandidos, grandes, medianos, pequeños y raros. La clasificación de clones basada en su expansión clonal se basa en los intervalos de frecuencia de un clonotipo en un repertorio: Raro ($0-10^{-5}$), Pequeño ($10^{-5}-10^{-4}$), Mediano ($10^{-4}-0.001$), Grande ($0.001-0.01$), Hiperexpandido ($0.01-1$). Para comparaciones de más de dos grupos, se utilizó la prueba de Kruskal-Wallis con la prueba a posteriori de Dunn. Para múltiples comparaciones post-hoc, el P valor se corrigió con el método de Bonferroni-Holm. Nivel de significancia de 0.05.

Al comparar el solapamiento entre pacientes dentro del mismo grupo, encontramos que los pacientes leves y severos mantienen cierto grado de solapamiento, particularmente con el índice de Morisita. Estos hallazgos sugieren que el grado de solapamiento es inversamente

proporcional al nivel de expansión clonal. Además, los pacientes leves menores de 55 años mostraron una mayor similitud en sus repertorios, en contraste con los pacientes severos mayores de 55 años. Además, observamos tendencias similares al considerar solo la secuencia de aminoácidos CDR3 como clonotipo.

4.4. Uso de alelos V y J

4.4.1 Frecuencias alélicas ponderadas

Las principales diferencias en la arquitectura del repertorio TCR están relacionadas con la gravedad y la edad. Esta sección estudia la gravedad, la edad y su combinación. El análisis del uso de alelos V reveló cambios significativos en las frecuencias de alelos al examinar la gravedad, los grupos de edad y los grupos combinados ([Figura 4.3](#)). Los resultados indicaron una disminución en la frecuencia del alelo TRBV19 en pacientes graves ($p=0.008$, $rrb=0.236$). Las frecuencias de TRBV12-3 ($p=0.002$, $rrb=0.271$) y TRBV10-3 ($p=0.045$, $rrb=0.177$) están reducidas en el grupo ≥ 55 años. Adicionalmente, las frecuencias de TRBV12-3 ($p=0.002$, $rrb=0.271$) y TRBV10-3 ($p=0.045$, $rrb=0.177$) se encontraron reducidas en el grupo de edad ≥ 55 años. Las frecuencias de TRBV14 exhibieron una disminución significativa entre los pacientes graves ($p=0.029$, $rrb=0.194$) y en aquellos de ≥ 55 años ($p=0.041$, $rrb=0.181$), con

cambios aún más pronunciados observados en TRBV6-4 ($p < 0.001$, $r_b = 0.314$ y $p < 0.001$, r_b). Al comparar los grupos combinados de edad y gravedad, los resultados son consistentes con los cambios previamente reportados en diversidad y clonalidad en las frecuencias de alelos V como TRBV14 ($p = 0.001$, $E2R = 0.09$), TRBV6-4 ($p < 0.001$, $E2R = 0.106$) y TRBV15 ($p = 0.0017$, $E2R = 0.09$) (Figura 4.4)

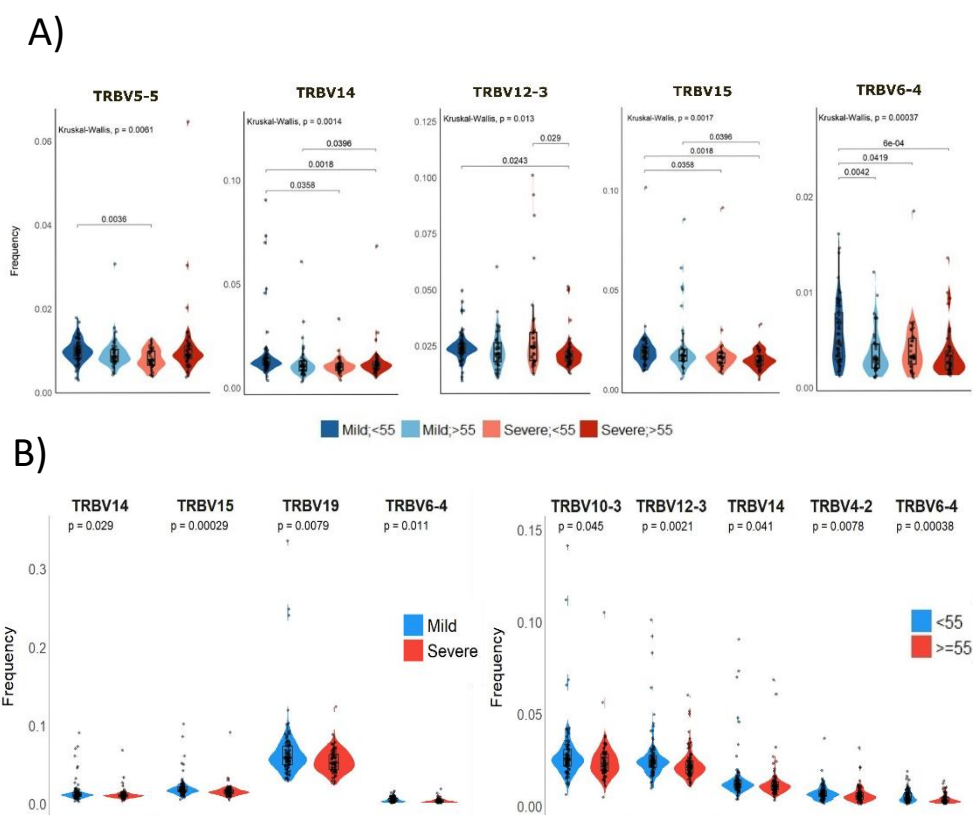


Figura 4.4. Frecuencias significativas de TRBV entre el grupo combinado de edad y severidad. La frecuencia del alelo se calcula ponderando el grado de expansión clonal del clonotipo en el que se encuentra. Para comparaciones de más de dos grupos, se utilizó la prueba de Kruskal-Wallis con la prueba a posteriori de Dunn. Para múltiples comparaciones post-hoc, el P valor se corrigió con el método de Bonferroni-Holm. Nivel de significancia de 0.05.

4.4.2 Distribuciones generales

El clustering jerárquico utilizando los recuentos generales de alelos V y J no diferencia claramente los grupos de pacientes por gravedad, grupo de edad o sexo ([Figura 4.5 A-B](#)). Sin embargo, algunos alelos V y J exhiben un uso más alto en todos los pacientes. TRBV5-1, TRBV20-1, TRBV29-1, TRBV19, TRBV7-2 y TRBV28 tienen la mayor representación, mientras que TRBV9 y TRBV10-3 son más frecuentes en ciertos individuos. TRBJ2-7, TRBJ2-1, TRBJ2-5 y TRBJ2-3 fueron los alelos J más prevalentes, mientras que otro grupo compuesto por TRBJ1-1, TRBJ1-2 y TRBJ2-2 mostró una distribución más diversa. El índice de divergencia de Jentsen-Shannon detectó subgrupos con distribuciones de uso de alelos comparables, específicamente para alelos V ([Figura suplementaria 8.8](#)). Los dendrogramas de clustering jerárquico proporcionaron clasificaciones más claras de clados, aunque la heterogeneidad en gravedad, edad y sexo persiste.

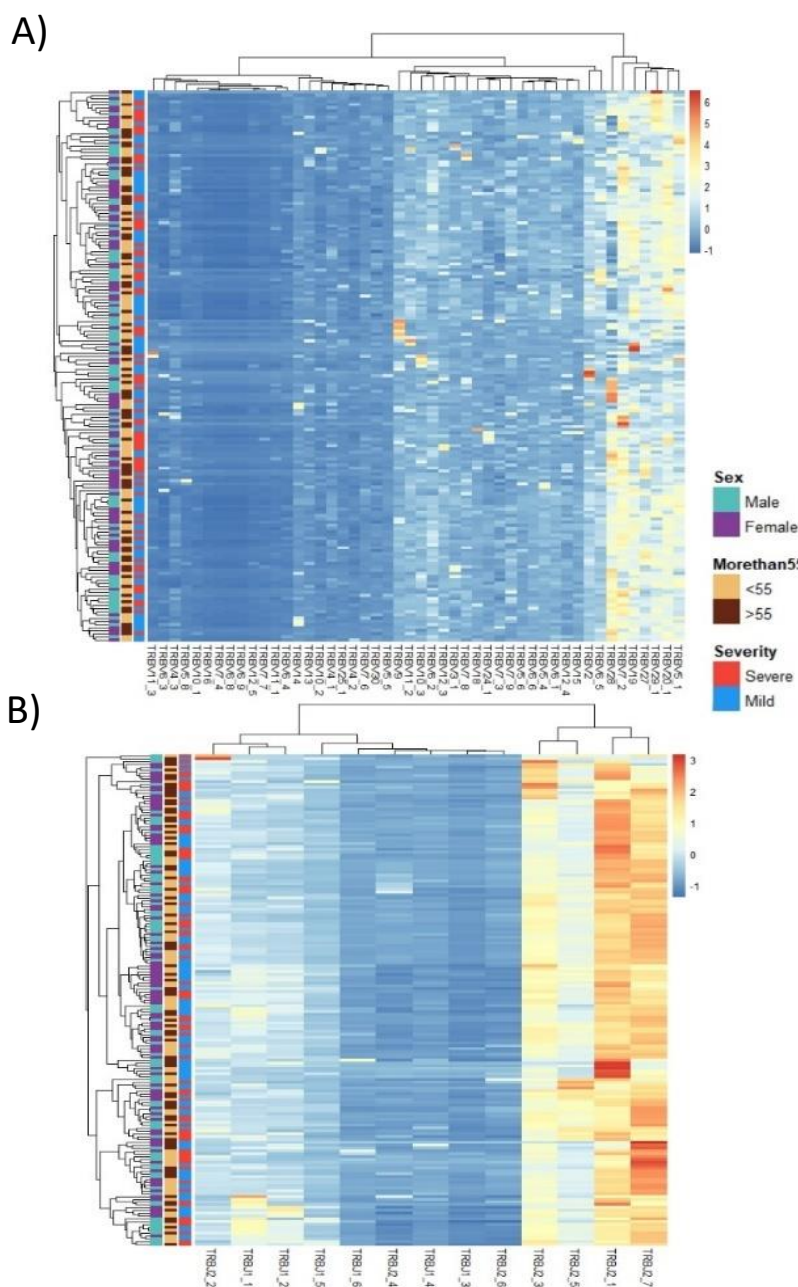


Figura 4.5. Heatmaps con agrupamiento jerárquico (distancia Euclídea) del número total de pacientes con el número total de alelos V (A) y alelos J (B) detectados. Los niveles de expresión en ambos mapas de calor se han representado mediante estandarización de puntuación z de los recuentos totales para cada alelo.

4.5 Inferencia sobre la secuencia CDR3b: agrupación de clonotipos mediante GLIPH2.

4.5.1 Distribución de los motivos estructurales GLIPH2 entre grupos de pacientes.

Aplicamos GLIPH2 a los 100 clonotipos más expandidos por repertorio, identificando 16 clusters de alineación local y 20 global ([MSO](#)). Cada clúster tiene en promedio 5-6 secuencias CDR3b, con algunos clústeres globales teniendo 11-13 secuencias ([MSO](#)). El análisis de red de estos 36 clústeres reveló algunos compuestos exclusivamente por CDR3b expandidos en pacientes leves (clúster LSPLA_4_22), severos (clúster S%GSYE_EGQST) o el subgrupo severo menor de 55 años combinado con pacientes de ≥ 55 años (clúster YSSGE_4_22) ([Figura 4.6](#)), alineándose con los resultados de diversidad, clonalidad y uso de genes. Sin embargo, la mayoría de los clústeres contenían secuencias CDR3b de los cuatro subgrupos de pacientes o subgrupos no relacionados.

4.5.2 Frecuencias de los motivos GLIPH2 con clonotipos reactivos al SARS-CoV-2.

Para mejorar la significancia funcional, filtramos las secuencias CDR3b de los 36 clústeres globales y locales, priorizando aquellas validadas experimentalmente frente a COVID-19 en el ensayo MIRA (etiquetadas como "COVID-19+"). Luego, extrajimos y sumamos las

frecuencias de estas secuencias COVID-19+ por repertorio, agrupándolas por su clúster GLIPH2. Después de esto, el número de clústeres se redujo de 36 a 24 eliminando los clústeres sin secuencias COVID-19+ CDR3b (MSO).

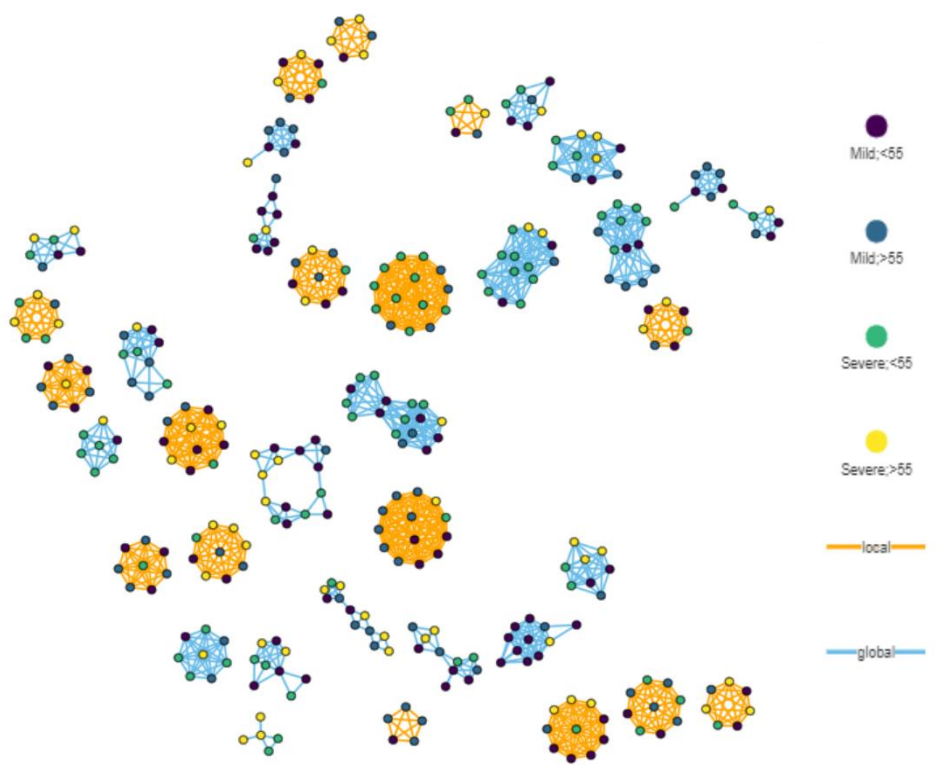


Figura 4.6 Principales resultados del análisis GLIPH2. Red con los 36 clusters globales y locales identificados inicialmente después de seleccionar los 100 clonotipos más expandidos de cada repertorio.

Nuestros resultados mostraron que cuatro clústeres con secuencias CDR3b "COVID-19+" (S%GYE_AEFGHLRSVWY, SS%YE_AGST, SP%YE_GHNRST, SL%SYE_DGNST) exhibieron diferencias

significativas de frecuencia entre grupos combinados de gravedad y edad, a pesar de no ser exclusivos en la red GLIPH2 ([Figura 4.7](#)). Los clústeres S%GYE_AEFGHLRSVWY y SL%SYE_DGNST fueron de particular interés ya que mostraron una tendencia en pacientes severos menores de 55 años a tener un perfil similar al de los pacientes mayores de 55, aunque no tan pronunciado como en las secciones anteriores. Esta tendencia menos pronunciada se refleja en los modelos de regresión logística, con un número variable de veces que cada clúster es seleccionado como relevante. Sin embargo, otro clúster GLIPH2 (SPT%AGE-ASV) se destaca como un motivo relevante en ambos grupos de edad y en los grupos combinados de edad-gravedad ([Tablas suplementarias 8.7 y 8.8](#)).

4.5.3 Distribución de las secuencias CDR3b reactivas a SARS-CoV-2 entre repertorios.

Las diferencias entre repertorios fueron notables al visualizar las frecuencias de secuencias "COVID-19+" CDR3b en cada grupo ([Figura suplementaria 8.8](#)). Estas diferencias permanecieron significativas después de eliminar los valores atípicos. El análisis WebLogo identificó variaciones características en posiciones intermedias en la secuencia CDR3b ([Figura 4.8](#)), influyendo en la afinidad del antígeno según el tipo de molécula o la fuerza de unión. Un examen adicional de las secuencias

CDR3b COVID-19+ reveló varios alelos V mencionados anteriormente. Notablemente, TRBV12-3 estaba presente en las cadenas CDR3b de los cuatro grupos seleccionados, con una presencia menor de TRBV5-5 y TRBV14/TRBV04-02, indicando una codificación ambigua de alelos V ([MSO](#)).

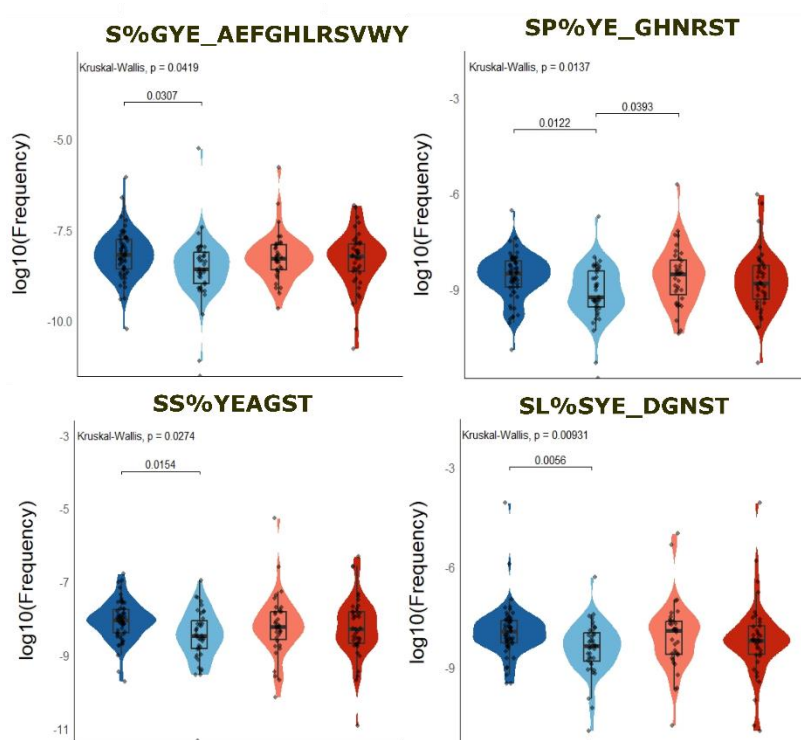


Figura 4.7 Gráfico de caja y violín de los 4 clusters identificados por GLIPH2 después de seleccionar y filtrar las secuencias CDR3b COVID-19+ con cambios significativos. Para las comparaciones de dos grupos, se aplicó el test de Wilcoxon para muestras independientes. Nivel de significancia de 0.05

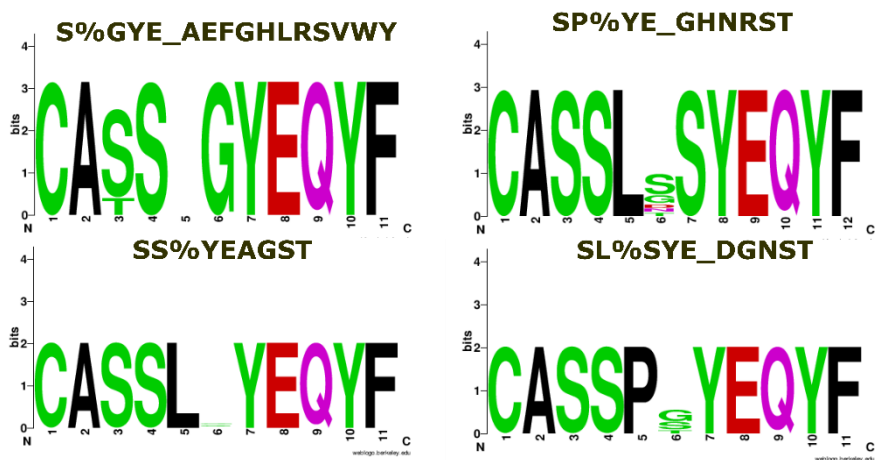


Figura 4.8. Logos de secuencias de las CDR3b COVID-19+ de cada clúster GLIPH2 mostrado en 4.7.

5. Resultados (II). La integración de datos de citometría, repertorio de TCR, genotipado y sintomatología revela nuevos subgrupos de pacientes COVID-19

5.1 Antecedentes

Se incluyó una subcohorte de 61 pacientes recuperados de COVID-19 (33 leves, 28 graves; edad media: 51 años). Se analizaron 143 variables (129 cuantitativas, 14 cualitativas) relacionadas con *TCRseq*, CyTOF, sintomatología y genotipado. Las variables cuantitativas se escalaron con puntuación z y las cualitativas se factorizaron. Se seleccionaron síntomas relevantes con <95% de datos faltantes, codificados como presencia/ausencia: fiebre, tos, disnea, astenia, mialgia, cefalea, anosmia, ageusia, afectación dermatológica, respuesta inflamatoria sistémica.

Del repertorio T se incluyeron métricas de diversidad, clonalidad, uso de alelos V/J y motivos GLIPH2 asociados a COVID-19. Los datos CyTOF consisten en los valores de porcentaje de 37 poblaciones de células del sistema inmunitario innato, con especial detalle en la línea monocito macrófago ([Tabla Suplementaria 9.1](#)). El genotipado analizó SNPs en *ACE2*, *MX1* y *TMPRSS2* mediante PCR en tiempo real. La integración se hizo con LCM-BIC para identificar variables relevantes y clústeres. Sobre los subgrupos, se usaron modelos de *deep learning*. El análisis no supervisado empleó VAE, mientras que para el análisis supervisado aplicó validación cruzada Monte Carlo.

Se seleccionaron secuencias CDR3b con $AUC \geq 0.99$, identificando motivos clave. Estas se evaluaron con TCRmatch para detectar reactividad a SARS-CoV-2. Las secuencias resultantes se alinearon con Clustal Omega y se agruparon por similitud mediante dendogramas. Los diferentes análisis quedan esquematizados en la [Figura 5.1](#).

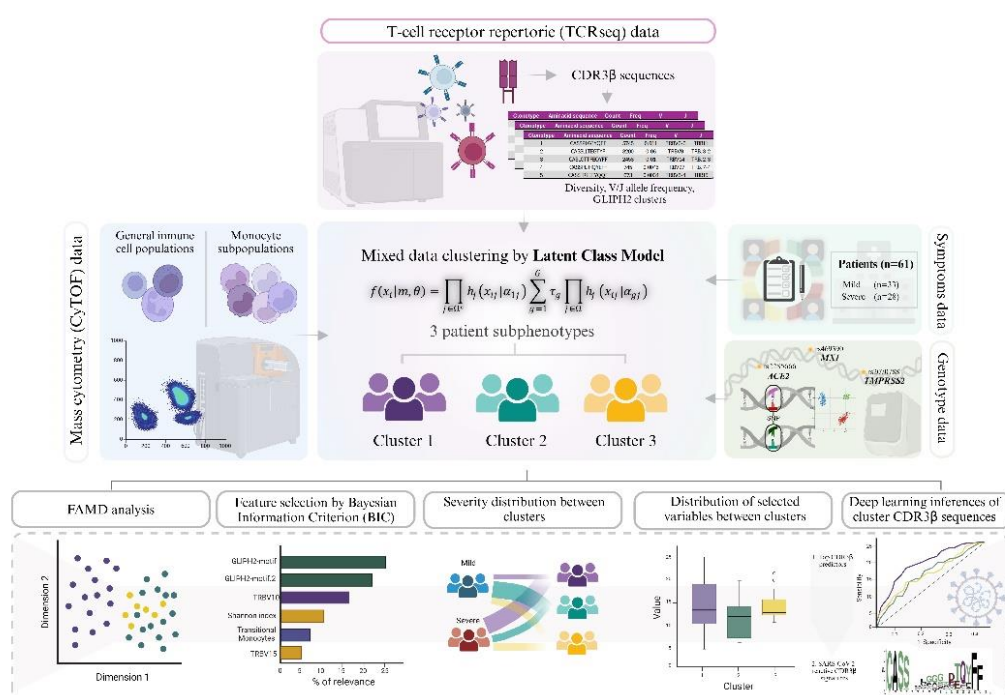


Figure 5.1 Resumen esquemático del segundo estudio de esta tesis doctoral.

5.2 Resultados del análisis FAMD en los conjuntos pre-LCM y post-LCM.

5.2.1 Conjunto pre-LCM

En el análisis FAMD del conjunto de datos pre-LCM, se evaluaron 142 variables (128 continuas y 14 categóricas). La dimensión 1 (8.1% de variabilidad) destacó variables relacionadas con la diversidad y clonalidad del repertorio TCR, con métricas clave como el índice de Shannon-Wiener, DE50, coeficiente de Gini y clonotipos hiperexpandidos siendo las más influyentes (valores $\cos^2$ 0.5-0.8) (Figura 5.2 A). Motivos GLIPH2 y algunos alelos V como TRBV25-1, TRBV5-1, TRBV19, TRBV6-8 y la presencia de neumonía también fueron relevantes. La dimensión 2 (6.5%) destacó contribuciones de poblaciones de monocitos y alelos V, con monocitos transicionales y no clásicos, monocitos CD33+Dim14, alelos V (TRBV2, TRBV6-5, TRBV28) y alelos J (TRBJ1-1, TRBJ1-4, TRBJ2-7) siendo clave. El genotipo Tmprss2-rs2070788 fue la única variable categórica en la dimensión 2. El gráfico FAMD mostró una débil separación entre casos leves y graves, particularmente en el componente 1, impulsado por la diversidad y clonalidad del repertorio TCR (Figura 5.2 B). Las pruebas PERMANOVA con 128 variables cuantitativas no revelaron diferencias significativas entre casos leves y graves ($p=0.893$) ni

con la gravedad combinada con el grupo de edad ($p=0.287$) (Figura Suplementaria 9.2).

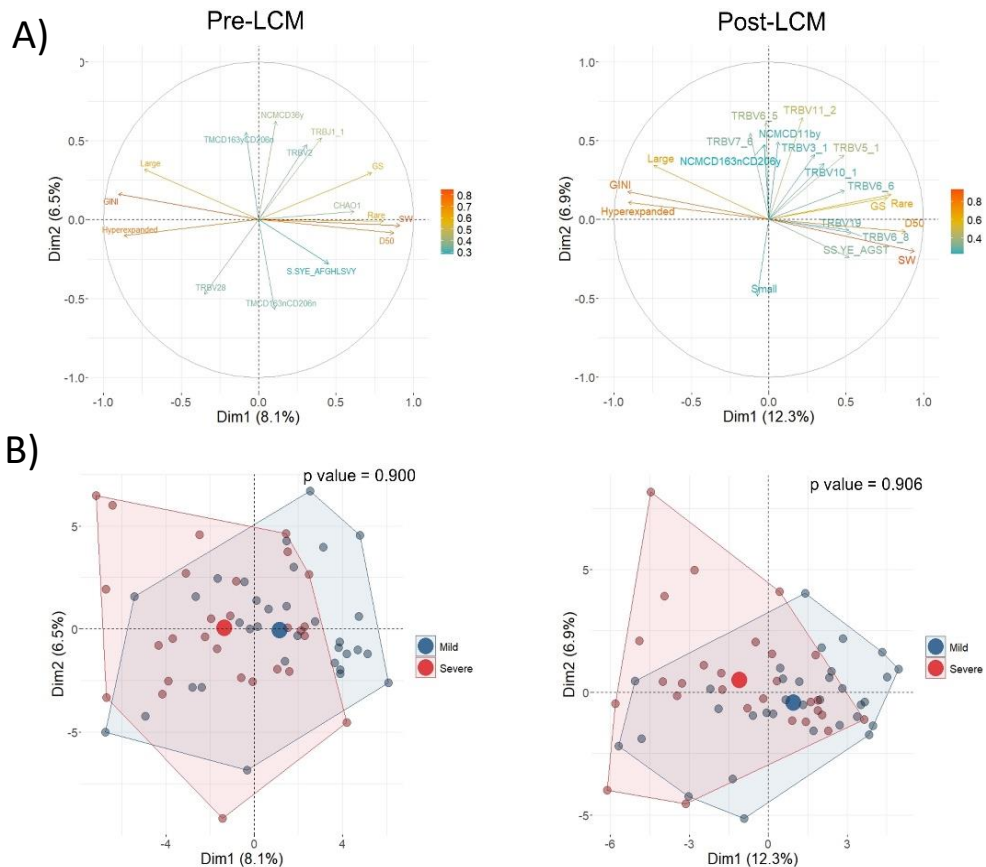


Figura 5.2. Resultados de FAMD para los conjuntos de datos pre y post-LCM. (A) Biplots de los FMAD para los conjuntos pre-LCM y post LCM. Se muestran las 20 variables más importantes según el valor de $\cos^2$. **(B)** Gráficos de las dos principales dimensiones obtenidas por FAMD. Los valores p representan los resultados de la prueba PERMANOVA con 1000 permutaciones. Nivel de significancia de 0.05.

Resultados similares se observaron para los genotipos *ACE2-rs2285666*, *MX1-rs469390* y *TMPRSS2-rs2070788*, con valores p de 0.128, 0.445 y 0.892, respectivamente ([Figura Suplementaria 9.3](#)). Para las variables cualitativas, la neumonía y el genotipo *TMPRSS2-rs2070788* mostraron resultados limitados con valores $\cos^2$ bajos ([Figura Suplementaria 9.4](#)).

5.2.2 Conjunto post-LCM

La agrupación no supervisada por LCM, con selección de variables mediante BIC, dividió las 61 muestras en 3 clústeres (Clúster 1: 28, Clúster 2: 20, Clúster 3: 13), clasificando 70 de las 142 variables como relevantes, siendo el desarrollo de neumonía la única variable cualitativa. El conjunto de datos de las 70 variables seleccionadas por LCM-BIC puede consultarse en la página web de la publicación ([LCM-BIC immunophenotyping clustering in COVID-19 patients Interactive Results and Databases](#)). La correspondiente tabla se presenta en formato interactivo, con la posibilidad de ordenar y filtrar valores de interés, así como de descargar la tabla resultante en .csv, .xlsx o .txt. En el análisis FAMD con el conjunto de datos post-LCM, las contribuciones a las dimensiones 1 (12.3%) y 2 (6.9%) aumentaron en comparación con los datos pre-LCM ([Figura 5.2](#), [Figura Suplementaria 9.1](#)). Las métricas de diversidad y clonalidad, especialmente el índice de Shannon-Wiener, DE50 y el coeficiente de

Gini, siguieron siendo significativas en la dimensión 1, con los alelos V siendo altamente influyentes.

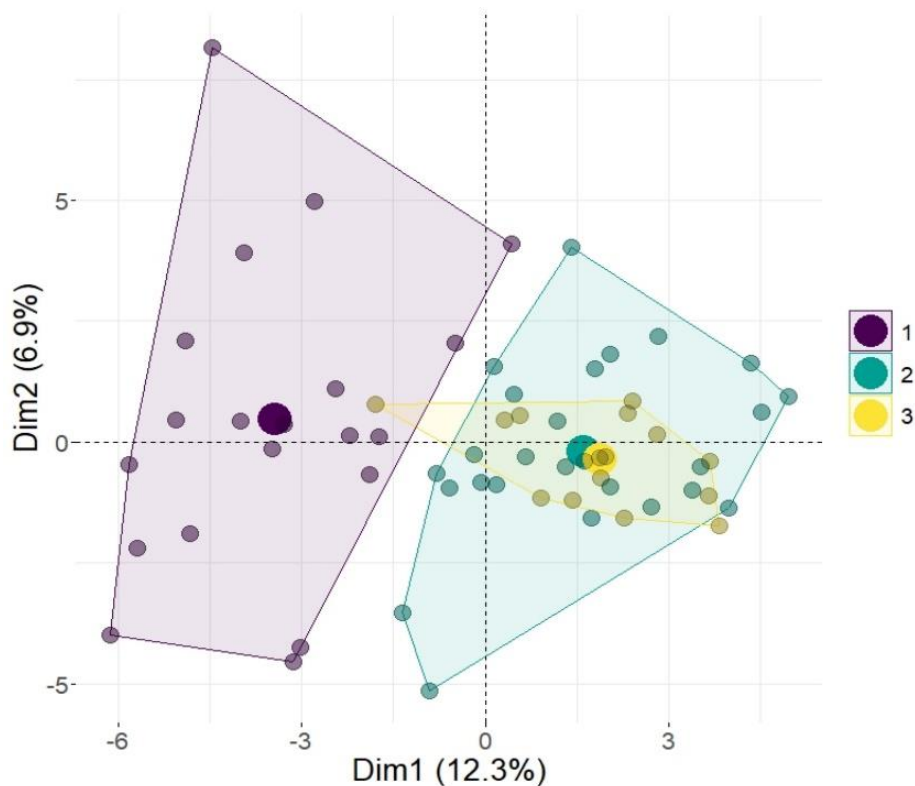


Figura 5.3. Resultados de FAMD para los conjuntos de datos post-LCM y agrupados. Las observaciones quedan divididas en función del correspondiente clúster LCM al que pertenezcan. Los valores p representan los resultados de la prueba PERMANOVA con 1000 permutaciones sobre las variables cuantitativas, con un nivel de significancia de 0.05.

En la dimensión 2, los alelos TRBV11-2, TRBV6-5 y TRBV7-6 hicieron contribuciones notables, junto con cambios en las poblaciones de monocitos, disminuyendo los monocitos transicionales y aumentando los

monocitos no clásicos ([Figura 5.2 A](#), [Figura Suplementaria 9.1](#)). Las métricas de diversidad y clonalidad, como la proporción de clonotipos grandes y pequeños, también fueron relevantes. La distribución post-LCM por gravedad en el gráfico FAMD mostró cierta separación entre pacientes graves a la derecha y pacientes leves a la izquierda, pero con alto grado de solapamiento ([Figura 5.2 B](#)). Las pruebas PERMANOVA no revelaron cambios significativos según la gravedad y la gravedad combinada con el grupo de edad ($p=0.910$, $p=0.312$) ([Figura Suplementaria 9.2](#)), pero se observaron diferencias significativas entre los tres grupos de clústeres ($p=0.001$). El gráfico FAMD muestra una clara división entre el Clúster 1 y los Clústeres 2 y 3 ([Figura 5.3](#)).

5.3 Distribución de las variables seleccionadas por LCM-BIC entre grupos de pacientes.

5.3.1 Descripción general

Siguiendo los resultados de FAMD y PERMANOVA, nos centramos en analizar la distribución de las 70 variables seleccionadas por el criterio BIC en nuestro modelo LCM. La selección afectó negativamente a las variables de síntomas y genotipos, dejando a la neumonía como la única variable cualitativa ([Figura 5.4 A](#)). El número de poblaciones de leucocitos obtenidas por CyTOF también se redujo. Sin

embargo, los motivos GLIPH2 y los alelos V ganaron relevancia en el nuevo conjunto de datos post-LCM, representando 11 de las 25 variables más relevantes ([Figura 5.4 B](#)). Motivos notables incluyen SYGGE_4_22, S%GGE-FIY, SY%GE-GST, y SL%SYE-DGNST, y alelos V como TRBV11-3, TRBV6-3, TRBV9, TRBV12-3, TRBV16, TRBV18 y TRBV15. CM-CD36+ fue la única variable CyTOF presente en el top 25. Al estudiar la distribución de pacientes por gravedad, se observó una distribución significativa ($p=0.009$), con una predominancia de pacientes graves en el Clúster 1 y pacientes leves en el Clúster 2, mientras que el Clúster 3 mostró una distribución más equilibrada ([Figura 5.4 C](#)). En el reducido grupo de pacientes leves en el Clúster 1, la mayoría eran mayores de 55 años, mientras que en el Clúster 2 eran menores de 55 años.

5.3.2 Distribución de la diversidad y la clonalidad

Los gráficos de radar mostraron una distribución sesgada entre los clústeres y los valores promedio de las principales seleccionadas por LCM-BIC ([Figura 5.5 A](#)). En cuanto a la distribución individual de las 70 variables seleccionadas entre los tres clústeres, las métricas de diversidad y clonalidad mostraron que los repertorios de células T eran significativamente menos diversos y más oligoclonales en el Clúster 1 en comparación con los Clústeres 2 y 3, que tenían valores similares ([Figura 5.5 B](#)), ver [MSO](#)). Además, todas las métricas de diversidad y clonalidad,

así como la proporción de clonotipos raros e hiperexpandidos, exhibieron un gran tamaño del efecto (MSO), consistente con las observaciones en el FAMD.

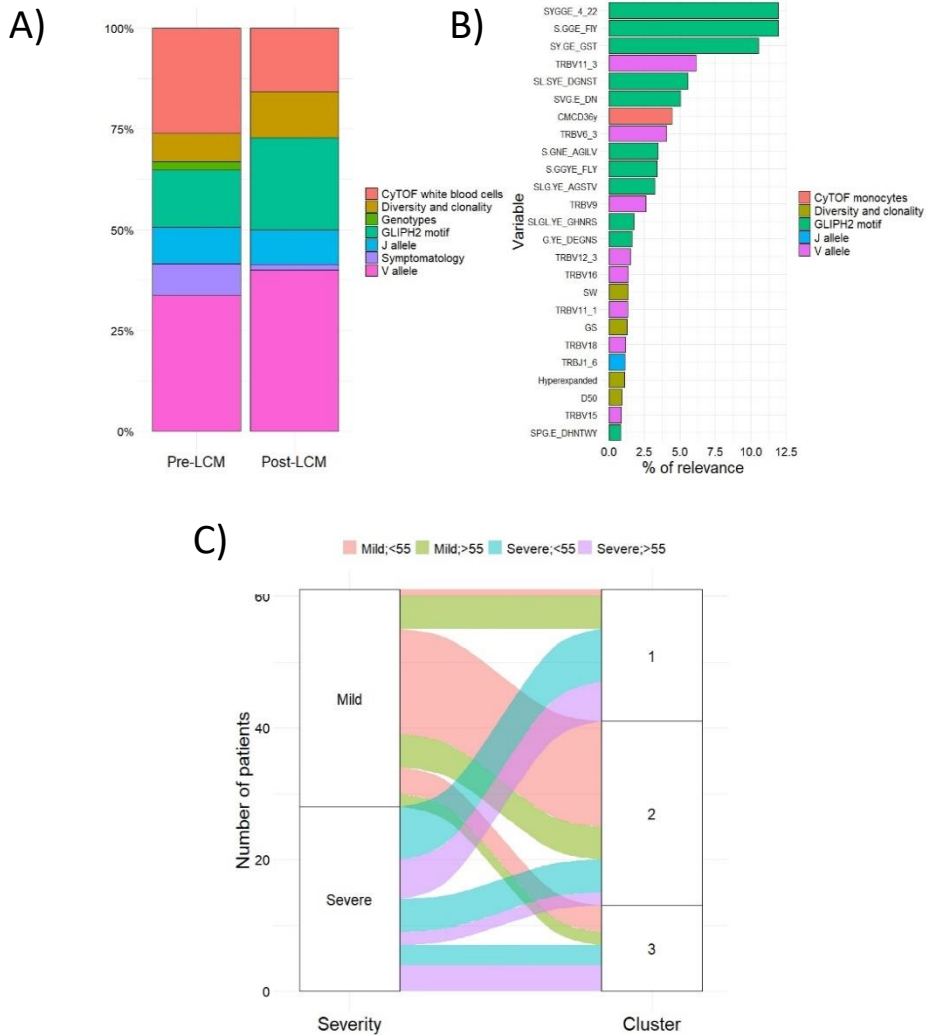


Figura 5.4. Resultados de FAMD para los conjuntos de datos post-LCM y agrupados. (A). Distribución de las variables agrupadas por tipo de dato (*TCRseq*, CyTOF, genotipado y sintomatología) en los conjuntos pre-LCM y post-LCM. **(B).** 20 primeras variables ordenadas por porcentaje de relevancia según LCM-BIC. **(C).** Diagrama de flujo entre las muestras leves y graves respecto a los clústeres generados por LCM, teniendo en cuenta además el grupo combinado de severidad y edad.

5.3.3 Distribución de los alelos V

En cuanto a los alelos V, encontramos un total de 15 con cambios significativos en la frecuencia ([MSO](#)). TRBV6-3, TRBV12-3 y TRBV15 fueron particularmente relevantes ([Figura 5.5 B](#)), mientras que TRBV19, TRBV6-6, TRBV5-1 y TRBV5-5 también mostraron un notable tamaño del efecto. Aunque el Clúster 1 mostró una frecuencia significativamente menor en comparación con los Clústeres 2 y 3 para algunos alelos V, hubo casos en los que el Clúster 3 tenía valores intermedios entre 1 y 2, como TRBV15, TRBV7-4 o TRBV6-8. En cuanto a los motivos GLIPH2, se encontró una representación notable de variables con cambios significativos ([MSO](#)) y [Figura 5.5 B](#)).

5.3.4 Distribución de los motivos estructurales GLIPH2

Los clústeres S%GGE_FIY y SYGGE-4-22 fueron las dos variables más relevantes. A pesar de tener secuencias CDR3b diferenciadas inicialmente, ambos clústeres presentaron los mismos CDR3b específicos de COVID-19 después del filtrado, reflejándose en resultados similares en gráficos de cajas y estadísticas. Entre los motivos GLIPH2 significativos con un mayor tamaño del efecto, encontramos SS%YE-AGST, S%GNE-AGLIV y S%GGYE-FLY.

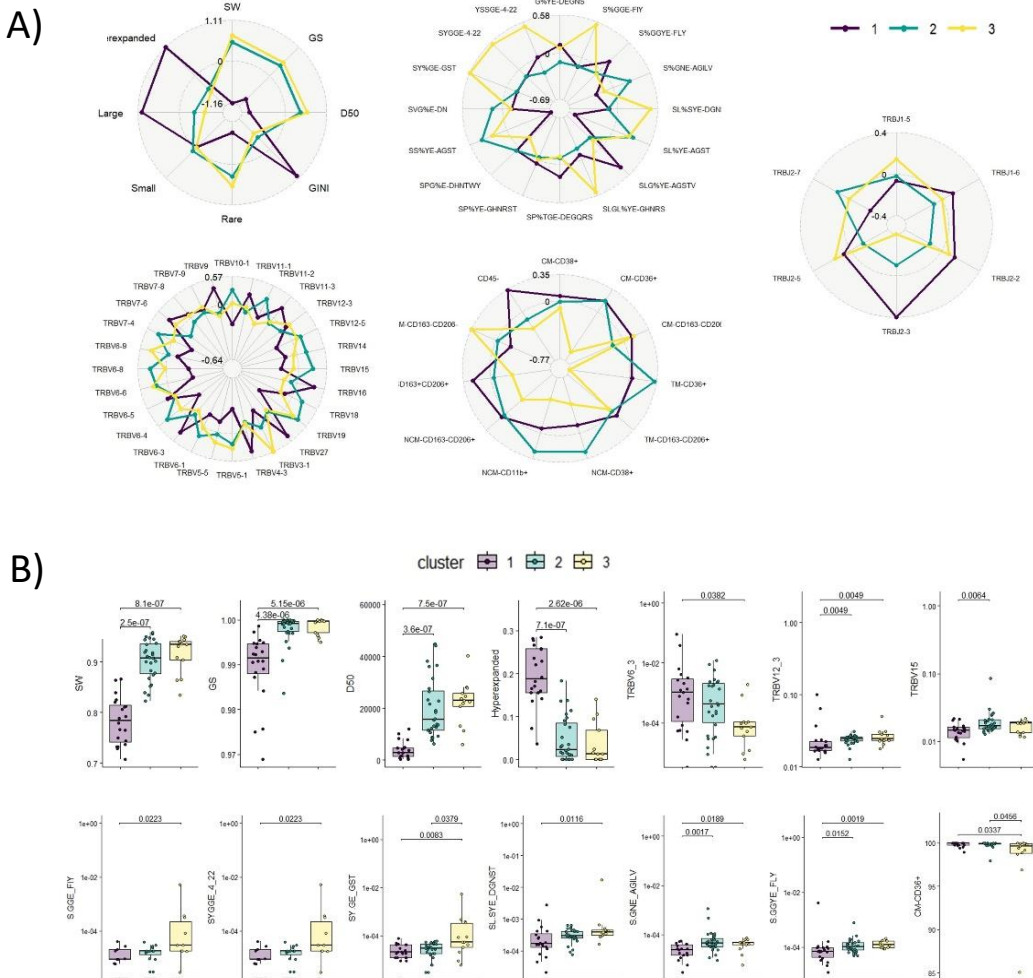


Figura 5.5. Principales resultados de la tipología y distribución de variables seleccionadas por el modelo LCM-BIC. (A). Gráficos de radar para las 70 variables seleccionadas en el modelo LCM-BIC. Cada gráfico de radar representa un conjunto de variables definidas en los grupos de las **Figuras 5.4 A y B**. Las variables de diversidad y clonalidad fueron estandarizadas previamente a z-score para una representación correcta. **(B)** Diagramas de caja que muestran las 25 variables principales con cambios significativos entre los tres clústeres tras realizar una prueba de Kruskal-Wallis de una vía. La prueba de Dunn se utilizó para post-hoc y los valores p se ajustaron usando el método de Benjamini-Hochberg para comparaciones múltiples. El nivel de significancia es 0.05.

5.3.5 Distribución de las subpoblaciones celulares por CyTOF

En cuanto a las variables CyTOF, encontramos 4 poblaciones de monocitos con cambios significativos: CM-CD36+, TM-CD36+, NCM-CD163-CD206+ y NCM-CD11b+ ([MSQ](#)). A diferencia de las variables anteriores, no exhibieron grandes tamaños del efecto, y los cambios reflejaron una presencia significativamente mayor de estas poblaciones en el Clúster 2 en comparación con los Clústeres 1 y 3, que tenían valores similares. Una excepción fue la población CM-CD36+, que aumentó a niveles similares al Clúster 2 y fue la única variable CyTOF en el top 25 ([Figura 5.4 B](#) y [5.5 B](#)).

5.4 Detección de firmas en las secuencias CDR3b entre clústeres LCM.

Para investigar más a fondo las diferencias entre los clústeres, analizamos el repertorio de células T usando la secuencia de aminoácidos de CDR3b y sus correspondientes TRBV y TRBJ. Seleccionamos los 1000 clonotipos más abundantes de cada repertorio y aplicamos *deep learning* no supervisado y supervisado utilizando las plataformas DeepTCR.

5.4.1 *Deep learning* no supervisado.

Los resultados del análisis no supervisado por VAE indicaron que el 99% de la varianza se explicó mediante las primeras 15 divisiones latentes. Al graficar las observaciones contra el P valor promedio de cada una de las 15 dimensiones latentes utilizando un mapa de calor ([Figura suplementaria 9.5 B](#)), pudimos diferenciar la mayoría de las muestras del Clúster 1 de los Clústeres 2 y 3. La evaluación del poder predictivo del modelo VAE usando el método K-nearest neighbour confirmó su capacidad predictiva sobre el Clúster 1, con un AUC promedio de 0.7 ([Figura Suplementaria 9.5 A](#)).

5.4.2 *Deep learning* supervisado

Los resultados del análisis supervisado mediante *deep learning* con DeepTCR generaron modelos capaces de distinguir moderadamente las muestras del Clúster 1 de los Clústeres 2 y 3, con un AUC de 0.645 y 0.796, respectivamente ([Figura 5.6](#)). Sin embargo, el modelo generado para la comparación entre los Clústeres 2 y 3 no logró detectar características con suficiente poder predictivo (AUC=0.442). Estos resultados se alinean con los obtenidos en el análisis no supervisado de DeepTCR utilizando VAE. Además, el modelo supervisado creado en

DeepTCR para la comparación entre casos leves y graves también carecía de poder predictivo ($AUC = 0.444$, [Figura Suplementaria 9.6](#)), subrayando la relevancia de la agrupación por LCM para detectar características previamente indistinguibles mientras se mantiene una división significativa por gravedad.

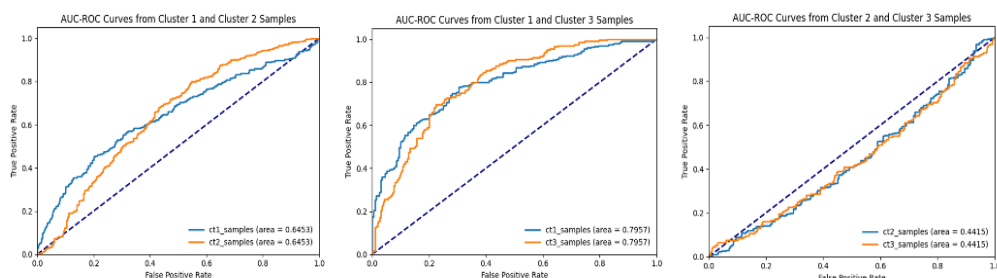


Figura 5.6 Principales resultados de los análisis de aprendizaje profundo supervisado utilizando DeepTCR. Curvas AUC-ROC de los modelos clúster 1 vs clúster 2 (izquierda), clúster 1 vs clúster 3 (centro) y clúster 2 vs clúster 3 (derecha). Los análisis fueron validados utilizando 100 simulaciones de Monte Carlo y una división de 75% de entrenamiento y 25% de prueba.

5.4.3 Residuos de aminoácidos críticos en los CDR3b

con mayor poder predictivo

Por otro lado, los RSL de las 25 secuencias de CDR3b con mayor poder predictivo ($AUC \geq 0.99$) para las comparaciones Clúster 1 vs Clúster 2 y Clúster 1 vs Clúster 3 revelaron la importancia de varios aminoácidos ubicados en las posiciones centrales debido a su alta sensibilidad ([Figura 5.7 A-B](#)). Es notable que en las 25 secuencias de CDR3b pertenecientes al Clúster 1, la relevancia se concentra principalmente en las posiciones 5, 6

y 7, con una mayor presencia de residuos de ácido glutámico (E), ácido aspártico (D) y prolina (P). Estas observaciones facilitaron la distinción de firmas significativas entre los Clústeres 1 ($p=0.005$) y 2 ($p=0.011$ y $p=0.007$) (Figura 5.7 C). Sin embargo, en la comparación entre el Clúster 1 y el Clúster 3, no se encontraron firmas representativas significativas.

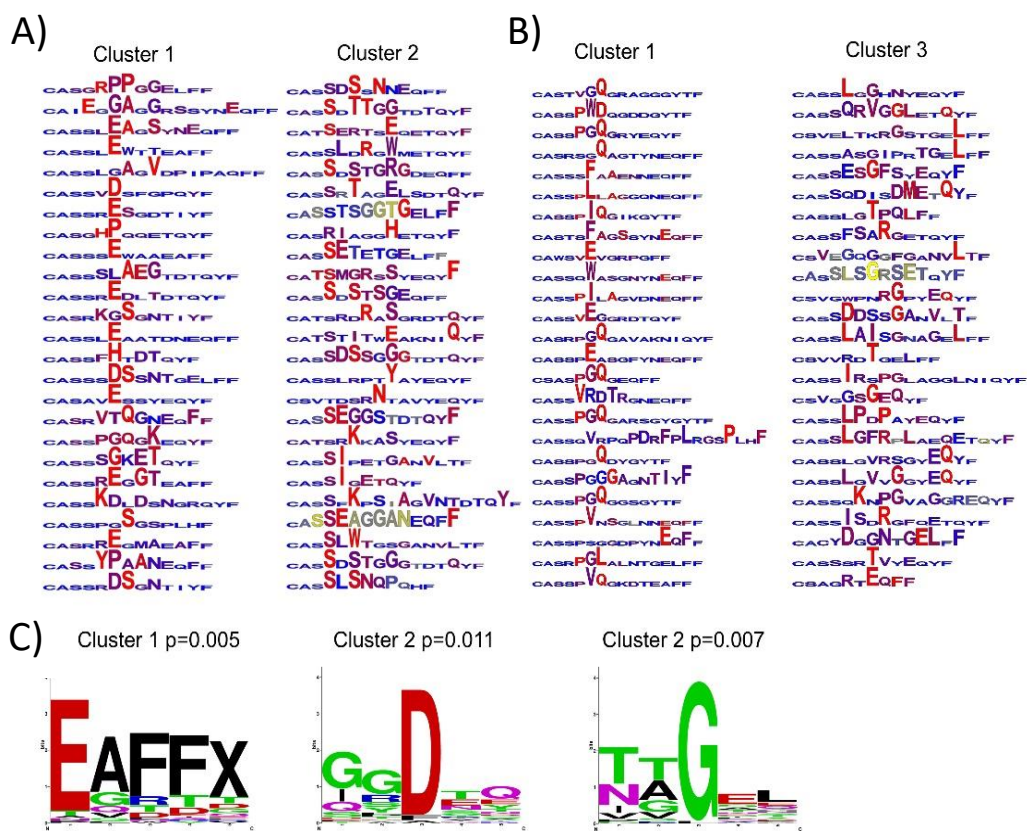


Figura 5.7. Principales resultados del estudio de residuos críticos en los CDR3b más predictivos. (A). Gráficos RSL que muestran los 25 CDR3b con mejor predicción para cada clúster en el modelo clúster 1 vs clúster 2 y (B) en el modelo clúster 1 vs clúster 3. (C) Logotipos de los motivos estructurales de los 50 CDR3b más representativos para cada clúster en el modelo clúster 1 vs clúster 2. P valor obtenido de la prueba de Mann-Whitney. Nivel de significancia de 0.05.

5.5 Distribución de las secuencias en CDR3bs predictivas reactivas al SARS-CoV-2.

Las secuencias más predictivas de las comparaciones entre el Clúster 1 vs. Clúster 2 y el Clúster 1 vs. Clúster 3 seleccionadas por DeepTCR ($AUC \geq 0.99$) fueron elegidas para estudiar su grado de compartición y su reactividad al SARS-CoV-2. Para identificar las secuencias CDR3b afines al SARS-CoV-2, se utilizó el programa TCRmatch con parámetros predeterminados. Los resultados muestran que entre el 23.52% (clúster 1) y el 28.91% (clúster 3) de las secuencias predictivas de CDR3b son reactivas al SARS-CoV-2 ([Figura Suplementaria 9.7](#)). Información adicional sobre las secuencias CDR3b analizadas y su afinidad a diferentes antígenos puede encontrarse en la web asociada: [LCM-BIC immunophenotyping clustering in COVID-19 patients Interactive Results and Databases](#). Los detalles de las bases de datos creadas en este recurso online se describen más adelante.

La distribución de secuencias por tipo de antígeno no difirió significativamente, excepto por algunos antígenos con baja cobertura ([Figura Suplementaria 9.8](#)). El número de secuencias CDR3b compartidas entre los clústeres indica una compartición casi nula entre el clúster 1 y los

clústeres 2 y 3, ya que no hay secuencias con alta fuerza predictiva compartidas entre ellos ([Figura 5.8 A](#)).

Para las secuencias afines al SARS-CoV-2, la superposición fue de 59 secuencias para el clúster 1 en cada comparación, y 115 para los clústeres 2 y 3. Las 59 secuencias en el clúster 1 con alta capacidad para discriminar los clústeres 2 y 3 provocaron una investigación adicional. El MSA de estas 59 secuencias CDR3b reactivas al SARS-CoV-2 reveló residuos centrales-distales conservados ([Figura 5.8 B](#) y [Figura Suplementaria 9.9 A](#)).

El MSA identificó dos grandes subgrupos según un dendrograma de la matriz de sustitución BLOSUM62 ([Figura 5.8 C](#)). Las diferencias entre los grupos se encuentran principalmente en posiciones centrales-distales ([Figura 5.8 D](#)). El subgrupo uno tiene la firma E/DTQYF y residuos de guanina en la posición 6 ([Figura Suplementaria 9.9 A](#)), mientras que el subgrupo dos es más numeroso y variable en posiciones centrales. Los RSLs para el total de 59 secuencias CDR3b muestran una mayor variabilidad en los residuos críticos en comparación con los 25 principales logotipos de sensibilidad de residuos, aunque la posición 6 sigue siendo importante ([Figura Suplementaria 9.9B-C](#)).

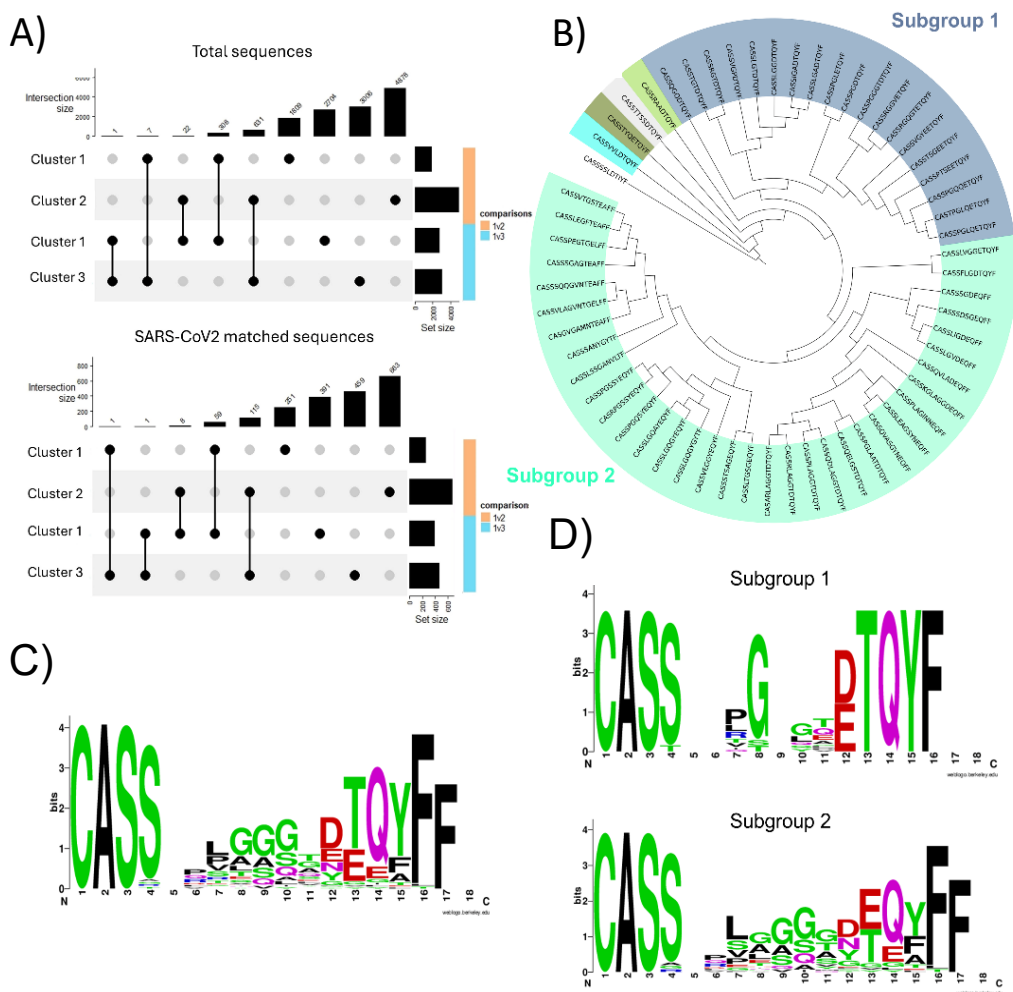


Figura 5.8. Secuencias del clúster 1 reactivas a SARS-CoV-2 (A). Gráficos UpSet con la intersección de los clústeres de las dos comparaciones de interés (1 vs 2 y 2 vs 3). El panel superior indica el número de secuencias totales de CDR3b intersectadas, mientras que el panel inferior muestra el número de CDR3b específicas de SARS-CoV-2 después del análisis de TCRmatch. (B) Dendrograma generado por MSA de las 59 secuencias comunes de CDR3b reactivas a SARS-CoV-2 en el clúster 1 en ambas comparaciones. Se utilizó el software Jalview2.11.3.2 con la matriz de sustitución BLOSUM62 para el diagrama del árbol. Los dos subgrupos principales identificados en el árbol se muestran en color junto con las secuencias de CDR3b que quedan fuera de ellos (C). Logotipo de secuencia generado con el total de las 59 secuencias de CDR3b seleccionadas del clúster 1. (D) Logotipo de secuencia de las secuencias de CDR3b pertenecientes a cada subgrupo obtenido del árbol en (B).

5.7 Creación de una base de datos interactiva

El segundo estudio enmarcado en esta tesis doctoral ha supuesto la obtención de resultados de interés como es el de un subconjunto de 70 variables de TCRseq, CyTOF y sintomatología (únicamente con la presencia/ausencia de neumonía) y de miles de secuencias CDR3b con alta capacidad predictiva entre clústeres y sus correspondientes antígenos afines. Con el objetivo de que el trabajo realizado gane mayor relevancia y utilidad para el conjunto de científicos y clínicos, se ha creado un portal web en el cual se presentan dos bases de datos interactivas de uso abierto ([LCM-BIC immunophenotyping clustering in COVID-19 patients Interactive Results and Databases.](#)).

La primera base de datos consiste en el conjunto de las 70 variables LCM-BIC, con su correspondiente valor por cada muestra analizada (filas). Los valores no se encuentran transformados por z-score para favorecer así una mejor interpretabilidad y reutilización de los datos por parte de nuevos usuarios interesados. Además de las 70 variables, también se ha incluido el correspondiente clúster al que pertenece la muestra, así como la clasificación inicial de gravedad y severidad.

La segunda base de datos consiste en el total de 7631 secuencias CDR3b, clasificadas como altamente predictivas tras los análisis

supervisados por DeepTCR, teniendo en cuenta un AUC por secuencia igual o superior al 0.99. La tabla resultante resume información clave para cada secuencia analizada, incluyendo: la secuencia CDR3 de aminoácidos con una coincidencia validada en IEDB, la versión recortada de la CDR3 β (sin el primer y último residuo) usada por compatibilidad con TCRmatch, la puntuación de similitud otorgada por TCRmatch, la secuencia del epítipo coincidente que sugiere especificidad antigénica, el nombre completo y abreviado de la proteína de origen del antígeno, el organismo patológico del cual proviene, el clúster de pacientes al que pertenece la secuencia CDR3b, y el modelo supervisado de DeepTCR (Clúster 1 vs 2 o Clúster 1 vs 3) donde se identificó originalmente. El epítipo, antígeno y patógeno de origen no se limita a la COVID-19, sino que también incluye otros organismos como el virus del dengue, hepatitis B y C, virus de Epstein Barr, influenza A (gripe), citomegalovirus e incluso autoantígenos humanos.

Las bases de datos permiten filtrar por variables (por búsqueda, orden o intervalos) y realizar búsquedas globales. Los resultados pueden descargarse en formatos como Excel, CSV, PDF o copiarse al portapapeles. Para facilitar la navegación, se ha fijado la columna con los identificadores clave (ID de muestra y secuencia CDR3b). Además, cada base incluye un resumen de resultados con gráficos interactivos, un

graphical abstract y enlaces a los repositorios de datos y metodología. La interfaz se ha diseñado para ser clara y fácil de usar. Todo ello aporta un valor añadido a la tesis, al ofrecer una herramienta útil para el estudio del repertorio de linfocitos T en COVID-19 y otras enfermedades.

6. Discusión

La COVID-19 presenta una gran variabilidad clínica, influida en parte por la respuesta de los linfocitos T. Esta tesis aplicó técnicas de inmunoinformática para caracterizar el repertorio TCR en pacientes españoles, integrando datos de *TCRseq*, CyTOF, genotipado y síntomas. Se identificaron subgrupos inmunológicos con perfiles distintos. A continuación, se discuten los principales hallazgos, la integración de datos con LCM, las limitaciones del estudio y perspectivas futuras.

6.1 Análisis del repertorio de linfocitos T

En este estudio, se detectó en pacientes graves una disminución en la diversidad pese a no ser significativa, así como un aumento en la clonalidad, en comparación con los pacientes leves. Estos cambios han sido también reportados en otros estudios [177,178]. Esta visible reducción en la diversidad y aumento en la clonalidad es atribuible principalmente a la característica linfopenia en pacientes graves, ya mencionada durante la introducción [178–181]. Por otro lado, la ausencia de cambios significativos en la diversidad y clonalidad puede atribuirse a la transición de los pacientes leves de una fase de alta a una de convalecencia [103]. Las diferencias en diversidad y clonalidad son más pronunciadas cuando los pacientes se dividen en grupos de edad <55 y ≥ 55 , observándose diferencias significativas. Estudios previos han demostrado que la diversidad de los repertorios TCR de pacientes con

COVID-19 disminuye con la edad [52,58,182]. Sin embargo, nuestro estudio ha revelado que los pacientes graves menores de 55 años tienen cambios significativos en diversidad y clonalidad, con valores muy similares a los de los pacientes mayores de 55 años, independientemente de la gravedad. Dada la producción constante de linfocitos T naïve en pacientes más jóvenes, una explicación plausible para este fenómeno podría ser la disfunción del interferón 1 ($INF\ 1$) en la gravedad de la enfermedad [50].

En cuanto al solapamiento de repertorios, nuestros hallazgos indican una disparidad significativa en la similitud de los repertorios entre pacientes leves y graves, particularmente con clones hiperexpandidos. El nivel de superposición de repertorios es bajo y disminuye a medida que aumenta la expansión clonal. Este resultado sigue los hallazgos de Chang *et al.*, donde se detectó un mínimo solapamiento de clonotipos agrupados por GLIPH2 entre pacientes leves y graves [177]. Una posible explicación para el limitado solapamiento de repertorios sería la propiedad de que TCRs diferentes pueden tener un repertorio de antígenos afines similares, llevando a la expansión de TCRs generalmente similares pero variables en ciertas regiones, aunque con una afinidad igual para un determinado antígeno [14,183].

En el análisis del repertorio de linfocitos T, identificamos un uso predominante de los alelos V TRBV7-2, TRBV20-1 y TRBV27 y de los alelos J como TRBJ2-1, TRBJ2-3 y TRBJ2-7 en casi todos los pacientes con COVID-19. Otros estudios del repertorio de linfocitos T señalan un sesgo en el uso de alelos V y J mencionados en pacientes COVID-19 respecto a controles sanos [179–181,183]. En nuestro estudio, los análisis muestran una disminución significativa en la frecuencia de TRBV19, TRBV15, TRBV6-4 y TRBV14 entre pacientes graves en comparación con casos leves. Estos hallazgos sugieren una posible asociación entre la frecuencia de estos alelos y la gravedad de la enfermedad. Como el uso diferencial de genes V/J entre diferentes grupos de gravedad es limitado [177], la presencia de alelos con cambios significativos en su frecuencia es relevante para la diferenciación de casos leves y graves.

Centrándonos en estos alelos V, los pacientes graves menores de 55 años tienen curiosamente una frecuencia de TRBV15, TRBV14 y TRBV6-4 similar a la de los pacientes mayores o con 55 años, siguiendo la misma línea que los resultados de diversidad y clonalidad. Varios estudios han reportado cambios notables en los alelos mencionados con respecto a la gravedad de COVID-19 [177,184]. Tanto nuestro estudio como los datos de Chen *et al.* encontraron frecuencias más bajas de los alelos TRBV19, TRBV15 y TRBV6-4 [184]. La excepción fue que Chen.

et al. encontraron un aumento pronunciado en la frecuencia del alelo TRBV14 entre los pacientes graves. Este evento se debe principalmente a la afinidad del TCR por las regiones superantigénicas del SARS-CoV-2 [184]. Por otro lado, Li. *et al.* reportaron una disminución de TRBV15 y otros alelos V en pacientes convalecientes [177]. Basándonos en sus hallazgos, sugerimos que estas discrepancias pueden deberse a cambios significativos en las frecuencias de TRBV14 durante las fases posteriores de la infección, lo que puede tener un mayor impacto. La menor presencia del alelo TRBV6-4 en pacientes graves en comparación con los leves está vinculada a una menor presencia de células Tc17/MAIT; tanto en la fase aguda de la enfermedad como en las fases post-infección, ya que estas células tienen un uso preferencial de TRBV6-4 [53].

Finalmente, el análisis por GLIPH2 sobre las secuencias más expandidas ha identificado clústeres con motivos distintos en las regiones centrales de CDR3, lo cual es consistente con hallazgos previos [179,185–187]. Observamos que ciertos clústeres de secuencias MIRA-positivas de CDR3b tienen una distribución de frecuencia en pacientes graves menores de 55 años que sigue la misma línea que los resultados de diversidad, clonalidad y uso de genes. Sin embargo, se debe tener precaución a la hora de interpretar los resultados, ya que la presencia de motivos de secuencia podría estar influenciada por características específicas de la cohorte. Este

fenómeno se detectó mediante análisis de *deep learning* con VAE [188] para encontrar firmas de CDR3b vinculadas a la gravedad del COVID-19 entre dos grandes conjuntos de datos de repertorios TCR como son las provenientes del Instituto Nacional de Salud de Estados Unidos (NIH/NIAID) y del (Instituto de Biología de Sistemas) ISB [189].

6.2 Integración de datos por medio de LCM

Estudios previos han demostrado la capacidad de las aproximaciones basadas en clases latentes para detectar nuevos subfenotipos de COVID-19 a través de la integración de datos mixtos. Por ejemplo, se ha aplicado Análisis de Clases Latentes (LCA) para identificar subgrupos radiológicos basados en imágenes de tomografía computarizada [190], también para estudiar la distribución de síntomas cardiovasculares durante el COVID-19 prolongado [191,192] y la identificación de subgrupos con diferentes pronósticos [34]. Al probar diferentes modelos de agrupamiento de datos mixtos en datos simulados y reales, LCM se distingue por su robustez frente a perturbaciones de datos y su superior capacidad para detectar subgrupos de pacientes distintos, superando a otras técnicas como K-prototypes, Kamila y Partitioning Around Medoids [169]. Además, la función de selección de variables de LCM en implementaciones de software de código abierto como R mejora la

precisión del agrupamiento y facilita un examen más detallado de las variables clínicas. A diferencia del LCA, que requiere la discretización de variables continuas, LCM integra los datos de manera más completa al acomodar tanto entradas cuantitativas como cualitativas [193].

El segundo estudio de esta tesis integró datos de CyTOF, repertorio T, genotipado y síntomas en 61 pacientes recuperados de COVID-19, aplicando agrupamiento no supervisado con LCM y selección de variables por BIC. Se identificaron tres subgrupos diferenciados por 70 variables, con distribución significativa entre casos leves y graves. Destacaron diferencias en SY%GE-GST, el alelo TRBV11-3 y monocitos CM-CD36+. Mediante *deep learning*, se identificaron secuencias CDR3b que distinguen al grupo 1 (predominantemente graves) del resto, asociado a repertorios menos diversos y más oligoclonales, en línea con lo observado en el Capítulo 4. También se detectaron alteraciones consistentes en alelos V como TRBV15, TRBV6-4 y TRBV19, y también en otros menos descritos como TRBV12-3, TRBV6-6, TRBV6-3 y TRBV5-5. Estos hallazgos refuerzan la relevancia del repertorio T en la evolución clínica y su potencial como marcador inmunológico.

En cuanto a las alteraciones de los monocitos, se ha observado que los casos graves exhiben una disminución de monocitos no clásicos y

transicionales junto con un aumento de monocitos clásicos [194]. Los monocitos que expresan marcadores CD11b+, CD163+ y CD206+ están asociados con un fenotipo de respuesta inflamatoria [195,196]. Concretamente, los casos COVID-19 graves presentan niveles elevados de CD163 plasmático (sCD163) [197], mientras que los niveles de CD11b+ bajan [195]. En nuestro estudio se han detectado cambios significativos en cuatro subtipos de monocitos (CM-CD36+, TM-CD36+, NCM-CD163-CD206+, NCM-CD11b+), que exhiben niveles reducidos en el grupo 3, aunque con un impacto moderado. Estudios previos utilizando muestras de esta cohorte no revelaron diferencias en CD163-/CD206+ entre grupos leves y graves [167], lo que sugiere la identificación de un nuevo subgrupo de pacientes con cambios distintivos en las poblaciones de monocitos.

En este estudio también figuran numerosos clústeres de CDR3b obtenidos por GLIPH2) identificados como distintivos entre pacientes leves y graves. Una de las principales características comunes de estas variables es la considerable variabilidad en las muestras dentro del mismo grupo. Algo que ha sido reportado en otros trabajos [189]. Esta alta variabilidad entre muestras también es observable en los resultados del primer estudio de esta tesis doctoral, en donde los clústeres seleccionados mostraban valores de frecuencia dispares entre muestras. Nuestros resultados sugieren que los pacientes en los grupos 2 y 3 demuestran un

mayor uso de secuencias reactivas a COVID-19, tanto en cantidad como en frecuencia, lo que indicaría una respuesta potencialmente más efectiva contra el SARS-CoV-2.

Por otro lado, los resultados del análisis por DeepTCR destacan la importancia de los aminoácidos centrales como firmas pronósticas, dada su función como la principal región de unión de antígenos, al igual que lo detectado en otros estudios sobre el CDR3 [184,185,188]. Los modelos de *deep learning* tanto supervisado como no supervisado son capaces de distinguir el grupo 1 respecto al 2 y al 3, algo que no sucede con la división inicial entre leves y graves. Además, identificamos un nuevo subconjunto de 59 secuencias CDR3b de alta predictibilidad y afines al SARS-CoV en el grupo 1, capaces de distinguir tanto los grupos 2 como 3. Estas secuencias predictivas mantienen la importancia de los residuos 5, 6 y 7. Estos resultados van en consonancia con los obtenidos por Sidhom y Baras en la aplicación de DeepTCR para distinguir firmas de TCR entre leves y graves en las cohortes NIH/NIAID e ISB [188]. Además de lograr una distinción entre leves y graves con muy buena capacidad predictiva entre individuos de la misma cohorte (AUC: 0.8-0.95), los RSL en pacientes graves muestran un patrón de uso de residuos muy similar a los obtenidos en este estudio con el clúster 1, en especial la cohorte de NIH/NIAID en cuanto a la centralidad de las posiciones críticas, aunque con diferente uso

de aminoácidos. Sin embargo, la comparación posterior entre pacientes graves de NIH/NIAID vs ISB (cohorte de pacientes de España) muestra una baja capacidad predictiva del modelo resultante, enfatizando que la buena capacidad para distinguir firmas entre pacientes leves y graves está condicionada a variaciones biológicas y técnicas en el estudio del repertorio.

6.3 Limitaciones

Los resultados y la bibliografía destacan la importancia de las técnicas bioinformáticas y su aporte al estudio de la respuesta inmunitaria en pacientes con COVID-19. No obstante, existen limitaciones que deben de ser mencionadas.

Una de las principales limitaciones viene dada por la recolección de muestras entre 2020 y 2021, que estuvo condicionada por la saturación del sistema sanitario. De hecho, la recolección de muestras durante las primeras olas de la pandemia en España puede introducir un sesgo de selección no sistemático, debido a la inclusión de pacientes asintomáticos confirmados por pruebas de serología o la inclusión y/o exclusión de pacientes leves o graves por discrepancias en los criterios de selección empleados en ese tiempo. En consecuencia, existe la posibilidad de que algunos pacientes hayan contribuido a que los grupos de leves o graves

cuenten con casos infra- o sobreestimados. Por otro lado, también existen diferencias intrínsecas a los diferentes grupos de pacientes analizados, debido a diferencias en el tiempo de recolección de las muestras respecto al tiempo de infección del virus en algunos de estos pacientes, con el consecuente cambio en la actividad de los linfocitos T de acuerdo con lo descrito en la Introducción.

Respecto al análisis del repertorio de linfocitos T, sobre la composición y dinámica de los mismos podría estar influenciado por el genotipo *HLA* [90,198,199] y los sesgos en la metodología de preparación de librerías de TCR [200]. Además, el análisis *bulkTCR β seq* no puede capturar la identidad clonal absoluta que comprende las cadenas α y β emparejada [87]. Otra limitación para considerar en este estudio es que los datos se recopilaban sin fraccionar las PBMCs (células mononucleares de sangre periférica), lo que dificulta distinguir las respuestas específicas de los linfocitos T CD8⁺ o CD4⁺. Finalmente, el filtrado de secuencias CDR3b en clústeres de GLIPH2 puede haber excluido secuencias no anotadas con especificidad desconocida para los antígenos del SARS-CoV-2.

Respecto los análisis de integración de datos y *deep learning*, la principal limitación de este estudio es el tamaño relativamente pequeño de la cohorte, en parte debido a los altos costos técnicos y económicos de los

protocolos experimentales desarrollados. Sin embargo, el descubrimiento de firmas características en CDR3 permite la creación de modelos enfocados únicamente en esta característica para clasificar muestras de repertorios de células T en los grupos 1-3. Este hallazgo, en especial cuando se han obtenido secuencias CDR3 reactivas a SARS-CoV-2, adquiere valor añadido ya que ha demostrado ser robusto ante los posibles sesgos de población mencionados anteriormente en las limitaciones.

6.4 Perspectivas de futuro

El principal estudio que se plantea en la actualidad es el de probar los modelos de *deep learning* creados en cohortes de repertorio de TCR β de mayor tamaño. Un primer paso será el de la validación con el resto de las muestras de *TCRseq* que no han sido incluidas en los análisis de integración (112 muestras), controlando de esta manera las posibles diferencias poblacionales que influyen en el repertorio del TCR. Tras ello, se emplearán cohortes mayores de diferentes poblaciones, etnias y distribuciones de edad para comprobar su robustez e identificar CDR3b reactivos a COVID-19 que puedan servir como firma global. Por otro lado, sobre estas nuevas secuencias CDR3b se hará una inferencia mayor empleando herramientas de predicción de alelos del *HLA* con afinidad de unión, añadiendo un nuevo conjunto de variables y precisando en la

estratificación. Los resultados ponen en valor el avance que suponen los estudios de esta tesis doctoral a la hora de comprender el papel de los linfocitos T en la respuesta inmunitaria frente a la COVID-19. El potencial que guarda la inmunoinformática, y en particular el análisis por *TCRseq* no solo se limita a esta enfermedad, sino a presentes y futuras patologías que afectan a la población.

7. Conclusiones

1. El análisis del repertorio de linfocitos T reveló que los pacientes graves menores de 55 poseen un repertorio poco diverso, oligoclonal y escasamente compartido, propio de personas mayores de 55 años con independencia de la severidad.
2. El estudio del uso de alelos V(D)J identificó a alelos V específicos como TRBV6-4, TRBV19, TRBV12-3, TRBV15 con cambios significativos respecto a la gravedad y/o la combinación de gravedad y edad.
3. La agrupación de clonotipos mediante la aplicación de GLIPH2 y posterior selección por modelos de regresión logística identificó a grupos de CDR3b relacionados con la severidad y la edad.
4. La integración de datos de *bulk-TCR β seq*, CyTOF, genotipado y sintomatología mediante LCM-BIC identificó 3 subgrupos de pacientes, estando el subgrupo 1 bien diferenciado respecto al 2 y 3.
5. El análisis de variables identificadas como relevantes mostró diferencias significativas entre los subgrupos. El clúster 1 está compuesto mayormente por pacientes graves y caracterizado por un repertorio menos diverso y oligoclonal, con perfiles distintivos de alelos V, clústeres de secuencias CDR3b y poblaciones monocitarias. El subgrupo 2 incluye principalmente pacientes

leves y el subgrupo 3 presenta una distribución balanceada de leves y graves.

6. El análisis con *deep learning* mostró que el subgrupo 1 tenía un conjunto de secuencias CDR3b altamente predictivas, con un perfil de residuos críticos bastante restringido y capaces de diferenciarse de los subgrupos 2 y 3. Objetivo que no se pudo lograr con la clasificación inicial de leves y graves.
7. Se identificaron computacionalmente 59 secuencias altamente predictivas y exclusivas al subgrupo 1, que presentaban afinidad por la unión a antígenos del SARS-CoV y que a su vez se dividían en dos clados en base al uso de aminoácidos centrales.

8.Apéndice I. Figuras y Tablas suplementarias de Resultados (I)

Tabla suplementaria 8.1. Resultados del estudio de asociación entre sintomatología y grupos de pacientes por edad, gravedad y sexo. El p-valor se calculó mediante la prueba Chi-cuadrado. La V de Crammer se calculó para el tamaño del efecto.

Síntomas		Gravedad					Mayores de 55 años			Sexo			
		Total (N=173)	Leve (N=98)	Grave (N=75)	p-valor	Asoc.	<55 (N=93)	≥55 (N=80)	p-valor	Hombre (N=85)	Mujer (N=88)	p-valor	Asoc.
Afección dermatológica					0.004	0.223			0.946			0.161	
	NA	5	5	0			2	3		2	3		
	no	153 (91.1%)	90 (96.8%)	63 (84.0%)			83 (91.2%)	70 (90.9%)		73 (88.0%)	80 (94.1%)		
	si	15 (8.9%)	3 (3.2%)	12 (16.0%)			8 (8.8%)	7 (9.1%)		10 (12.0%)	5 (5.9%)		
Anosmia					0.482				0.822			0.145	
	NA	5	5	0			2	3		2	3		
	no	134 (79.8%)	76 (81.7%)	58 (77.3%)			72 (79.1%)	62 (80.5%)		70 (84.3%)	64 (75.3%)		
	si	34 (20.2%)	17 (18.3%)	17 (22.7%)			19 (20.9%)	15 (19.5%)		13 (15.7%)	21 (24.7%)		
Ageusia					0.04	0.158			0.691			0.211	
	NA	5	5	0			2	3		2	3		
	no	133 (79.2%)	79 (84.9%)	54 (72.0%)			71 (78.0%)	62 (80.5%)		69 (83.1%)	64 (75.3%)		
	si	35 (20.8%)	14 (15.1%)	21 (28.0%)			20 (22.0%)	15 (19.5%)		14 (16.9%)	21 (24.7%)		
Mialgias					< 0.001	0.377			0.082			0.122	
	NA	6	5	1			2	4		3	3		
	no	100 (59.9%)	71 (76.3%)	29 (39.2%)			49 (53.8%)	51 (67.1%)		54 (65.9%)	46 (54.1%)		
	si	67 (40.1%)	22 (23.7%)	45 (60.8%)			42 (46.2%)	25 (32.9%)		28 (34.1%)	39 (45.9%)		
Dolor de cabeza					0.191				0.985			0.011	0.197
	NA	6	5	1			3	3		3	3		
	no	117 (70.1%)	69 (74.2%)	48 (64.9%)			63 (70.0%)	54 (70.1%)		65 (79.3%)	52 (61.2%)		
	si	50 (29.9%)	24 (25.8%)	26 (35.1%)			27 (30.0%)	23 (29.9%)		17 (20.7%)	33 (38.8%)		
Fiebre					< 0.001	0.66			0.402			0.945	
	NA	6	6	0			2	4		2	4		
	no	7 (44.3%)	68 (73.9%)	6 (8.0%)			43 (47.3%)	31 (40.8%)		37 (44.6%)	37 (44.0%)		
	si	93 (55.7%)	24 (26.1%)	69 (92.0%)			48 (52.7%)	45 (59.2%)		46 (55.4%)	47 (56.0%)		
Disnea					< 0.001	0.59			0.187			0.319	
	NA	5	5	0			2	3		2	3		
	no	105 (62.5%)	82 (88.2%)	23 (30.7%)			61 (67.0%)	44 (57.1%)		55 (66.3%)	50 (58.8%)		
	si	63 (37.5%)	11 (11.8%)	52 (69.3%)			30 (33.0%)	33 (42.9%)		28 (33.7%)	35 (41.2%)		
Astenia					< 0.001	0.508			0.313			0.16	
	NA	5	5	0			2	3		2	3		
	no	90 (53.6%)	71 (76.3%)	19 (25.3%)			52 (57.1%)	38 (49.4%)		49 (59.0%)	41 (48.2%)		
	si	78 (46.4%)	22 (23.7%)	56 (74.7%)			39 (42.9%)	39 (50.6%)		34 (41.0%)	44 (51.8%)		
Tos					< 0.001	0.334			0.754				
	NA	0	0	0			0	0		0	0	0.617	
	no	141 (81.5%)	91 (92.9%)	50 (66.7%)			75 (80.6%)	66 (82.5%)		68 (80.0%)	73 (83.0%)		
	si	32 (18.5%)	7 (7.1%)	25 (33.3%)			18 (19.4%)	14 (17.5%)		17 (20.0%)	15 (17.0%)		

Tabla suplementaria 8.2. Resultados del estudio de asociación entre grupos de pacientes en relación con la clasificación por gravedad.

	Leve (n=98)	Grave (n=75)	Total (n=173)	p-valor
Sexo			0.334	0.334
Hombre	45 (45.90%)	40 (53.30%)	85 (49.10%)	
Mujer	53 (54.10%)	35 (46.70%)	88 (50.90%)	
Mayores de 55 años n (%)			0.102	0.102
<55	58 (59.20)	35 (46.70%)	93 (53.80%)	
≥ 55	40 (40.80%)	40 (53.30%)	80 (46.20%)	
Grupo combinado de mayores de 55 años y sexo del paciente (%)			0.151	0.151
Hombre;<55	27 (27.60%)	15 (20%)	42 (24.30%)	
Hombre;≥ 55	18 (18.40%)	25 (33,30%)	43 (24.90%)	
Mujer;<55	31 (31.60%)	20 (26.70%)	51 (29.50%)	
Mujer;≥ 55	22 (22.40%)	15 (20%)	37 (21.40%)	

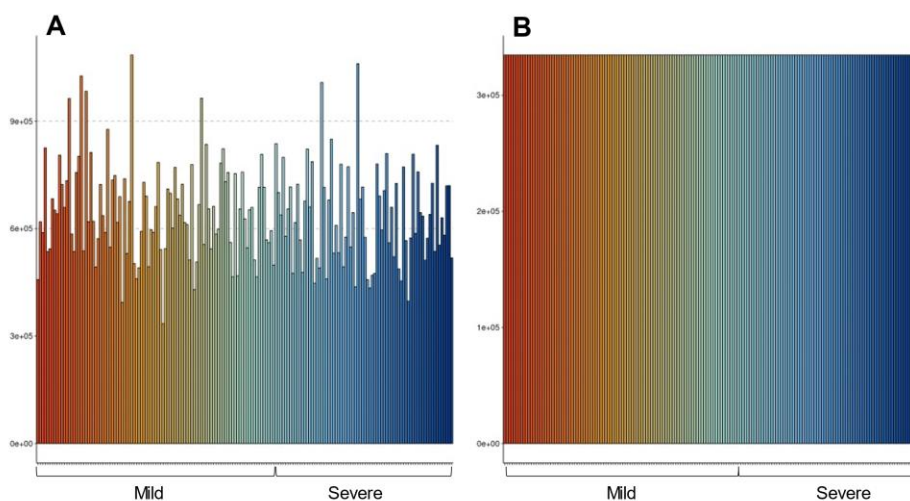


Figura suplementaria 8.1. Recuento de clonotipos leídos (profundidad del repertorio) para cada paciente antes (A) y después del *downsampling* (B).

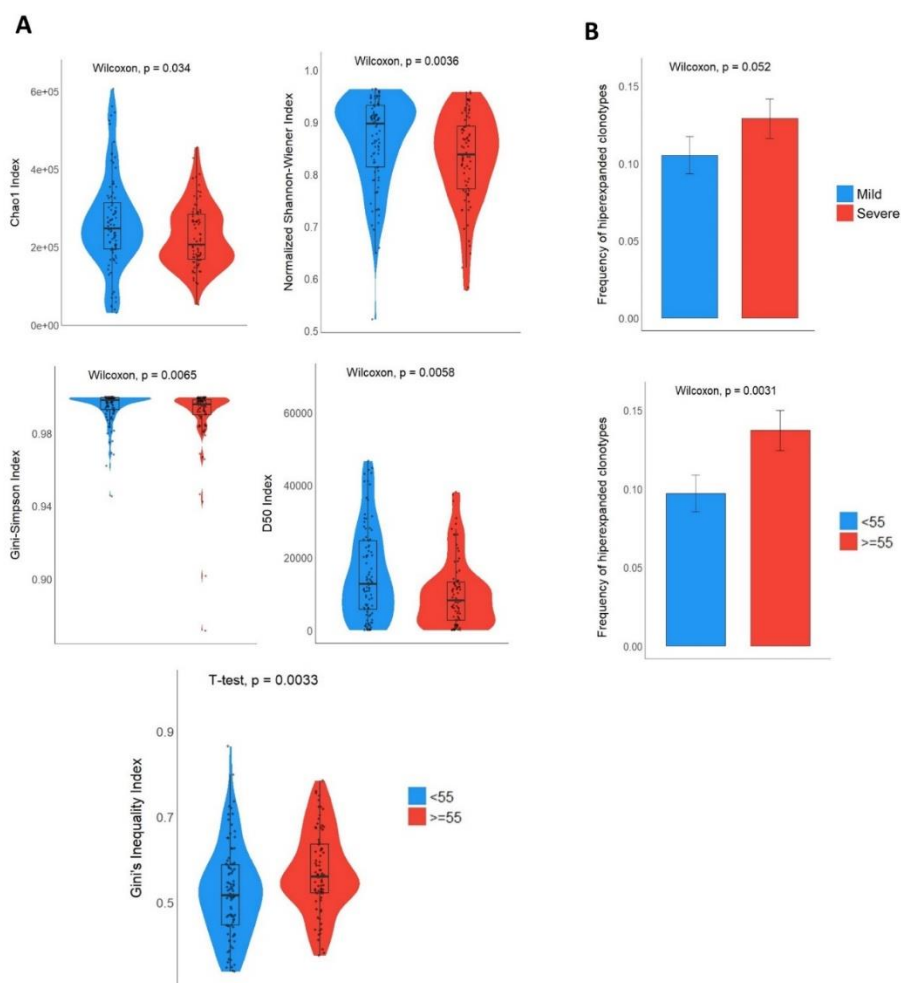


Figura suplementaria 8.2. Principales resultados de los análisis de α -diversidad y clonalidad. (A) Estimador de riqueza CHAO1, índices de α -diversidad de Shannon-Wiener normalizados e índice de Gini-Simpson, índice de clonalidad DE50 y coeficiente de desigualdad de Gini para grupos de pacientes divididos entre mayores y menores de 55 años. (B) Diferencias en la proporción de clonotipos hiperexpandidos entre pacientes leves y graves y entre pacientes mayores y menores de 55 años. La clasificación de los clones en función de su expansión clonal se basa en los intervalos de frecuencia de un clonotipo en un repertorio: Raro ($0-1E-5$), Pequeño ($1E-5-1E-4$), Mediano ($1E-4-0,001$), Grande ($0,001-0,01$), Hiperexpandido ($0,01-1$). Para los contrastes entre dos grupos, se utilizó la prueba U de Mann-Whitney para los datos no paramétricos y la prueba t para los datos paramétricos. Nivel de significancia de 0.05

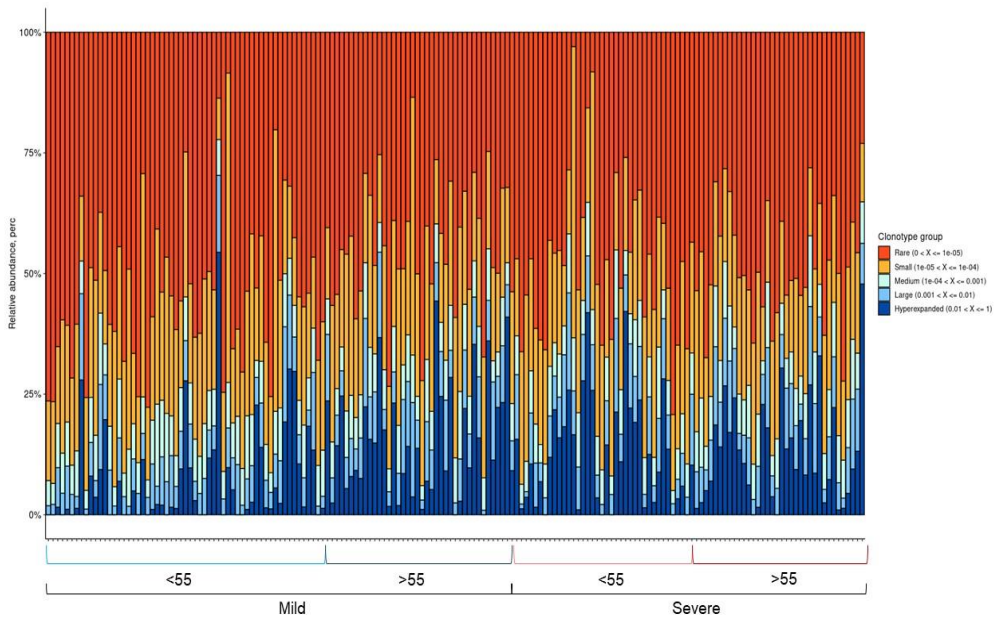


Figura suplementaria 8.3. Distribución de frecuencias en % de clones raros, medianos, grandes e hiperexpandidos en el total de los repertorios. El criterio del grado de expansión clonal se define por los intervalos de frecuencia que un clonotipo dado tiene en un repertorio.

Tabla suplementaria 8.3 Resultados de las comparaciones de las métricas de riqueza de especies, diversidad y clonalidad entre los grupos combinados de edad y gravedad, edad y sexo del paciente y gravedad y sexo del paciente. Para todos ellos se aplicó la prueba de Dunn con p-valor corregido por el método de Bonferroni-Holm. Nivel de significación de 0.05.

Índice	Grupo 1	Grupo 2	n1	n2	Estadístico	p-valor	p.adj
Shannon Wiener	Leve;<55	Leve;≥ 55	58	40	-3.697	0.000	0.001
	Leve;<55	Grave;<55	58	35	-2.703	0.007	0.014
	Leve;<55	Grave;≥ 55	58	40	-2.750	0.006	0.014
	Leve;≥ 55	Grave;<55	40	35	0.783	0.434	0.520
	Leve;≥ 55	Grave;≥ 55	40	40	0.871	0.384	0.520
	Grave;<55	Grave;≥ 55	35	40	0.058	0.954	0.954
CHAO1	Leve;<55	Leve;≥ 55	58	40	-3.257	0.001	0.007
	Leve;<55	Grave;<55	58	35	-2.405	0.016	0.049
	Leve;<55	Grave;≥ 55	58	40	-1.781	0.075	0.150
	Leve;≥ 55	Grave;<55	40	35	0.669	0.504	0.521
	Leve;≥ 55	Grave;≥ 55	40	40	1.357	0.175	0.262
	Grave;<55	Grave;≥ 55	35	40	0.642	0.521	0.521
Gini-Simpson	Leve;<55	Leve;≥ 55	58	40	-3.779	0.000	0.001
	Leve;<55	Grave;<55	58	35	-2.944	0.003	0.010
	Leve;<55	Grave;≥ 55	58	40	-2.572	0.010	0.020
	Leve;≥ 55	Grave;<55	40	35	0.633	0.526	0.632
	Leve;≥ 55	Grave;≥ 55	40	40	1.109	0.267	0.401
	Grave;<55	Grave;≥ 55	35	40	0.438	0.661	0.661

Tabla suplementaria 8.3 (cont.)

Índice	Grupo 1	Grupo 2	n1	n2	Estadístico	p-valor	p.adj
D50	Leve;<55	Leve;≥ 55	58	40	-3.707	0.000	0.001
	Leve;<55	Grave;<55	58	35	-2.773	0.006	0.017
	Leve;<55	Grave;≥ 55	58	40	-2.560	0.010	0.021
	Leve;≥ 55	Grave;<55	40	35	0.727	0.467	0.561
	Leve;≥ 55	Grave;≥ 55	40	40	1.054	0.292	0.438
	Grave;<55	Grave;≥ 55	35	40	0.291	0.771	0.771
Coefficiente de Gini	Leve;<55	Leve;≥ 55	58	40	3.963	0.000	0.000
	Leve;<55	Grave;<55	58	35	2.820	0.005	0.012
	Leve;<55	Grave;≥ 55	58	40	2.761	0.006	0.012
	Leve;≥ 55	Grave;<55	40	35	-0.911	0.362	0.435
	Leve;≥ 55	Grave;≥ 55	40	40	-1.105	0.269	0.404
	Grave;<55	Grave;≥ 55	35	40	-0.157	0.876	0.876

Tabla suplementaria 8.3 (cont.)

Índice	Grupo 1	Grupo 2	n1	n2	Estadístico	p-valor	p.adj
Shannon Wiener	Hombre;<55	Hombre;≥ 55	42	43	-0.417	0.677	0.677
	Hombre;<55	Mujer;<55	42	51	2.540	0.011	0.022
	Hombre;<55	Mujer;≥ 55	42	37	-1.015	0.310	0.465
	Hombre;≥ 55	Mujer;<55	43	51	2.994	0.003	0.008
	Hombre;≥ 55	Mujer;≥ 55	43	37	-0.618	0.537	0.644
	Mujer;<55	Mujer;≥ 55	51	37	-3.511	0.000	0.003
CHAO1	Hombre;<55	Hombre;≥ 55	42	43	-0.441	0.659	0.791
	Hombre;<55	Mujer;<55	42	51	1.977	0.048	0.096
	Hombre;<55	Mujer;≥ 55	42	37	-0.448	0.654	0.791
	Hombre;≥ 55	Mujer;<55	43	51	2.452	0.014	0.053
	Hombre;≥ 55	Mujer;≥ 55	43	37	-0.023	0.982	0.982
	Mujer;<55	Mujer;≥ 55	51	37	-2.375	0.018	0.053
Gini-Simpson	Hombre;<55	Hombre;≥ 55	42	43	-0.179	0.858	0.858
	Hombre;<55	Mujer;<55	42	51	2.464	0.014	0.028
	Hombre;<55	Mujer;≥ 55	42	37	-1.085	0.278	0.417
	Hombre;≥ 55	Mujer;<55	43	51	2.667	0.008	0.023
	Hombre;≥ 55	Mujer;≥ 55	43	37	-0.918	0.359	0.430
	Mujer;<55	Mujer;≥ 55	51	37	-3.510	0.000	0.003

Tabla suplementaria 8.3 (cont.)

Index	Grupo 1	Grupo 2	n1	n2	Estadístico	p-valor	p.adj
D50	Hombre;<55	Hombre;≥ 55	42	43	-0.504	0.614	0.667
	Hombre;<55	Mujer;<55	42	51	2.335	0.020	0.039
	Hombre;<55	Mujer;≥ 55	42	37	-0.912	0.362	0.543
	Hombre;≥ 55	Mujer;<55	43	51	2.878	0.004	0.012
	Hombre;≥ 55	Mujer;≥ 55	43	37	-0.430	0.667	0.667
	Mujer;<55	Mujer;≥ 55	51	37	-3.205	0.001	0.008
Gini Inequality	Hombre;<55	Hombre;≥ 55	42	43	0.632	0.528	0.633
	Hombre;<55	Mujer;<55	42	51	-2.547	0.011	0.022
	Hombre;<55	Mujer;≥ 55	42	37	0.951	0.342	0.513
	Hombre;≥ 55	Mujer;<55	43	51	-3.225	0.001	0.004
	Hombre;≥ 55	Mujer;≥ 55	43	37	0.345	0.730	0.730
	Mujer;<55	Mujer;≥ 55	51	37	3.450	0.001	0.003

Tabla suplementaria 8.3 (cont.)

Index	Grupo 1	Grupo 2	n1	n2	Estadístico	p-valor	p.adj
Shannon Wiener	Hombre;Leve;	Hombre;Grave;	45	40	-1.006	0.314	0.437
	Hombre;Leve;	Mujer;Leve;	45	53	1.284	0.199	0.434
	Hombre;Leve;	Mujer;Grave;	45	35	-0.037	0.970	0.970
	Hombre;Grave;	Mujer;Leve;	40	53	2.286	0.022	0.133
	Hombre;Grave;	Mujer;Grave;	40	35	0.908	0.364	0.437
	Mujer;Leve;	Mujer;Grave;	53	35	-1.234	0.217	0.434
CHAO1	Hombre;Leve;	Hombre;Grave;	45	40	-1.161	0.246	0.491
	Hombre;Leve;	Mujer;Leve;	45	53	0.694	0.488	0.732
	Hombre;Leve;	Mujer;Grave;	45	35	0.461	0.645	0.774
	Hombre;Grave;	Mujer;Leve;	40	53	1.876	0.061	0.364
	Hombre;Grave;	Mujer;Grave;	40	35	1.539	0.124	0.372
	Mujer;Leve;	Mujer;Grave;	53	35	-0.169	0.866	0.866
Gini_Simpson	Hombre;Leve;	Hombre;Grave;	45	40	-1.005	0.315	0.472
	Hombre;Leve;	Mujer;Leve;	45	53	1.081	0.280	0.472
	Hombre;Leve;	Mujer;Grave;	45	35	-0.228	0.820	0.820
	Hombre;Grave;	Mujer;Leve;	40	53	2.089	0.037	0.220
	Hombre;Grave;	Mujer;Grave;	40	35	0.722	0.471	0.565
	Mujer;Leve;	Mujer;Grave;	53	35	-1.242	0.214	0.472

Tabla suplementaria 8.3 (cont.)

Index	Grupo 1	Grupo 2	n1	n2	Estadístico	p-valor	p.adj
D50	Hombre;Leve	Hombre;Grave	45	40	-1.039	0.299	0.385
	Hombre;Leve	Mujer;Leve	45	53	1.170	0.242	0.385
	Hombre;Leve	Mujer;Grave	45	35	0.019	0.985	0.985
	Hombre;Grave	Mujer;Leve	40	53	2.210	0.027	0.163
	Hombre;Grave	Mujer;Grave	40	35	0.994	0.320	0.385
	Mujer;Leve	Mujer;Grave	53	35	-1.070	0.285	0.385
Gini Inequality	Hombre;Leve	Hombre;Grave	45	40	0.971	0.332	0.398
	Hombre;Leve	Mujer;Leve	45	53	-1.421	0.155	0.398
	Hombre;Leve	Mujer;Grave	45	35	-0.156	0.876	0.876
	Hombre;Grave	Mujer;Leve	40	53	-2.383	0.017	0.103
	Hombre;Grave	Mujer;Grave	40	35	-1.064	0.288	0.398
	Mujer;Leve	Mujer;Grave	53	35	1.161	0.246	0.398

Tabla suplementaria 8.4. Resultados de las comparaciones de las métricas de expansión clonal del repertorio entre los grupos combinados de edad y gravedad, para todos ellos se aplicó la prueba de Dunn con p-valor corregido con el método de Bonferroni-Holm. Nivel de significación de 0.05.

Metric	Grupo1	Grupo2	n1	n2	statistic	p-valor	p.adj
Rare	Leve;<55	Leve;>55	58	40	-3.545	0.000	0.002
	Leve;<55	Grave;<55	58	35	-2.385	0.017	0.051
	Leve;<55	Grave;>55	58	40	-2.011	0.044	0.089
	Leve;>55	Grave;<55	40	35	0.943	0.346	0.415
	Leve;>55	Grave;>55	40	40	1.411	0.158	0.237
	Grave;<55	Grave;>55	35	40	0.420	0.674	0.674
Small	Leve;<55	Leve;>55	58	40	0.620	0.535	0.937
	Leve;<55	Grave;<55	58	35	0.297	0.766	0.937
	Leve;<55	Grave;>55	58	40	-0.080	0.937	0.937
	Leve;>55	Grave;<55	40	35	-0.275	0.783	0.937
	Leve;>55	Grave;>55	40	40	-0.643	0.520	0.937
	Grave;<55	Grave;>55	35	40	-0.346	0.730	0.937
Medium	Leve;<55	Leve;>55	58	40	-0.257	0.798	0.798
	Leve;<55	Grave;<55	58	35	-0.696	0.487	0.798
	Leve;<55	Grave;>55	58	40	0.590	0.555	0.798
	Leve;>55	Grave;<55	40	35	-0.416	0.678	0.798
	Leve;>55	Grave;>55	40	40	0.778	0.437	0.798
	Grave;<55	Grave;>55	35	40	1.167	0.243	0.798
Large	Leve;<55	Leve;>55	58	40	1.996	0.046	0.138
	Leve;<55	Grave;<55	58	35	1.759	0.079	0.157
	Leve;<55	Grave;>55	58	40	2.134	0.033	0.138
	Leve;>55	Grave;<55	40	35	-0.145	0.884	0.899
	Leve;>55	Grave;>55	40	40	0.127	0.899	0.899
	Grave;<55	Grave;>55	35	40	0.268	0.788	0.899
Hyperexpanded	Leve;<55	Leve;>55	58	40	4.006	0.000	0.000
	Leve;<55	Grave;<55	58	35	3.183	0.001	0.004
	Leve;<55	Grave;>55	58	40	2.884	0.004	0.008
	Leve;>55	Grave;<55	40	35	-0.614	0.539	0.647
	Leve;>55	Grave;>55	40	40	-1.031	0.303	0.454
	Grave;<55	Grave;>55	35	40	-0.382	0.702	0.702

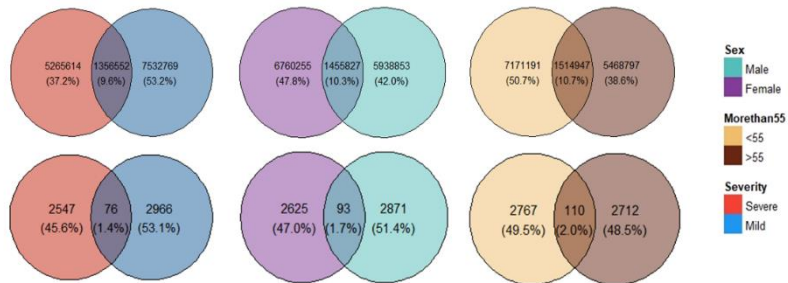
Tabla suplementaria 8.5. Resultados de las comparaciones de las métricas de expansión clonal del repertorio entre los grupos combinados de sexo y gravedad, para todos ellos se aplicó la prueba de Dunn con p-valor corregido con el método de Bonferroni-Holm. Nivel de significación de 0.05

Métrica	Grupo1	Grupo2	n1	n2	Estadístico	p-valor	p.adj
Raro	Hombre;Leve	Hombre;Grave	45	40	-0.989	0.323	0.629
	Hombre;Leve	Mujer;Leve	45	53	0.807	0.419	0.629
	Hombre;Leve	Mujer;Grave	45	35	0.399	0.690	0.735
	Hombre;Grave	Mujer;Leve	40	53	1.808	0.071	0.424
	Hombre;Grave	Mujer;Grave	40	35	1.317	0.188	0.563
	Mujer;Leve	Mujer;Grave	53	35	-0.339	0.735	0.735
Pequeño	Hombre;Leve	Hombre;Grave	45	40	0.256	0.798	0.931
	Hombre;Leve	Mujer;Leve	45	53	0.471	0.638	0.931
	Hombre;Leve	Mujer;Grave	45	35	-0.086	0.931	0.931
	Hombre;Grave	Mujer;Leve	40	53	0.190	0.849	0.931
	Hombre;Grave	Mujer;Grave	40	35	-0.324	0.746	0.931
	Mujer;Leve	Mujer;Grave	53	35	-0.527	0.598	0.931
Mediano	Hombre;Leve	Hombre;Grave	45	40	0.668	0.504	0.504
	Hombre;Leve	Mujer;Leve	45	53	-0.857	0.391	0.504
	Hombre;Leve	Mujer;Grave	45	35	-1.471	0.141	0.283
	Hombre;Grave	Mujer;Leve	40	53	-1.522	0.128	0.283
	Hombre;Grave	Mujer;Grave	40	35	-2.059	0.039	0.237
	Mujer;Leve	Mujer;Grave	53	35	-0.724	0.469	0.504
Expandido	Hombre;Leve	Hombre;Grave	45	40	0.876	0.381	0.572
	Hombre;Leve	Mujer;Leve	45	53	-1.021	0.307	0.572
	Hombre;Leve	Mujer;Grave	45	35	0.274	0.784	0.784
	Hombre;Grave	Mujer;Leve	40	53	-1.896	0.058	0.347
	Hombre;Grave	Mujer;Grave	40	35	-0.555	0.579	0.695
	Mujer;Leve	Mujer;Grave	53	35	1.234	0.217	0.572
Hiperexpandido	Hombre;Leve	Hombre;Grave	45	40	1.247	0.212	0.425
	Hombre;Leve	Mujer;Leve	45	53	-1.069	0.285	0.428
	Hombre;Leve	Mujer;Grave	45	35	0.345	0.730	0.730
	Hombre;Grave	Mujer;Leve	40	53	-2.329	0.020	0.119
	Hombre;Grave	Mujer;Grave	40	35	-0.835	0.404	0.484
	Mujer;Leve	Mujer;Grave	53	35	1.352	0.176	0.425

Tabla suplementaria 8.6. Resultados de las comparaciones de las métricas de expansión clonal del repertorio entre la edad combinada y los grupos de edad, para todos ellos se aplicó la prueba de Dunn con el P valor corregido con el método de Bonferroni-Holm. Nivel de significación de 0.05.

Métrica	Grupo1	Grupo2	n1	n2	Estadístico	p-valor	p.adj
Raro	Hombre;<55	Hombre;>55	42	43	-0.476	0.634	0.752
	Hombre;<55	Mujer;<55	42	51	2.124	0.034	0.067
	Hombre;<55	Mujer;>55	42	37	-0.773	0.440	0.659
	Hombre;>55	Mujer;<55	43	51	2.637	0.008	0.025
	Hombre;>55	Mujer;>55	43	37	-0.316	0.752	0.752
	Mujer;<55	Mujer;>55	51	37	-2.856	0.004	0.025
Pequeño	Hombre;<55	Hombre;>55	42	43	0.376	0.707	0.957
	Hombre;<55	Mujer;<55	42	51	0.338	0.735	0.957
	Hombre;<55	Mujer;>55	42	37	0.253	0.800	0.957
	Hombre;>55	Mujer;<55	43	51	-0.054	0.957	0.957
	Hombre;>55	Mujer;>55	43	37	-0.109	0.913	0.957
	Mujer;<55	Mujer;>55	51	37	-0.062	0.950	0.957
Mediano	Hombre;<55	Hombre;>55	42	43	0.102	0.919	0.919
	Hombre;<55	Mujer;<55	42	51	-1.624	0.104	0.313
	Hombre;<55	Mujer;>55	42	37	-1.028	0.304	0.456
	Hombre;>55	Mujer;<55	43	51	-1.742	0.082	0.313
	Hombre;>55	Mujer;>55	43	37	-1.133	0.257	0.456
	Mujer;<55	Mujer;>55	51	37	0.494	0.622	0.746
Expandido	Hombre;<55	Hombre;>55	42	43	-0.003	0.998	0.998
	Hombre;<55	Mujer;<55	42	51	-1.991	0.046	0.093
	Hombre;<55	Mujer;>55	42	37	0.533	0.594	0.713
	Hombre;>55	Mujer;<55	43	51	-2.001	0.045	0.093
	Hombre;>55	Mujer;>55	43	37	0.539	0.590	0.713
	Mujer;<55	Mujer;>55	51	37	2.477	0.013	0.079
Hiperexpandido	Hombre;<55	Hombre;>55	42	43	0.476	0.634	0.634
	Hombre;<55	Mujer;<55	42	51	-2.382	0.017	0.034
	Hombre;<55	Mujer;>55	42	37	1.189	0.234	0.352
	Hombre;>55	Mujer;<55	43	51	-2.896	0.004	0.011
	Hombre;>55	Mujer;>55	43	37	0.735	0.462	0.555
	Mujer;<55	Mujer;>55	51	37	3.540	0.000	0.002

A)



B)



Figura suplementaria 8.4 (A). Superposición del repertorio con el número de clonotipos totales e hiperexpandidos compartidos entre los grupos de gravedad, edad y sexo y (B) los grupos combinados de gravedad-edad, gravedad-sexo y sexo-edad.

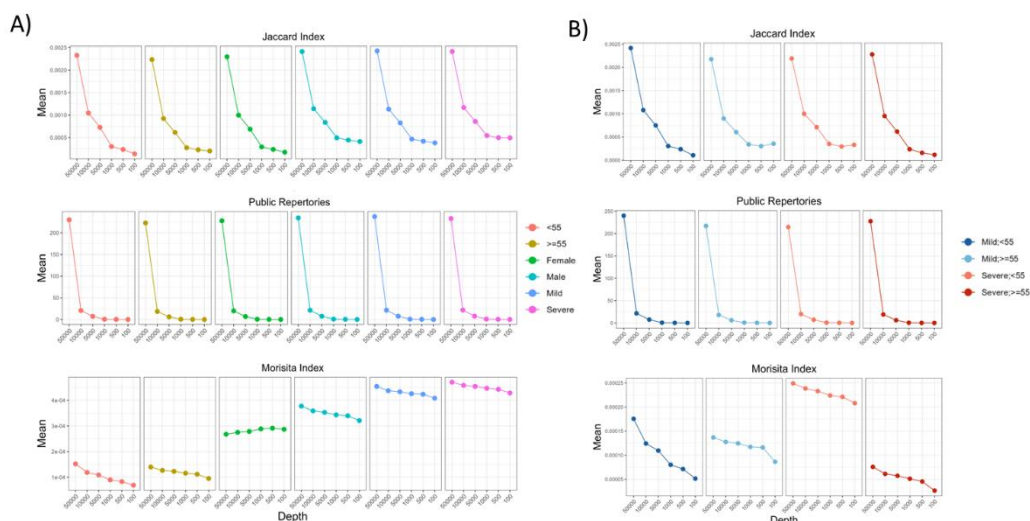


Figura suplementaria 8.5. Índices promedio de Jaccard, Morisita y recuento de repertorios en los diferentes grupos de pacientes (A) y en el grupo conjunto de edad y gravedad (B), seleccionando los 100, 500, 1000, 5000, 10000 y 50000 mejores clones expandidos de cada repertorio. En este caso, se ha considerado la secuencia de aminoácidos de los alelos CDR3 y V-J.

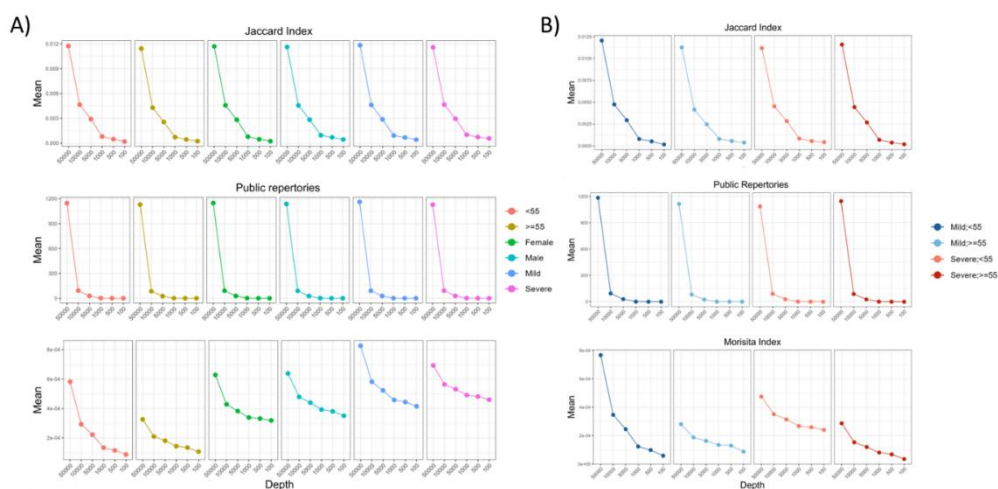


Figura suplementaria 8.6. Índices promedio de Jaccard, Morisita y recuento de repertorios en los diferentes grupos de pacientes (A) y en el grupo conjunto de edad y gravedad (B), seleccionando los 100, 500, 1000, 5000, 10000 y 50000 mejores clones expandidos de cada repertorio. En este caso, sólo se ha considerado la secuencia de aminoácidos de CDR3.

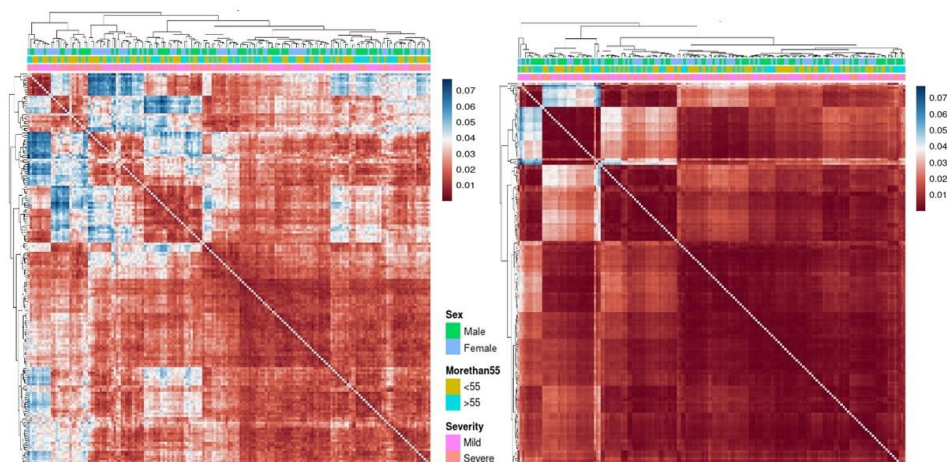


Figura suplementaria 8.7. Índice de Jaccard-Shannon calculado para la distribución de alelos V (A) y J (B).

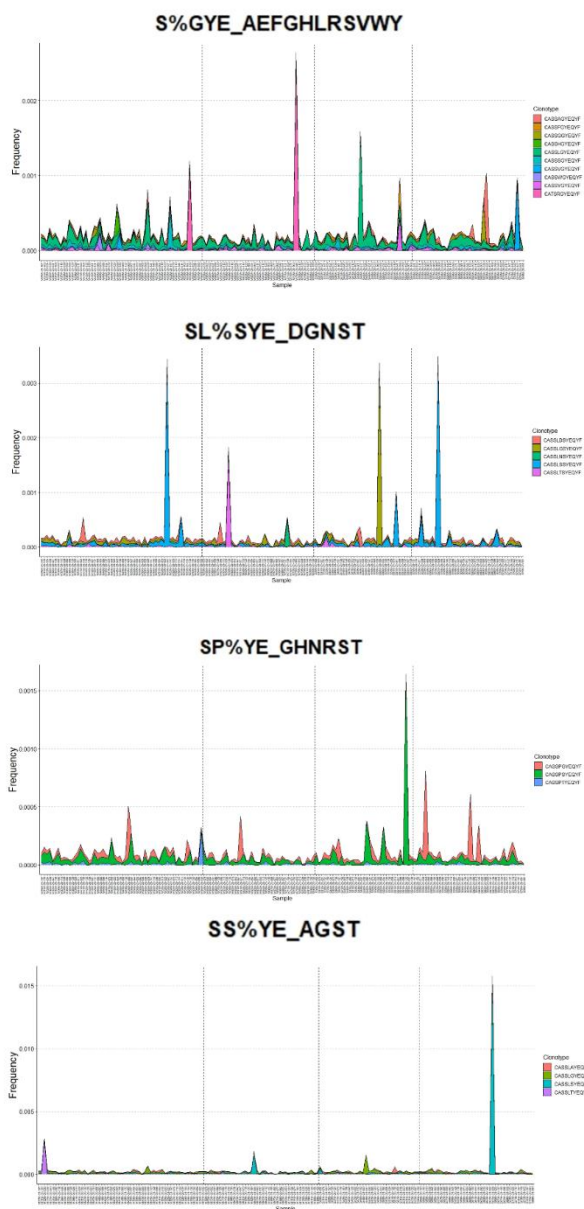


Figura suplementaria 8.8. Distribución de frecuencias por repertorio/muestra de cada secuencia CDR3b COVID-19+ para cada clúster GLIPH2 significativo entre los grupos combinados de edad y gravedad.

Tabla suplementaria 8.7. Se muestran los resultados de la validación cruzada con simulación Monte Carlo de 1000 repeticiones. Se muestra el número de veces que cada motivo COVID-19+ es seleccionado en el modelo basado en el cálculo AIC. Además, se proporciona el AUC medio de las 1000 simulaciones para cada comparación.

Motivos estructurales	Severe;<55 v Leve;<55 AUC=0.595	Severe;<55 v Leve; ≥55 AUC=0.59	Severe;<55 vs Severe; ≥55 AUC=0.592
S%GET_AFLVWY [†]	921	518	220
SPT%AGE_ASV*	889	756	843
SLGL%YE_GHNRS [†]	633	133	938
YSSGE_4_22	623	216	150
G%YE_DEGNS	251	414	191
S%GSYE_EGQST	250	491	373
SLG%YE_AGSTV	236	450	398
SL%SYE_DGNST [‡]	220	708	392
S%GGE_FIY	183	753	794
SS%YE_AGST	182	245	599
SP%TGE_DEGQRS	175	330	194
S%GYE_AEFGHLRSVWY [‡]	173	223	129
SL%YE_AGST [‡]	172	468	655
S%GGYE_FLY	167	407	280
SPG%E_DHNTWY	142	499	216
S%SYE_AFGHLSVY	141	386	733
SY%GE_GST	128	852	425
SVG%E_DN	112	139	301
SP%YE_GHNRST	87	317	155
S%GNE_AGILV	38	673	263

* SPT%AGE_ASV se selecciona en aproximadamente el 75-89% de los modelos tras la selección de características.

[†] SLGL%YE_GHNRS y S%GET_AFLVWY destacan para las comparaciones entre pacientes graves mayores de 55 años frente a menores de 55 años, y pacientes leves frente a graves menores de 55 años, respectivamente.

[‡] SL%SYE_DGNST y SL%YE_AGST, tienen un efecto de moderado a alto en comparaciones específicas, mientras que S%GYE-AEFGHLRSVWY muestra una relevancia baja.

Tabla suplementaria 8.8. Se muestran los resultados de la validación cruzada con simulación Monte Carlo de 1000 repeticiones para las comparaciones de grupos de gravedad y edad. Se muestra el número de veces que cada motivo COVID-19+ es seleccionado en el modelo basado en el cálculo AIC. Además, se proporciona el AUC medio de las 1000 simulaciones para cada comparación.

COVID-19 + motifs	Leve vs Grave AUC=0.54	<55 vs ≥55 AUC=0.56
SPT%AGE_ASV*	825	679
SLG%YE_AGSTV†	739	407
S%GGYE_FLY	633	57
S%GSYE_EGQST	503	707
S%GET_AFLVWY	449	517
S%SYE_AFGHLSVY‡	398	787
YSSGE_4_22	378	417
SPG%E_DHNTWY	347	109
SY%GE_GST	238	244
S%GGE_FIY	234	177
SS%YE_AGST‡	207	721
G%YE_DEGNS	203	73
SL%SYE_DGNST	197	116
SP%YE_GHNRST	161	75
SL%YE_AGST	153	636
S%GNE_AGILV	130	552
SVG%E_DN	129	348
SLGL%YE_GHNRS	90	120
SP%TGE_DEGQRS	67	187
S%GYE_AEFGHLRSVWY	45	177

* SPT%AGE-ASV se selecciona 82,5% y 67,9% en modelos de regresión logística leve frente a grave y <55 frente a ≥ 55 años, respectivamente.

† SLG%YE_AGSTV sigue siendo relevante y modelos leves vs grave.

‡ S%SYE_AFGHLSVY y SS%YE_AGST son relevantes en los modelos de regresión logística <55 frente a ≥ 55 años.

9. Apéndice II. Figuras y tablas suplementarias de Resultados (II)

Población CyTOF (Q1, mediana, Q3)	Leves (n = 33)	Graves (n = 28)	Total (n = 61)
Monocitos Clásicos			
CM-CD38+	96.23, 98.48, 99.23	95.96, 98.26, 99.32	96.17, 98.33, 99.26
CM-CD33+	86, 99, 100	86, 95, 99	86, 98, 100
CM-CD11b+	80, 93, 97	89, 96, 98	84, 95, 98
CM-CD36+	99.70, 99.91, 99.97	99.80, 99.94, 99.99	99.73, 99.93, 99.98
CM-CD163+CD206-	69, 78, 85	74, 83, 87	73, 80, 86
CM-CD163-CD206+	0.22, 0.44, 0.67	0.17, 0.34, 0.59	0.17, 0.38, 0.65
CM-CD163+CD206+	4.02, 5.25, 6.59	3.92, 4.97, 7.29	4.02, 5.21, 6.76
CM-CD163-CD206-	8, 15, 23	6, 11, 20	8, 13, 20
Monocitos Transicionales			
TM-CD38+	68, 81, 94	67, 82, 95	68, 81, 94
TM-CD33+	93, 98, 99	88, 96, 99	90, 97, 99
TM-CD11b+	56, 65, 74	50, 63, 80	55, 65, 77
TM-CD36+	96.08, 98.42, 99.72	96.53, 98.36, 99.71	96.41, 98.42, 99.72
TM-CD163+CD206-	19, 53, 62	46, 62, 75	34, 55, 64
TM-CD163-CD206+	1.1, 2.3, 5.5	0.7, 1.1, 2.7	0.8, 1.7, 4.0
TM-CD163+CD206+	3, 8, 18	8, 13, 21	6, 10, 20
TM-CD163-CD206-	12, 38, 71	9, 16, 38	11, 29, 50
Monocitos No Clásicos			
NCM-CD38+	4, 9, 11	5, 6, 16	5, 8, 13
NCM-CD33+	37, 49, 74	18, 38, 66	27, 44, 68
NCM-CD11b+	0.6, 2.5, 5.1	1.5, 2.4, 5.4	1.4, 2.5, 5.4
NCM-CD36+	51, 55, 66	53, 56, 64	52, 55, 64
NCM-CD163+CD206-	3, 5, 9	3, 7, 11	3, 6, 10
NCM-CD163-CD206+	0.73, 1.38, 2.54	1.10, 1.46, 3.25	0.79, 1.44, 2.64
NCM-CD163+CD206+	0.08, 0.41, 0.67	0.26, 0.54, 0.99	0.12, 0.49, 0.86
NCM-CD163-CD206-	87, 92, 94	82, 91, 93	85, 91, 94

Tabla suplementaria 9.1 (cont.)

Población CyTOF (Q1, mediana, Q3)	Leves (n = 33)	Graves (n = 28)	Total (n = 61)
Monocitos 14 dim			
14DIM-CD38+	59, 76, 85	54, 69, 80	56, 71, 83
14DIM-CD33+	32, 45, 53	26, 34, 46	27, 38, 50
14DIM-CD11b+	9, 17, 26	12, 15, 19	11, 16, 22
14DIM-CD36+	42, 51, 59	44, 50, 59	43, 50, 59
14DIM- CD163+CD206-	11, 15, 23	9, 14, 19	10, 15, 21
14DIM-CD163- CD206+	6, 9, 11	6, 8, 13	6, 9, 12
14DIM- CD163+CD206+	4.8, 7.5, 10.9	3.8, 5.5, 10.6	4.3, 6.4, 10.9
14DIM-CD163- CD206-	60, 65, 72	54, 70, 77	59, 68, 75
Linfocitos T totales (CD3+)	21, 23, 28	21, 25, 29	21, 23, 28
Granulocitos	47, 54, 59	47, 54, 58	47, 54, 59
Linfocitos B	2.19, 2.94, 3.78	1.96, 2.78, 3.69	2.14, 2.82, 3.75
Células Natural Killer	2.37, 3.24, 4.23	1.92, 2.84, 4.85	2.20, 3.18, 4.58
Células no leucocitarias (CD45-)	1.5, 4.1, 6.8	0.9, 1.5, 3.8	1.1, 2.6, 5.1

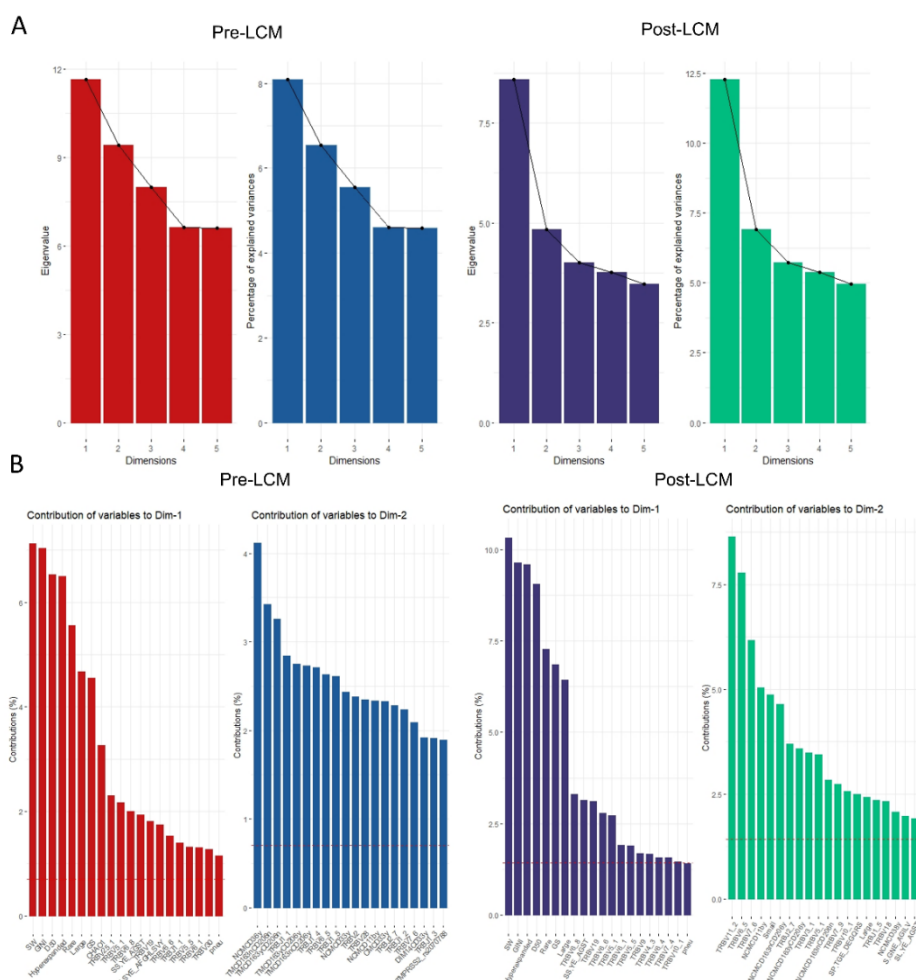


Figura suplementaria 9.1 Resultados de la FMAD en los conjuntos pre-LCM y post-LCM. (A) Valores propios y porcentajes de varianza explicada por las cinco primeras dimensiones. (B) Porcentaje de contribución en las dimensiones 1 y 2 de las 20 variables principales.

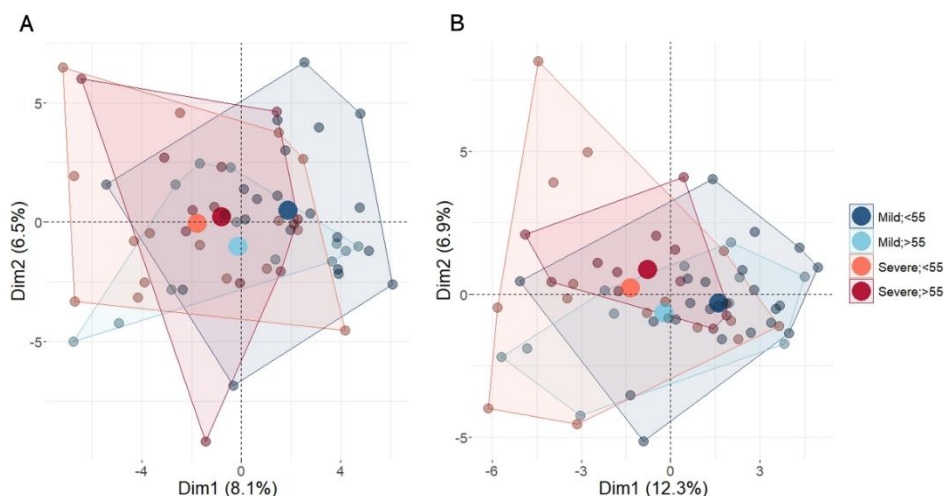


Figura suplementaria 9.2. Gráficos FAMD para la reducción de la dimensionalidad de las observaciones etiquetadas por grupos combinados de gravedad y edad en (A) el conjunto de datos pre-LCM y (B) post-LCM.

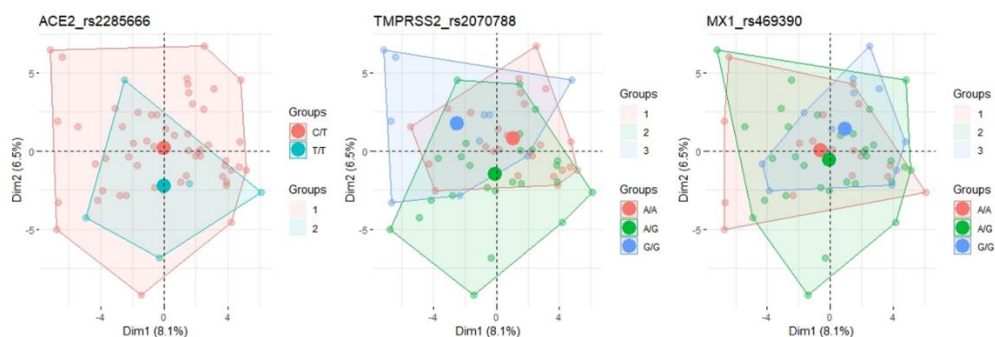


Figura suplementaria 9.3. Gráfico FAMD para la reducción de la dimensionalidad de las observaciones etiquetadas por su genotipo en los tres SNP estudiados.

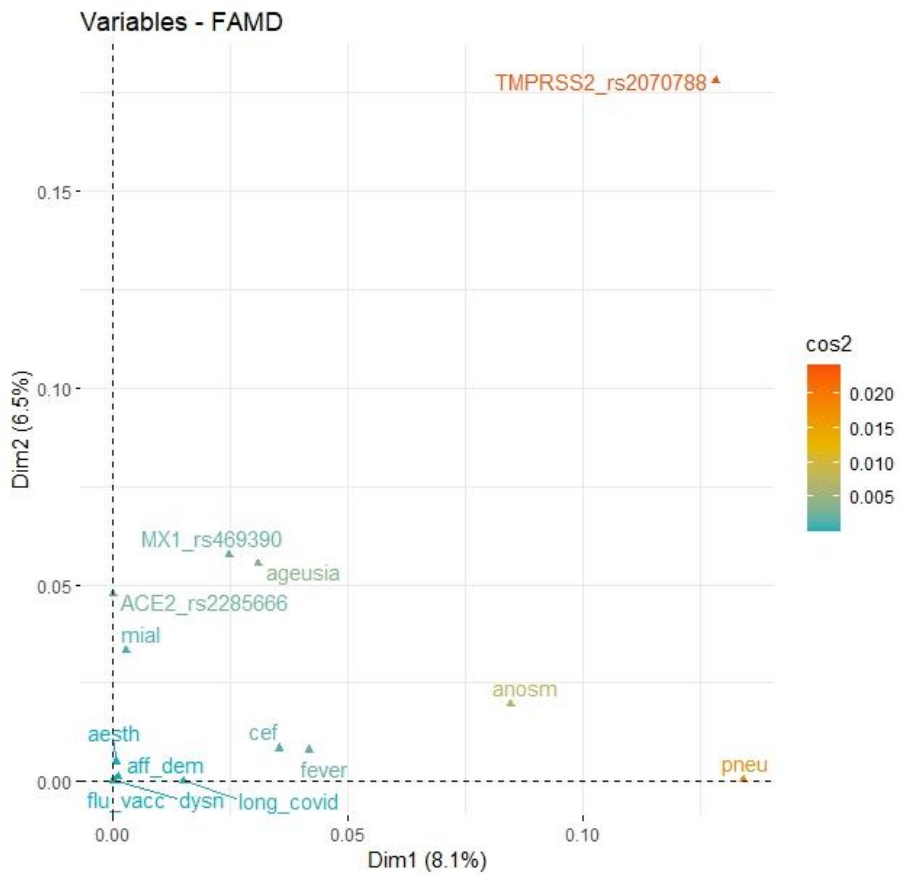


Figura suplementaria 9.4. Gráfico de contribución de las variables cualitativas en el conjunto de datos pre-FAMD. Las variables con valores cos2 cercanos a 1 están bien representadas en el eje.

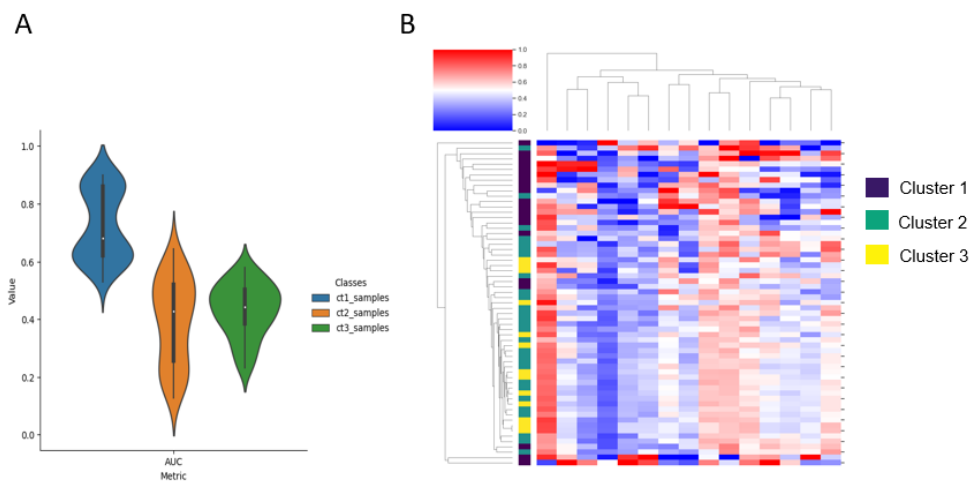


Figura suplementaria 9.5. Principales resultados del análisis no supervisado utilizando la Autocodificación Variacional (VAE) de DeepTCR en los 1000 clonotipos más expandidos de cada muestra. (A) Valores AUC de cada clase tras aplicar una clasificación no supervisada K-Nearest Neighbourhood con 5 pliegues, utilizando la distancia euclidiana, al modelo VAE entrenado. (B) Mapa de calor con un dendrograma basado en la distancia euclidiana de los valores medios de las 15 dimensiones latentes principales del modelo VAE entrenado.

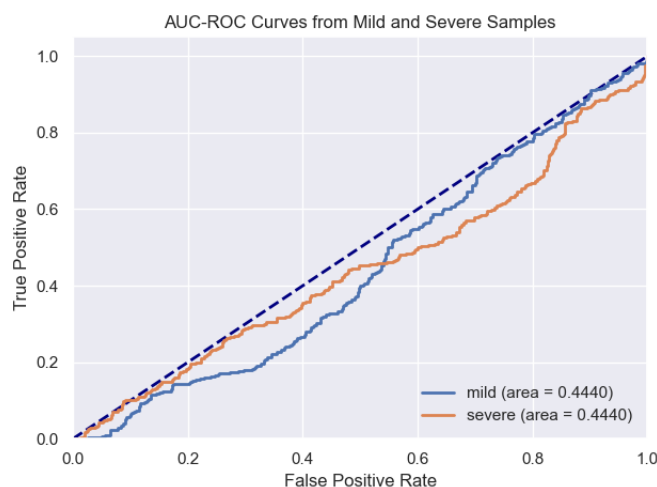


Figura suplementaria 9.6. Curva AUC-ROC de DeepTCR del modelo de deep learning supervisado para comparar muestras divididas por casos leves o graves. El modelo se entrenó utilizando los 1000 clonotipos principales de cada muestra. Los análisis se validaron utilizando 100 simulaciones de Monte Carlo y una división del 75 % de entrenamiento y el 25 % de prueba.

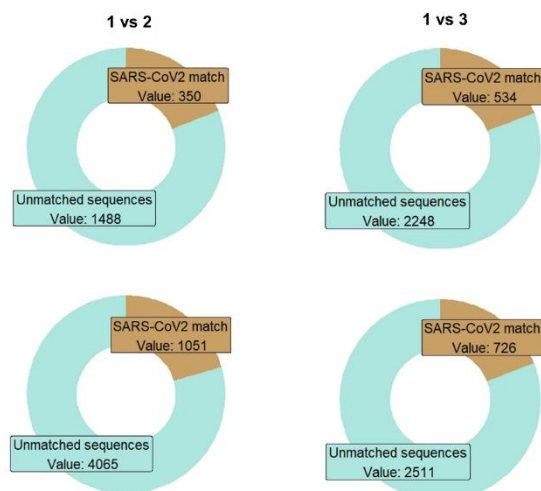


Figura suplementaria 9.7. Distribución del número de secuencias CDR3b seleccionadas (AUC igual o superior a 0,99) de las comparaciones clúster 1 vs clúster 2 (izquierda) y clúster 1 vs clúster 3 (derecha). División basada en el número de CDR3b identificadas como reactivas al SARS-CoV2 por TCRmatch

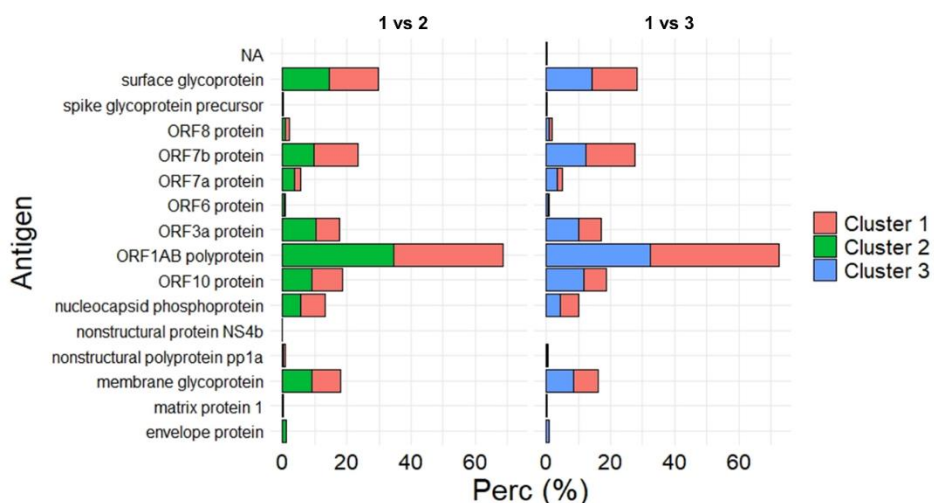


Figura suplementaria 9.8. Distribución de frecuencias en el uso de antígenos CDR3b seleccionados reactivos al SARS-Cov2, por clúster y en cada comparación (clúster 1 vs clúster 2 y clúster 1 vs clúster 3). La relación CDR3b-antígeno se extrajo de la base de datos IEDB.

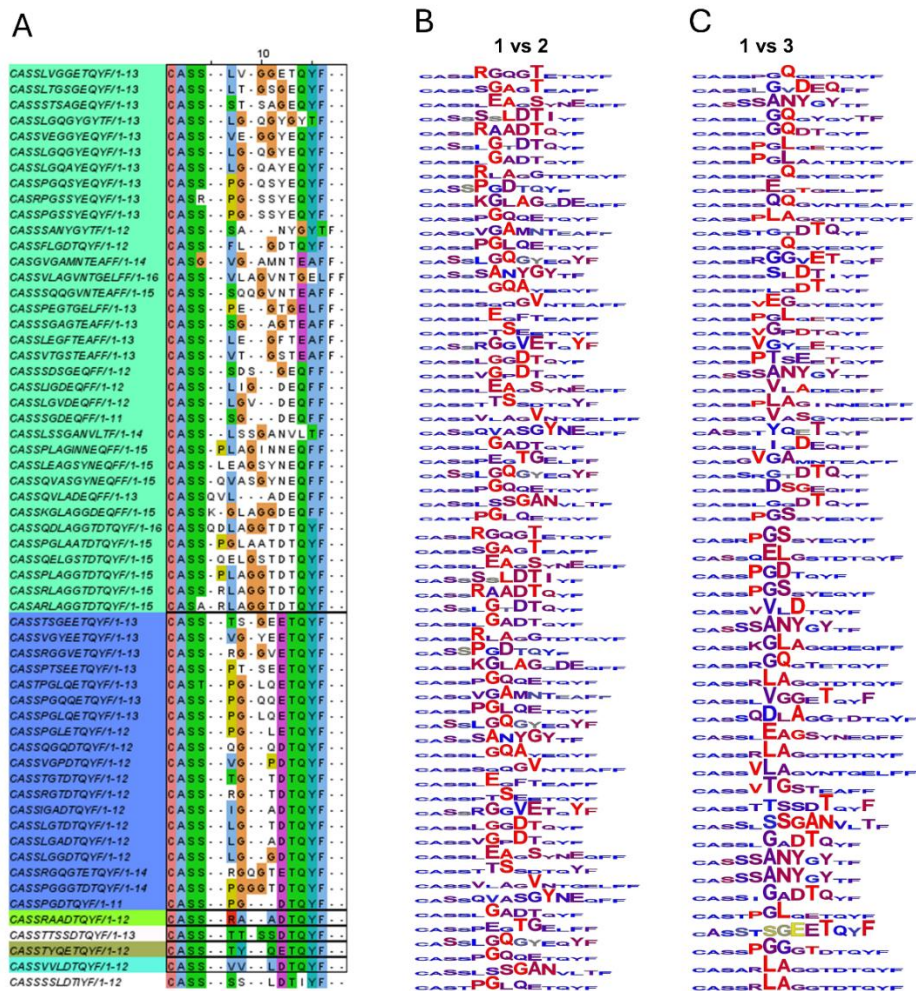


Figura suplementaria 9.9. Características de la secuencia de aminoácidos de las CDR3 reactivas y comunes del SARS-CoV2 en el clúster 1. (A) Visualización de los resultados del MSA con distinción de los clades principales. (B) Gráfico RSL de las selecciones de CDR3b dentro del clúster 1 frente a la comparación del clúster 2 y (C) clúster 2 frente a clúster 3.

10.Apéndice III. Publicaciones científicas

RESEARCH

Open Access



The T-cell repertoire of Spanish patients with COVID-19 as a strategy to link T-cell characteristics to the severity of the disease

Fernando Marín-Benesiu^{1,2†}, Lucia Chica-Redecillas^{1,2†}, Verónica Arenas-Rodríguez³, Esperanza de Santiago², Silvia Martínez-Diz³, Ginesa López-Torres⁴, Ana Isabel Cortés-Valverde⁴, Catalina Romero-Cachinero⁵, Carmen Entrala-Bernal⁶, Francisco Javier Fernandez-Rosado⁶, Luis Javier Martínez-González^{1,2†} and Maria Jesus Alvarez-Cubero^{1,2,7†}

Abstract

Background The architecture and dynamics of T cell populations are critical in orchestrating the immune response to SARS-CoV-2. In our study, we used T Cell Receptor sequencing (TCRseq) to investigate TCR repertoires in 173 post-infection COVID-19 patients.

Methods The cohort included 98 mild and 75 severe cases with a median age of 53. We amplified and sequenced the TCR β chain Complementary Determining Region 3 (CDR3b) and performed bioinformatic analyses to assess repertoire diversity, clonality, and V/J allelic usage between age, sex and severity groups. CDR3b amino acid sequence inference was performed by clustering structural motifs and filtering validated reactive CDR3b to COVID-19.

Results Our results revealed a pronounced decrease in diversity and an increase in clonal expansion in the TCR repertoires of severe COVID-19 patients younger than 55 years old. These results reflect the observed trends in patients older than 55 years old (both mild and severe). In addition, we identified a significant reduction in the usage of key V alleles (TRBV14, TRBV19, TRBV15 and TRBV6-4) associated with disease severity. Notably, severe patients under 55 years old had allelic patterns that resemble those over 55 years old, accompanied by a skewed frequency of COVID-19-related motifs.

Conclusions Present results suggest that severe patients younger than 55 may have a compromised TCR repertoire contributing to a worse disease outcome.

Keywords T cells, SARS-Cov2, Immunoinformatics, COVID-19, Adaptative immunology, TCR repertoire

[†]Luis Javier Martínez-González and Maria Jesus Alvarez-Cubero contributed equally to this work.

*Correspondence:
 Luis Javier Martínez-González
luisjavier.martinez@genyo.es

Full list of author information is available at the end of the article



© The Author(s) 2024 **Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.



Contents lists available at ScienceDirect

Computational and Structural Biotechnology Journal

journal homepage: www.elsevier.com/locate/csbj

Research article

Integration of T cell repertoire, CyTOF, genotyping and symptomatology data reveals subphenotypic variability in COVID-19 patients

Fernando Marín-Benesi^{a,b}, Lucia Chica-Redecillas^{a,b}, Sergio Cuenca-López^{c,d},
Carmen Entrala-Bernal^e, Sara Martín-Esteban^d, Maria Jesús Alvarez-Cubero^{a,b,e,*},
Luis Javier Martínez-González^{a,b,1}

^a Department of Biochemistry, Molecular Biology III and Immunology, Faculty of Medicine, University of Granada, Parque Tecnológico de la Salud, Granada 18016, Spain

^b GENYO, Centre for Genomics and Oncological Research: Pfizer, University of Granada, Andalusian Regional Government, Parque Tecnológico de la Salud, Granada 18016, Spain

^c LORGEN G.P., PI, Ciencias de la Salud - Business Innovation Centre (BIC), Armilla 18100, Spain

^d Montefrío Health Center, Metropolitan District of Granada, Montefrío 18009, Spain

^e Ibs Granada, Biosanitary Research Institute of Granada, Granada 18012, Spain



ARTICLE INFO

Keywords:
Cellular immunology
bioinformatics
COVID-19
T-cell repertoire
Latent Class Model

ABSTRACT

COVID-19 manifests a broad spectrum of clinical outcomes, from asymptomatic cases to severe disease. While several biomarkers have been proposed, comprehensive immunological analyses integrating mass cytometry (CyTOF) and T-cell receptor sequencing (TCRseq) data remain limited. In this study, we applied the Latent Class Model based on the Bayesian Information Criterion (LCM-BIC) algorithm to integrate immunophenotyping, including monocyte-macrophage counts from CyTOF, T-cell receptor repertoire data via TCRseq, SNPs data from ACE2 (rs2285666), MX1 (rs469390), and TMPRSS2 (rs2070788), and symptomatology data from 61 Spanish COVID-19 patients (33 mild, 28 severe). We identified three novel and distinct patient clusters with significant differences in TCR diversity, monocyte subpopulations, and V allele usage and disease outcome. Cluster 1 was predominantly enriched in severe cases, characterized by unique immunological features. Deep learning analysis of TCR amino acid sequences further distinguished Cluster 1 from the others, identifying SARS-CoV-2-specific TCR sequences associated with disease severity. In addition, analysis of residue sensitivity of cluster 1 SARS-CoV-2-specific TCR sequences further identified conserved aminoacids located in key central positions of the complementarity-determining region 3. This study highlights the value of integrating immunophenotyping and genetic profiling to identify novel immunological markers and patterns, aiding in the stratification and management of COVID-19 patients based on their immune profiles and genetic background.

1. Introduction

COVID-19, caused by SARS-CoV-2, exhibits a wide spectrum of symptoms ranging from asymptomatic cases to critical, even fatal outcomes. Numerous studies have proposed molecular biomarkers, including classical serum markers such as LDH, CRP and IL-6 [1], as well as genetic and gene expression analyses of key genes such as *TMPRSS2*,

ACE2, *MX1*, *IFITM3* and *CXCL10* [2–5]. It should be noted, however, that these markers are not fully specific to COVID-19 and only cover a small part of the complex development of the disease. Therefore research is required to gain a better understanding of the immune response.

A promising approach for patient stratification involves studying the immune response, focusing particularly on the innate and adaptive

Abbreviations: ACE2, Angiotensin-Converting Enzyme 2; AUC, Area Under the Curve; CDR3, Complementarity Determining Region 3; CEI, Ethics Committee; COVID-19, Coronavirus Disease 2019; CRP, C-Reactive Protein; CyTOF, Cytometry by Time of Flight; HIF1A, Hypoxia-Inducible Factor 1-Alpha; IL-6, Interleukin 6; LDH, Lactate Dehydrogenase; LCA, Latent Class Analysis; LCM, Latent Class Model; MX1, MX Dynamine Like GTPase 1; RT-PCR, Reverse Transcription Polymerase Chain Reaction; TCR, T-Cell Receptor; TCRseq, T-Cell Receptor Sequencing; TMPRSS2, Transmembrane Protease, Serine 2.

* Corresponding author at: Medicine Faculty, Av. de la Investigación nº 11, Tower C, 11th floor 18071, Granada, Spain.

E-mail address: njesusac@ugr.es (M.J. Alvarez-Cubero).

¹ The authors have contributed equally to this work

<https://doi.org/10.1016/j.csbj.2025.05.016>

Received 26 December 2024; Received in revised form 13 May 2025; Accepted 13 May 2025

Available online 14 May 2025

2001-0370/© 2025 The Author(s). Published by Elsevier B.V. on behalf of Research Network of Computational and Structural Biotechnology. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

Referencias

- [1] Gorbalenya AE, Baker SC, Baric RS, de Groot RJ, Drosten C, Gulyaeva AA, et al. The species Severe acute respiratory syndrome-related coronavirus: classifying 2019-nCoV and naming it SARS-CoV-2. *Nat Microbiol* 2020;5:536–44. <https://doi.org/10.1038/s41564-020-0695-z>.
- [2] Saberian M, Karimi E, Khademi Z, Movahhed P, Safi A, Mehri-Ghahfarrokhi A. SARS-CoV-2: phenotype, genotype, and characterization of different variants. *Cell Mol Biol Lett* 2022;27. <https://doi.org/10.1186/s11658-022-00352-6>.
- [3] Jahirul Islam M, Nawal Islam N, Siddik Alom M, Kabir M, Halim MA. A review on structural, non-structural, and accessory proteins of SARS-CoV-2: Highlighting drug target sites. *Immunobiology* 2023;228. <https://doi.org/10.1016/j.imbio.2022.152302>.
- [4] Yang H, Rao Z. Structural biology of SARS-CoV-2 and implications for therapeutic development. *Nat Rev Microbiol* 2021;19:685–700. <https://doi.org/10.1038/s41579-021-00630-8>.
- [5] Huang Y, Yang C, Xu X feng, Xu W, Liu S wen. Structural and functional properties of SARS-CoV-2 spike protein: potential antiviral drug development for COVID-19. *Acta Pharmacol Sin* 2020;41:1141–9. <https://doi.org/10.1038/s41401-020-0485-4>.
- [6] Rahimi A, Mirzazadeh A, Tavakolpour S. Genetics and genomics of SARS-CoV-2: A review of the literature with the special focus on genetic diversity and SARS-CoV-2 genome detection. *Genomics* 2021;113:1221–32. <https://doi.org/10.1016/j.ygeno.2020.09.059>.
- [7] Emrani J, Ahmed M, Jeffers-Francis L, Teleha JC, Mowa N, Newman RH, et al. SARS-COV-2, infection, transmission, transcription, translation, proteins, and treatment: A review. *Int J Biol Macromol* 2021;193:1249–73. <https://doi.org/10.1016/j.ijbiomac.2021.10.172>.
- [8] Naqvi AAT, Fatima K, Mohammad T, Fatima U, Singh IK, Singh A, et al. Insights into SARS-CoV-2 genome, structure, evolution, pathogenesis and therapies: Structural genomics approach. *Biochim Biophys Acta Mol Basis Dis* 2020;1866. <https://doi.org/10.1016/j.bbadis.2020.165878>.
- [9] Malone B, Urakova N, Snijder EJ, Campbell EA. Structures and functions of coronavirus replication–transcription complexes and their relevance for SARS-CoV-2 drug design. *Nat Rev Mol Cell Biol* 2022;23:21–39. <https://doi.org/10.1038/s41580-021-00432-z>.

- [10] Jackson CB, Farzan M, Chen B, Choe H. Mechanisms of SARS-CoV-2 entry into cells. *Nat Rev Mol Cell Biol* 2022;23:3–20. <https://doi.org/10.1038/s41580-021-00418-x>.
- [11] Zhao MM, Yang WL, Yang FY, Zhang L, Huang WJ, Hou W, et al. Cathepsin L plays a key role in SARS-CoV-2 infection in humans and humanized mice and is a promising target for new drug development. *Signal Transduct Target Ther* 2021;6. <https://doi.org/10.1038/s41392-021-00558-8>.
- [12] Zhang J. Immune responses in COVID-19 patients: Insights into cytokine storms and adaptive immunity kinetics. *Heliyon* 2024;10. <https://doi.org/10.1016/j.heliyon.2024.e34577>.
- [13] Yang L, Xie X, Tu Z, Fu J, Xu D, Zhou Y. The signal pathways and treatment of cytokine storm in COVID-19. *Signal Transduct Target Ther* 2021;6. <https://doi.org/10.1038/s41392-021-00679-0>.
- [14] Jing X, Xu M, Song D, Yue T, Wang Y, Zhang P, et al. Association between inflammatory cytokines and anti-SARS-CoV-2 antibodies in hospitalized patients with COVID-19. *Immunity and Ageing* 2022;19. <https://doi.org/10.1186/s12979-022-00271-2>.
- [15] Çelik I, Öztürk R. From asymptomatic to critical illness: decoding various clinical stages of COVID-19. *Turk J Med Sci* 2021;51:3284–300. <https://doi.org/10.3906/sag-2107-137>.
- [16] Wiersinga WJ, Rhodes A, Cheng AC, Peacock SJ, Prescott HC. Pathophysiology, Transmission, Diagnosis, and Treatment of Coronavirus Disease 2019 (COVID-19): A Review. *JAMA - Journal of the American Medical Association* 2020;324:782–93. <https://doi.org/10.1001/jama.2020.12839>.
- [17] Walsh KA, Spillane S, Comber L, Cardwell K, Harrington P, Connell J, et al. The duration of infectiousness of individuals infected with SARS-CoV-2. *Journal of Infection* 2020;81:847–56. <https://doi.org/10.1016/j.jinf.2020.10.009>.
- [18] Puhach O, Meyer B, Eckerle I. SARS-CoV-2 viral load and shedding kinetics. *Nat Rev Microbiol* 2023;21:147–61. <https://doi.org/10.1038/s41579-022-00822-w>.
- [19] Dhar Chowdhury S, Oommen AM. Epidemiology of COVID-19. *Journal of Digestive Endoscopy* 2020;11:03–7. <https://doi.org/10.1055/s-0040-1712187>.

- [20] Byrne AW, McEvoy D, Collins AB, Hunt K, Casey M, Barber A, et al. Inferred duration of infectious period of SARS-CoV-2: Rapid scoping review and analysis of available evidence for asymptomatic and symptomatic COVID-19 cases. *BMJ Open* 2020;10. <https://doi.org/10.1136/bmjopen-2020-039856>.
- [21] Bajaj V, Gadi N, Spihlman AP, Wu SC, Choi CH, Moulton VR. Aging, Immunity, and COVID-19: How Age Influences the Host Immune Response to Coronavirus Infections? *Front Physiol* 2021;11. <https://doi.org/10.3389/fphys.2020.571416>.
- [22] Davies NG, Klepac P, Liu Y, Prem K, Jit M, Pearson CAB, et al. Age-dependent effects in the transmission and control of COVID-19 epidemics. *Nat Med* 2020;26:1205–11. <https://doi.org/10.1038/s41591-020-0962-9>.
- [23] Hosseini A, Hashemi V, Shomali N, Asghari F, Gharibi T, Akbari M, et al. Innate and adaptive immune responses against coronavirus. *Biomedicine and Pharmacotherapy* 2020;132. <https://doi.org/10.1016/j.biopha.2020.110859>.
- [24] Zhu Q, Xu Y, Wang T, Xie F. Innate and adaptive immune response in SARS-CoV-2 infection-Current perspectives. *Front Immunol* 2022;13. <https://doi.org/10.3389/fimmu.2022.1053437>.
- [25] Merad M, Martin JC. Pathological inflammation in patients with COVID-19: a key role for monocytes and macrophages. *Nat Rev Immunol* 2020;20:355–62. <https://doi.org/10.1038/s41577-020-0331-4>.
- [26] Wang X, Bai H, Ma J, Qin H, Zeng Q, Hu F, et al. Identification of Distinct Immune Cell Subsets Associated With Asymptomatic Infection, Disease Severity, and Viral Persistence in COVID-19 Patients. *Front Immunol* 2022;13. <https://doi.org/10.3389/fimmu.2022.812514>.
- [27] Iyer A, Hamers AAJ, Pillai AB. CyTOF® for the Masses. *Front Immunol* 2022;13. <https://doi.org/10.3389/fimmu.2022.815828>.
- [28] Cheng ZJ, Huang H, Liu Q, Zhong R, Liang Z, Xue M, et al. Immunoassay and mass cytometry revealed immunological profiles induced by inactivated BBIBP COVID-19 vaccine. *J Med Virol* 2022;94:5206–16. <https://doi.org/10.1002/jmv.27983>.
- [29] Zhou D, Zhao S, He K, Liu Q, Zhang F, Pu Z, et al. Longitudinal dynamic single-cell mass cytometry analysis of peripheral blood mononuclear cells in COVID-19 patients within 6 months after viral

- RNA clearance. *BMC Infect Dis* 2024;24.
<https://doi.org/10.1186/s12879-024-09464-0>.
- [30] Geanon D, Lee B, Gonzalez-Kozlova E, Kelly G, Handler D, Upadhyaya B, et al. A streamlined whole blood CyTOF workflow defines a circulating immune cell signature of COVID-19. *Cytometry Part A* 2021;99:446–61. <https://doi.org/10.1002/cyto.a.24317>.
- [31] Burnett CE, Okholm TLH, TenVooren I, Marquez DM, Tamaki S, Munoz Sandoval P, et al. Mass cytometry reveals a conserved immune trajectory of recovery in hospitalized COVID-19 patients. *Immunity* 2022;55:1284–1298.e3. <https://doi.org/10.1016/j.immuni.2022.06.004>.
- [32] Lapuente D, Winkler TH, Tenbusch M. B-cell and antibody responses to SARS-CoV-2: infection, vaccination, and hybrid immunity. *Cell Mol Immunol* 2024;21:144–58. <https://doi.org/10.1038/s41423-023-01095-w>.
- [33] Cyster JG, Allen CDC. B Cell Responses: Cell Interaction Dynamics and Decisions. *Cell* 2019;177:524–40.
<https://doi.org/10.1016/j.cell.2019.03.016>.
- [34] Röltgen K, Boyd SD. Antibody and B cell responses to SARS-CoV-2 infection and vaccination. *Cell Host Microbe* 2021;29:1063–75.
<https://doi.org/10.1016/j.chom.2021.06.009>.
- [35] Sun L, Su Y, Jiao A, Wang X, Zhang B. T cells in health and disease. *Signal Transduct Target Ther* 2023;8. <https://doi.org/10.1038/s41392-023-01471-y>.
- [36] Molina-París C, Lythe G. Mathematical, computational and experimental T cell immunology. Springer International Publishing; 2021. <https://doi.org/10.1007/978-3-030-57204-4>.
- [37] Moss P. The T cell immune response against SARS-CoV-2. *Nat Immunol* 2022;23:186–93. <https://doi.org/10.1038/s41590-021-01122-w>.
- [38] Poon MM, Rybkina K, Kato Y, Kubota M, Matsumoto R, Bloom NI, et al. SARS-CoV-2 infection generates tissue-localized immunological memory in humans. vol. 6. 2021.
- [39] Sette A, Sidney J, Crotty S. Annual Review of Immunology T Cell Responses to SARS-CoV-2 2023. <https://doi.org/10.1146/annurev-immunol-101721>.
- [40] Grifoni A, Weiskopf D, Ramirez SI, Mateus J, Dan JM, Moderbacher CR, et al. Targets of T Cell Responses to SARS-CoV-2 Coronavirus in

- Humans with COVID-19 Disease and Unexposed Individuals. *Cell* 2020;181:1489-1501.e15. <https://doi.org/10.1016/j.cell.2020.05.015>.
- [41] Keeton R, Tincho MB, Ngomti A, Baguma R, Benede N, Suzuki A, et al. T cell responses to SARS-CoV-2 spike cross-recognize Omicron. *Nature* 2022;603:488–92. <https://doi.org/10.1038/s41586-022-04460-3>.
- [42] Sette A, Sidney J, Crotty S. Annual Review of Immunology T Cell Responses to SARS-CoV-2 2023. <https://doi.org/10.1146/annurev-immunol-101721>.
- [43] Grifoni A, Sidney J, Zhang Y, Scheuermann RH, Peters B, Sette A. A Sequence Homology and Bioinformatic Approach Can Predict Candidate Targets for Immune Responses to SARS-CoV-2. *Cell Host Microbe* 2020;27:671-680.e2. <https://doi.org/10.1016/j.chom.2020.03.002>.
- [44] Molina-París C, Lythe G. Mathematical, computational and experimental T cell immunology. Springer International Publishing; 2021. <https://doi.org/10.1007/978-3-030-57204-4>.
- [45] Chung KJ, Nati M, Chavakis T, Chatzigeorgiou A. Innate immune cells in the adipose tissue. *Rev Endocr Metab Disord* 2018;19:283–92. <https://doi.org/10.1007/s11154-018-9451-6>.
- [46] Almendro-Vázquez P, Laguna-Goya R, Paz-Artal E. Defending against SARS-CoV-2: The T cell perspective. *Front Immunol* 2023;14. <https://doi.org/10.3389/fimmu.2023.1107803>.
- [47] Flanagan KL, Best E, Crawford NW, Giles M, Koirala A, Macartney K, et al. Progress and Pitfalls in the Quest for Effective SARS-CoV-2 (COVID-19) Vaccines. *Front Immunol* 2020;11. <https://doi.org/10.3389/fimmu.2020.579250>.
- [48] Crouse J, Kalinke U, Oxenius A. Regulation of antiviral T cell responses by type I interferons. *Nat Rev Immunol* 2015;15:231–42. <https://doi.org/10.1038/nri3806>.
- [49] Bergamaschi L, Mescia F, Turner L, Hanson AL, Kotagiri P, Dunmore BJ, et al. Longitudinal analysis reveals that delayed bystander CD8+ T cell activation and early immune pathology distinguish severe COVID-19 from mild disease. *Immunity* 2021;54:1257-1275.e8. <https://doi.org/10.1016/j.immuni.2021.05.010>.
- [50] Notarbartolo S, Ranzani V, Bandera A, Gruarin P, Bevilacqua V, Rita Putignano A, et al. Integrated longitudinal immunophenotypic,

- transcriptional, and repertoire analyses delineate immune responses in patients with COVID-19. vol. 6. 2021.
- [51] Braun J, Loyal L, Frentsch M, Wendisch D, Georg P, Kurth F, et al. SARS-CoV-2-reactive T cells in healthy donors and patients with COVID-19. *Nature* 2020;587:270–4. <https://doi.org/10.1038/s41586-020-2598-9>.
- [52] Rydyznski Moderbacher C, Ramirez SI, Dan JM, Grifoni A, Hastie KM, Weiskopf D, et al. Antigen-Specific Adaptive Immunity to SARS-CoV-2 in Acute COVID-19 and Associations with Age and Disease Severity. *Cell* 2020;183:996–1012.e19. <https://doi.org/10.1016/j.cell.2020.09.038>.
- [53] Notarbartolo S, Ranzani V, Bandera A, Gruarin P, Bevilacqua V, Rita Putignano A, et al. Integrated longitudinal immunophenotypic, transcriptional, and repertoire analyses delineate immune responses in patients with COVID-19. vol. 6. 2021.
- [54] Peng Y, Felce SL, Dong D, Penkava F, Mentzer AJ, Yao X, et al. An immunodominant NP105–113-B*07:02 cytotoxic T cell response controls viral replication and is associated with less severe COVID-19 disease. *Nat Immunol* 2022;23:50–61. <https://doi.org/10.1038/s41590-021-01084-z>.
- [55] Mahmoud Salehi Khesht A, Karpisheh V, Qubais Saeed B, Olegovna Zekiy A, Yapanto LM, Nabi Afjadi M, et al. Different T cell related immunological profiles in COVID-19 patients compared to healthy controls. *Int Immunopharmacol* 2021;97. <https://doi.org/10.1016/j.intimp.2021.107828>.
- [56] Pavel AB, Glickman JW, Michels JR, Kim-Schulze S, Miller RL, Guttman-Yassky E. Th2/Th1 Cytokine Imbalance Is Associated With Higher COVID-19 Risk Mortality. *Front Genet* 2021;12. <https://doi.org/10.3389/fgene.2021.706902>.
- [57] Stephenson E, Reynolds G, Botting RA, Calero-Nieto FJ, Morgan MD, Tuong ZK, et al. Single-cell multi-omics analysis of the immune response in COVID-19. *Nat Med* 2021;27:904–16. <https://doi.org/10.1038/s41591-021-01329-2>.
- [58] Gutierrez L, Beckford J, Alachkar H. Deciphering the TCR Repertoire to Solve the COVID-19 Mystery. *Trends Pharmacol Sci* 2020;41:518–30. <https://doi.org/10.1016/j.tips.2020.06.001>.

- [59] Pinti M, Appay V, Campisi J, Frasca D, Fülöp T, Sauce D, et al. Aging of the immune system: Focus on inflammation and vaccination. *Eur J Immunol* 2016;46:2286–301. <https://doi.org/10.1002/eji.201546178>.
- [60] De Cecco M, Ito T, Petrashen AP, Elias AE, Skvir NJ, Criscione SW, et al. L1 drives IFN in senescent cells and promotes age-associated inflammation. *Nature* 2019;566:73–8. <https://doi.org/10.1038/s41586-018-0784-9>.
- [61] Thomas C, Tampé R. Structure and mechanism of immunoreceptors: New horizons in T cell and B cell receptor biology and beyond. *Curr Opin Struct Biol* 2023;80. <https://doi.org/10.1016/j.sbi.2023.102570>.
- [62] Alcover A, Alarcón B, Alarcón A, Bartolo V Di. Cell Biology of T Cell Receptor Expression and Regulation. *Annu Rev Immunol* 2018;12:40. <https://doi.org/10.1146/annurev-immunol>.
- [63] Gutierrez L, Beckford J, Alachkar H. Deciphering the TCR Repertoire to Solve the COVID-19 Mystery. *Trends Pharmacol Sci* 2020;41:518–30. <https://doi.org/10.1016/j.tips.2020.06.001>.
- [64] Nikolich-Zugich J, Slifka MK, Messaoudi I. The many important facets of T-cell repertoire diversity. *Nat Rev Immunol* 2004;4:123–32. <https://doi.org/10.1038/nri1292>.
- [65] Attaf M, Huseby E, Sewell AK. $\alpha\beta$ T cell receptors as predictors of health and disease. *Cell Mol Immunol* 2015;12:391–9. <https://doi.org/10.1038/cmi.2014.134>.
- [66] Tanno H, Gould TM, Mcdaniel JR, Cao W, Tanno Y, Durrett RE, et al. Determinants governing T cell receptor α/β -chain pairing in repertoire formation of identical twins n.d. <https://doi.org/10.1073/pnas.1915008117/-/DCSupplemental>.
- [67] Lawand M, Déchanet-Merville J, Dieu-Nosjean MC. Key features of gamma-delta T-cell subsets in human diseases and their immunotherapeutic implications. *Front Immunol* 2017;8. <https://doi.org/10.3389/fimmu.2017.00761>.
- [68] Morath A, Schamel WW. $\alpha\beta$ and $\gamma\delta$ T cell receptors: Similar but different. *J Leukoc Biol* 2020;107:1045–55. <https://doi.org/10.1002/JLB.2MR1219-233R>.
- [69] Shah K, Al-Haidari A, Sun J, Kazi JU. T cell receptor (TCR) signaling in health and disease. *Signal Transduct Target Ther* 2021;6. <https://doi.org/10.1038/s41392-021-00823-w>.

- [70] Vantourout P, Hayday A. Six-of-the-best: Unique contributions of $\gamma\delta$ T cells to immunology. *Nat Rev Immunol* 2013;13:88–100. <https://doi.org/10.1038/nri3384>.
- [71] Paul S, Shilpi, Lal G. Role of gamma-delta ($\gamma\delta$) T cells in autoimmunity . *J Leukoc Biol* 2015;97:259–71. <https://doi.org/10.1189/jlb.3ru0914-443r>.
- [72] Mariuzza RA, Agnihotri P, Orban J. The structural basis of T-cell receptor (TCR) activation: An enduring enigma. *Journal of Biological Chemistry* 2020;295:914–25. <https://doi.org/10.1074/jbc.REV119.009411>.
- [73] Xu X, Li H, Xu C. Structural understanding of T cell receptor triggering. *Cell Mol Immunol* 2020;17:193–202. <https://doi.org/10.1038/s41423-020-0367-1>.
- [74] Dong D, Zheng L, Lin J, Zhang B, Zhu Y, Li N, et al. Structural basis of assembly of the human T cell receptor–CD3 complex. *Nature* 2019;573:546–52. <https://doi.org/10.1038/s41586-019-1537-0>.
- [75] Chi H, Pepper M, Thomas PG. Principles and therapeutic applications of adaptive immunity. *Cell* 2024;187:2052–78. <https://doi.org/10.1016/j.cell.2024.03.037>.
- [76] Schatz DG, Swanson PC. V(D)J recombination: Mechanisms of initiation. *Annu Rev Genet* 2011;45:167–202. <https://doi.org/10.1146/annurev-genet-110410-132552>.
- [77] Little AJ, Matthews A, Oettinger M, Roth DB, Schatz DG. The Mechanism of V(D)J Recombination. *Molecular Biology of B Cells: Second Edition*, Elsevier Inc.; 2015, p. 13–34. <https://doi.org/10.1016/B978-0-12-397933-9.00002-3>.
- [78] Roth DB. V(D)J Recombination: Mechanism, Errors, and Fidelity. *Microbiol Spectr* 2014;2. <https://doi.org/10.1128/microbiolspec.mdna3-0041-2014>.
- [79] Iyer A, Hennessey D, Gniadecki R. Clonotype pattern in T-cell lymphomas map the cell of origin to immature lymphoid precursors. *Blood Adv* 2022;6:2334–45. <https://doi.org/10.1182/bloodadvances.2021005884>.
- [80] Schatz DG, Swanson PC. V(D)J recombination: Mechanisms of initiation. *Annu Rev Genet* 2011;45:167–202. <https://doi.org/10.1146/annurev-genet-110410-132552>.

- [81] Weng N ping. Numbers and odds: TCR repertoire size and its age changes impacting on T cell functions. *Semin Immunol* 2023;69. <https://doi.org/10.1016/j.smim.2023.101810>.
- [82] Ashby KM, Hogquist KA. A guide to thymic selection of T cells. *Nat Rev Immunol* 2024;24:103–17. <https://doi.org/10.1038/s41577-023-00911-8>.
- [83] Bensouda Koraichi M, Ferri S, Walczak AM, Mora T. Inferring the T cell repertoire dynamics of healthy individuals 2023. <https://doi.org/10.1073/pnas>.
- [84] Yassai MB, Naumov YN, Naumova EN, Gorski J. A clonotype nomenclature for T cell receptors. *Immunogenetics* 2009;61:493–502. <https://doi.org/10.1007/s00251-009-0383-x>.
- [85] Sofou E, Vlachonikola E, Zaragoza-Infante L, Brüggemann M, Darzentas N, Groenen PJTA, et al. Clonotype definitions for immunogenetic studies: proposals from the EuroClonality NGS Working Group. *Leukemia* 2023;37:1750–2. <https://doi.org/10.1038/s41375-023-01952-7>.
- [86] Mandric I, Rotman J, Yang HT, Strauli N, Montoya DJ, Van Der Wey W, et al. Profiling immunoglobulin repertoires across multiple human tissues using RNA sequencing. *Nat Commun* 2020;11. <https://doi.org/10.1038/s41467-020-16857-7>.
- [87] Valkiers S, de Vrij N, Gielis S, Verbandt S, Ogunjimi B, Laukens K, et al. Recent advances in T-cell receptor repertoire analysis: Bridging the gap with multimodal single-cell RNA sequencing. *ImmunoInformatics* 2022;5:100009. <https://doi.org/10.1016/j.immuno.2022.100009>.
- [88] Chi H, Pepper M, Thomas PG. Principles and therapeutic applications of adaptive immunity. *Cell* 2024;187:2052–78. <https://doi.org/10.1016/j.cell.2024.03.037>.
- [89] de Greef PC, Oakes T, Gerritsen B, Ismail M, Heather JM, Hermesen R, et al. The naive t-cell receptor repertoire has an extremely broad distribution of clone sizes. *Elife* 2020;9. <https://doi.org/10.7554/eLife.49900>.
- [90] Attaf M, Huseby E, Sewell AK. $\alpha\beta$ T cell receptors as predictors of health and disease. *Cell Mol Immunol* 2015;12:391–9. <https://doi.org/10.1038/cmi.2014.134>.

- [91] Krangel MS. Mechanics of T cell receptor gene rearrangement. *Curr Opin Immunol* 2009;21:133–9. <https://doi.org/10.1016/j.coi.2009.03.009>.
- [92] Hosokawa H, Rothenberg E V. How transcription factors drive choice of the T cell fate. *Nat Rev Immunol* 2021;21:162–76. <https://doi.org/10.1038/s41577-020-00426-6>.
- [93] Quigley MF, Greenaway HY, Venturi V, Lindsay R, Quinn KM, Seder RA, et al. Convergent recombination shapes the clonotypic landscape of the naïve T-cell repertoire. *Proc Natl Acad Sci U S A* 2010;107:19414–9. <https://doi.org/10.1073/pnas.1010586107>.
- [94] Klein L, Kyewski B, Allen PM, Hogquist KA. Positive and negative selection of the T cell repertoire: What thymocytes see (and don't see). *Nat Rev Immunol* 2014;14:377–91. <https://doi.org/10.1038/nri3667>.
- [95] Basu J, Reis BS, Peri S, Zha J, Hua X, Ge L, et al. Essential role of a ThPOK autoregulatory loop in the maintenance of mature CD4⁺ T cell identity and function. *Nat Immunol* 2021;22:969–82. <https://doi.org/10.1038/s41590-021-00980-8>.
- [96] Luckey MA, Kimura MY, Waickman AT, Feigenbaum L, Singer A, Park JH. The transcription factor ThPOK suppresses Runx3 and imposes CD4⁺ lineage fate by inducing the SOCS suppressors of cytokine signaling. *Nat Immunol* 2014;15:638–45. <https://doi.org/10.1038/ni.2917>.
- [97] Vrisekoop N, Monteiro JP, Mandl JN, Germain RN. Revisiting thymic positive selection and the mature T cell repertoire for antigen. *Immunity* 2014;41:181–90. <https://doi.org/10.1016/j.immuni.2014.07.007>.
- [98] Takaba H, Takayanagi H. The Mechanisms of T Cell Selection in the Thymus. *Trends Immunol* 2017;38:805–16. <https://doi.org/10.1016/j.it.2017.07.010>.
- [99] Xing Y, Wang X, Jameson SC, Hogquist KA. Late stages of T cell maturation in the thymus involve NF- κ B and tonic type I interferon signaling. *Nat Immunol* 2016;17:565–73. <https://doi.org/10.1038/ni.3419>.
- [100] Adams NM, Grassmann S, Sun JC. Clonal expansion of innate and adaptive lymphocytes. *Nat Rev Immunol* 2020;20:694–707. <https://doi.org/10.1038/s41577-020-0307-4>.
- [101] Kretschmer L, Flossdorf M, Mir J, Cho YL, Plambeck M, Treise I, et al. Differential expansion of T central memory precursor and effector

- subsets is regulated by division speed. *Nat Commun* 2020;11.
<https://doi.org/10.1038/s41467-019-13788-w>.
- [102] Zhu X, Ma E, Ning K, Feng X, Quan W, Wang F, et al. A comparative analysis of TCR immune repertoire in COVID-19 patients. *Hum Immunol* 2024;85. <https://doi.org/10.1016/j.humimm.2024.110795>.
- [103] Luo L, Liang W, Pang J, Xu G, Chen Y, Guo X, et al. Dynamics of TCR repertoire and T cell function in COVID-19 convalescent individuals. *Cell Discov* 2021;7. <https://doi.org/10.1038/s41421-021-00321-x>.
- [104] Zhang F, Gan R, Zhen Z, Hu X, Li X, Zhou F, et al. Adaptive immune responses to SARS-CoV-2 infection in severe versus mild individuals. *Signal Transduct Target Ther* 2020;5. <https://doi.org/10.1038/s41392-020-00263-y>.
- [105] Chatanaka MK, Uldreaj A, Sohaei D, Prassas I. Immunoinformatics: Pushing the boundaries of immunology research and medicine. *ImmunoInformatics* 2022;5:100007.
<https://doi.org/10.1016/j.immuno.2021.100007>.
- [106] De RK, Tomar N. *Immuno-informatics Second Edition Methods in Molecular Biology* 1184. n.d.
- [107] Oli AN, Obialor WO, Ositadimma M, Ifeanyichukwu, Odimegwu DC, Okoyeh JN, et al. Immunoinformatics and vaccine development: An overview. *Immunotargets Ther* 2020;9:13–30.
<https://doi.org/10.2147/ITT.S241064>.
- [108] Basmenj ER, Pajhouh SR, Ebrahimi Fallah A, naijian R, Rahimi E, Atighy H, et al. Computational epitope-based vaccine design with bioinformatics approach; a review. *Heliyon* 2025;11.
<https://doi.org/10.1016/j.heliyon.2025.e41714>.
- [109] Khorsand B, Savadi A, Naghibzadeh M. Comprehensive host-pathogen protein-protein interaction network analysis. *BMC Bioinformatics* 2020;21. <https://doi.org/10.1186/s12859-020-03706-z>.
- [110] Ewald J, Sieber P, Garde R, Lang SN, Schuster S, Ibrahim B. Trends in mathematical modeling of host–pathogen interactions. *Cellular and Molecular Life Sciences* 2020;77:467–80.
<https://doi.org/10.1007/s00018-019-03382-0>.
- [111] Charoentong P, Angelova M, Efremova M, Gallasch R, Hackl H, Galon J, et al. Bioinformatics for cancer immunology and immunotherapy. *Cancer Immunology, Immunotherapy* 2012;61:1885–903.
<https://doi.org/10.1007/s00262-012-1354-x>.

- [112] Sidhom J-W, Oliveira G, Ross-Macdonald P, Wind-Rotolo M, Wu CJ, Pardoll DM, et al. Deep learning reveals predictive sequence concepts within immune repertoires to immunotherapy. vol. 8. 2022.
- [113] Frank ML, Lu K, Erdogan C, Han Y, Hu J, Wang T, et al. T-Cell Receptor Repertoire Sequencing in the Era of Cancer Immunotherapy. *Clinical Cancer Research* 2023;29:994–1008. <https://doi.org/10.1158/1078-0432.CCR-22-2469>.
- [114] Doxey AC, Mansfield MJ, Lobb B. Exploring the Evolution of Virulence Factors through Bioinformatic Data Mining Exploring virulence factor evolution and diversity using bioinformatics: A perspective on our work mining genomic databases for homologs, adaptations, and new virulence factor families. 2019.
- [115] de Nies L, Lopes S, Busi SB, Galata V, Heintz-Buschart A, Laczny CC, et al. PathoFact: a pipeline for the prediction of virulence factors and antimicrobial resistance genes in metagenomic data. *Microbiome* 2021;9. <https://doi.org/10.1186/s40168-020-00993-9>.
- [116] Thorball CW, Fellay J, Borghesi A. Immunological lessons from genome-wide association studies of infections. *Curr Opin Immunol* 2021;72:87–93. <https://doi.org/10.1016/j.coi.2021.03.017>.
- [117] Amadori L, Calcagno C, Fernandez DM, Koplev S, Fernandez N, Kaur R, et al. Systems immunology-based drug repurposing framework to target inflammation in atherosclerosis. *Nature Cardiovascular Research* 2023;2:550–71. <https://doi.org/10.1038/s44161-023-00278-y>.
- [118] Curion F, Theis FJ. Machine learning integrative approaches to advance computational immunology. *Genome Med* 2024;16. <https://doi.org/10.1186/s13073-024-01350-3>.
- [119] Pertseva M, Gao B, Neumeier D, Yermanos A, Reddy ST. Annual Review of Chemical and Biomolecular Engineering Applications of Machine and Deep Learning in Adaptive Immunity 2025;17:53. <https://doi.org/10.1146/annurev-chembioeng>.
- [120] Zaslavsky ME, Craig E, Michuda JK, Sehgal N, Ram-Mohan N, Lee J-Y, et al. Disease diagnostics using machine learning of B cell and T cell receptor sequences. *Science* (1979) 2025;387. <https://doi.org/10.1126/science.adp2407>.
- [121] Seo K, Choi JK. Comprehensive Analysis of TCR and BCR Repertoires: Insights into Methodologies, Challenges, and Applications. *Genomics Inform* 2025;23. <https://doi.org/10.1186/s44342-024-00034-z>.

- [122] Valkiers S, de Vrij N, Gielis S, Verbandt S, Ogunjimi B, Laukens K, et al. Recent advances in T-cell receptor repertoire analysis: Bridging the gap with multimodal single-cell RNA sequencing. *ImmunoInformatics* 2022;5:100009. <https://doi.org/10.1016/j.immuno.2022.100009>.
- [123] Redmond D, Poran A, Elemento O. Single-cell TCRseq: Paired recovery of entire T-cell alpha and beta chain transcripts in T-cell receptors from single-cell RNAseq. *Genome Med* 2016;8:1. <https://doi.org/10.1186/s13073-016-0335-7>.
- [124] Pai JA, Satpathy AT. High-throughput and single-cell T cell receptor sequencing technologies. *Nat Methods* 2021;18:881–92. <https://doi.org/10.1038/s41592-021-01201-8>.
- [125] Tian R, Yu Z, Xue Z, Wu J, Wu L, Cai S, et al. Evaluation of T Cell Receptor Construction Methods from scRNA-Seq Data. *Genomics Proteomics Bioinformatics* 2025;22. <https://doi.org/10.1093/gpbjnl/qzae086>.
- [126] Mahdy AKH, Lokes E, Schöpfel V, Kriukova V, Britanova O V., Steiert TA, et al. Bulk T cell repertoire sequencing (TCR-Seq) is a powerful technology for understanding inflammation-mediated diseases. *J Autoimmun* 2024;149. <https://doi.org/10.1016/j.jaut.2024.103337>.
- [127] Xu J, Li XX, Yuan N, Li C, Yang JG, Cheng LM, et al. T cell receptor β repertoires in patients with COVID-19 reveal disease severity signatures. *Front Immunol* 2023;14. <https://doi.org/10.3389/fimmu.2023.1190844>.
- [128] Kockelbergh H, Evans S, Deng T, Clyne E, Kyriakidou A, Economou A, et al. Utility of Bulk T-Cell Receptor Repertoire Sequencing Analysis in Understanding Immune Responses to COVID-19. *Diagnostics* 2022;12. <https://doi.org/10.3390/diagnostics12051222>.
- [129] Kockelbergh H, Evans S, Deng T, Clyne E, Kyriakidou A, Economou A, et al. Utility of Bulk T-Cell Receptor Repertoire Sequencing Analysis in Understanding Immune Responses to COVID-19. *Diagnostics* 2022;12. <https://doi.org/10.3390/diagnostics12051222>.
- [130] Pogorelyy M V., Fedorova AD, McLaren JE, Ladell K, Bagaev D V., Eliseev A V., et al. Exploring the pre-immune landscape of antigen-specific T cells. *Genome Med* 2018;10. <https://doi.org/10.1186/s13073-018-0577-7>.
- [131] Springer I, Besser H, Tickotsky-Moskovitz N, Dvorkin S, Louzoun Y. Prediction of Specific TCR-Peptide Binding From Large Dictionaries of

- TCR-Peptide Pairs. *Front Immunol* 2020;11.
<https://doi.org/10.3389/fimmu.2020.01803>.
- [132] Rosati E, Dowds CM, Liaskou E, Henriksen EKK, Karlsen TH, Franke A. Overview of methodologies for T-cell receptor repertoire analysis. *BMC Biotechnol* 2017;17. <https://doi.org/10.1186/s12896-017-0379-9>.
- [133] Genolet R, Bobisse S, Chiffelle J, Arnaud M, Petremand R, Queiroz L, et al. TCR sequencing and cloning methods for repertoire analysis and isolation of tumor-reactive TCRs. *Cell Reports Methods* 2023;3.
<https://doi.org/10.1016/j.crmeth.2023.100459>.
- [134] Bolotin DA, Poslavsky S, Mitrophanov I, Shugay M, Mamedov IZ, Putintseva E V, et al. MiXCR: software for comprehensive adaptive immunity profiling. *Nat Methods* 2015;12:380–1.
<https://doi.org/10.1038/nmeth.3364>.
- [135] Smakaj E, Babrak L, Ohlin M, Shugay M, Briney B, Tosoni D, et al. Benchmarking immunoinformatic tools for the analysis of antibody repertoire sequences. *Bioinformatics* 2020;36:1731–9.
<https://doi.org/10.1093/bioinformatics/btz845>.
- [136] Marbac M, Sedki M. VarSelLCM: an R/C++ package for variable selection in model-based clustering of mixed-data with missing values. *Bioinformatics* 2019;35:1255–7.
<https://doi.org/10.1093/bioinformatics/bty786>.
- [137] Chiffelle J, Genolet R, Perez MA, Coukos G, Zoete V, Harari A. T-cell repertoire analysis and metrics of diversity and clonality. *Curr Opin Biotechnol* 2020;65:284–95.
<https://doi.org/10.1016/j.copbio.2020.07.010>.
- [138] Kockelbergh H, Evans S, Deng T, Clyne E, Kyriakidou A, Economou A, et al. Utility of Bulk T-Cell Receptor Repertoire Sequencing Analysis in Understanding Immune Responses to COVID-19. *Diagnostics* 2022;12.
<https://doi.org/10.3390/diagnostics12051222>.
- [139] Hudson D, Lubbock A, Basham M, Koohy H. A comparison of clustering models for inference of T cell receptor antigen specificity. *ImmunoInformatics* 2024;13:100033.
<https://doi.org/10.1016/j.immuno.2024.100033>.
- [140] Feng X, Huo M, Li H, Yang Y, Jiang Y, He L, et al. A comprehensive benchmarking for evaluating TCR embeddings in modeling TCR-epitope interactions. *Brief Bioinform* 2024;26.
<https://doi.org/10.1093/bib/bbaf030>.

- [141] Pertseva M, Follonier O, Scarcella D, Reddy ST. TCR clustering by contrastive learning on antigen specificity. *Brief Bioinform* 2024;25. <https://doi.org/10.1093/bib/bbae375>.
- [142] Vujovic M, Degn KF, Marin FI, Schaap-Johansen AL, Chain B, Andresen TL, et al. T cell receptor sequence clustering and antigen specificity. *Comput Struct Biotechnol J* 2020;18:2166–73. <https://doi.org/10.1016/j.csbj.2020.06.041>.
- [143] Huang H, Wang C, Rubelt F, Scriba TJ, Davis MM. Analyzing the *Mycobacterium tuberculosis* immune response by T-cell receptor clustering with GLIPH2 and genome-wide antigen screening. *Nat Biotechnol* 2020;38:1194–202. <https://doi.org/10.1038/s41587-020-0505-4>.
- [144] Glanville J, Huang H, Nau A, Hatton O, Wagar LE, Rubelt F, et al. Identifying specificity groups in the T cell receptor repertoire. *Nature* 2017;547:94–8. <https://doi.org/10.1038/nature22976>.
- [145] Zhang H, Zhan X, Li B. GIANA allows computationally-efficient TCR clustering and multi-disease repertoire classification by isometric transformation. *Nat Commun* 2021;12. <https://doi.org/10.1038/s41467-021-25006-7>.
- [146] Zhang H, Liu L, Zhang J, Chen J, Ye J, Shukla S, et al. Investigation of antigen-specific T-cell receptor clusters in human cancers. *Clinical Cancer Research* 2020;26:1359–71. <https://doi.org/10.1158/1078-0432.CCR-19-3249>.
- [147] Seamann A, Bennett-Boehm M, Ehrlich R, Gil A, Selin L, Ghersi D. TRain: T-cell receptor automated immunoinformatics. *BMC Bioinformatics* 2025;26:76. <https://doi.org/10.1186/s12859-025-06074-8>.
- [148] Sidhom JW, Larman HB, Pardoll DM, Baras AS. DeepTCR is a deep learning framework for revealing sequence concepts within T-cell repertoires. *Nat Commun* 2021;12. <https://doi.org/10.1038/s41467-021-21879-w>.
- [149] Croce G, Bobisse S, Moreno DL, Schmidt J, Guillame P, Harari A, et al. Deep learning predictions of TCR-epitope interactions reveal epitope-specific chains in dual alpha T cells. *Nat Commun* 2024;15. <https://doi.org/10.1038/s41467-024-47461-8>.
- [150] Pavlović M, Scheffer L, Motwani K, Kanduri C, Kompova R, Vazov N, et al. The immuneML ecosystem for machine learning analysis of

- adaptive immune receptor repertoires. *Nat Mach Intell* 2021;3:936–44.
<https://doi.org/10.1038/s42256-021-00413-z>.
- [151] Napolitani G, Kurupati P, Teng KWW, Gibani MM, Rei M, Aulicino A, et al. Clonal analysis of Salmonella-specific effector T cells reveals serovar-specific and cross-reactive T cell responses. *Nat Immunol* 2018;19:742–54. <https://doi.org/10.1038/s41590-018-0133-z>.
- [152] Wang Y, Li R, Tong R, Chen T, Sun M, Luo L, et al. Integrating single-cell RNA and T cell/B cell receptor sequencing with mass cytometry reveals dynamic trajectories of human peripheral immune cells from birth to old age. *Nat Immunol* 2025;26:308–22.
<https://doi.org/10.1038/s41590-024-02059-6>.
- [153] Chu Y, Milner J, Lamb M, Maryamchik E, Rigot O, Ayello J, et al. Manufacture and Characterization of Good Manufacturing Practice-Compliant SARS-COV-2 Cytotoxic T Lymphocytes. *Journal of Infectious Diseases* 2023;227:788–99.
<https://doi.org/10.1093/infdis/jiac500>.
- [154] clinical-management-of-novel-cov n.d.
- [155] Shugay M, Bagaev D V., Turchaninova MA, Bolotin DA, Britanova O V., Putintseva E V., et al. VDJtools: Unifying Post-analysis of T Cell Receptor Repertoires. *PLoS Comput Biol* 2015;11.
<https://doi.org/10.1371/journal.pcbi.1004503>.
- [156] Pagliuca S, Gurnari C, Hong S, Zhao R, Kongkiatkamon S, Terkawi L, et al. Clinical and basic implications of dynamic T cell receptor clonotyping in hematopoietic cell transplantation. *JCI Insight* 2021;6.
<https://doi.org/10.1172/jci.insight.149080>.
- [157] Bioinformatics Analysis of T-Cell and B-Cell Immune Repertoires • immunarch n.d. <https://immunarch.com/> (accessed October 17, 2023).
- [158] Chiffelle J, Genolet R, Perez MA, Coukos G, Zoete V, Harari A. T-cell repertoire analysis and metrics of diversity and clonality. *Curr Opin Biotechnol* 2020;65:284–95.
<https://doi.org/10.1016/j.copbio.2020.07.010>.
- [159] Chaudhary N, Wesemann DR. Analyzing immunoglobulin repertoires. *Front Immunol* 2018;9. <https://doi.org/10.3389/fimmu.2018.00462>.
- [160] Chiffelle J, Genolet R, Perez MA, Coukos G, Zoete V, Harari A. T-cell repertoire analysis and metrics of diversity and clonality. *Curr Opin Biotechnol* 2020;65:284–95.
<https://doi.org/10.1016/j.copbio.2020.07.010>.

- [161] turboGlyph: turboGlyph: Grouping of Lymphocyte Interactions by Paratope... in HetzDra/turboGlyph: Find Specificity Groups with GLIPH and GLIPH2 Method n.d.
<https://rdrr.io/github/HetzDra/turboGlyph/man/turboGlyph.html> (accessed October 17, 2023).
- [162] Nolan S, Vignali M, Klinger M, Dines JN, Kaplan IM, Svejnoha E, et al. A large-scale database of T-cell receptor beta (TCR β) sequences and binding associations from natural and synthetic exposure to SARS-CoV-2. *Res Sq* 2020. <https://doi.org/10.21203/rs.3.rs-51964/v1>.
- [163] Crooks GE, Hon G, Chandonia JM, Brenner SE. WebLogo: A sequence logo generator. *Genome Res* 2004;14:1188–90.
<https://doi.org/10.1101/gr.849004>.
- [164] Yanagihara H, Sekiguchi R, Fujikoshi Y. Bias correction of AIC in logistic regression models. vol. 115. 2003.
- [165] Ben-Shachar M, Lüdtke D, Makowski D. effectsize: Estimation of Effect Size Indices and Standardized Parameters. *J Open Source Softw* 2020;5:2815. <https://doi.org/10.21105/joss.02815>.
- [166] MASS: Support Functions and Datasets for Venables and Ripley's MASS n.d.
- [167] Martínez-Diz S, Marín-Benesiu F, López-Torres G, Santiago O, Díaz-Cuellar JF, Martín-Esteban S, et al. Relevance of TMPRSS2, CD163/CD206, and CD33 in clinical severity stratification of COVID-19. *Front Immunol* 2022;13.
<https://doi.org/10.3389/fimmu.2022.1094644>.
- [168] Martinez-Diz S, Morales-Álvarez CM, Garcia-Iglesias Y, Guerrero-González JM, Romero-Cachinero C, González-Cabezuelo JM, et al. Analyzing the role of ACE2, AR, MX1 and TMPRSS2 genetic markers for COVID-19 severity. *Hum Genomics* 2023;17.
<https://doi.org/10.1186/s40246-023-00496-2>.
- [169] Preud'homme G, Duarte K, Dalleau K, Lacomblez C, Bresso E, Smaïl-Tabbone M, et al. Head-to-head comparison of clustering methods for heterogeneous data: a simulation-driven benchmark. *Sci Rep* 2021;11:4202. <https://doi.org/10.1038/s41598-021-83340-8>.
- [170] Sinnott RO, Duan H, Sun Y. A Case Study in Big Data Analytics: Exploring Twitter Sentiment Analysis and the Weather. *Big Data: Principles and Paradigms* 2016:357–88. <https://doi.org/10.1016/B978-0-12-805394-2.00015-5>.

- [171] FactoMineR: Exploratory Multivariate Data Analysis with R n.d. <http://factominer.free.fr/> (accessed March 10, 2025).
- [172] Extract and Visualize the Results of Multivariate Data Analyses • factoextra n.d. <https://rpkgs.datanovia.com/factoextra/> (accessed March 10, 2025).
- [173] Chronister WD, Crinklaw A, Mahajan S, Vita R, Koşaloğlu-Yalçın Z, Yan Z, et al. TCRMatch: Predicting T-Cell Receptor Specificity Based on Sequence Similarity to Previously Characterized Receptors. *Front Immunol* 2021;12. <https://doi.org/10.3389/fimmu.2021.640725>.
- [174] IEDB.org: Free epitope database and prediction resource n.d. <https://www.iedb.org/> (accessed March 10, 2025).
- [175] Oksanen J, Simpson GL, Blanchet FG, Kindt R, Legendre P, Minchin PR, et al. Community Ecology Package [R package vegan version 2.6-10]. CRAN: Contributed Packages 2025. <https://doi.org/10.32614/CRAN.PACKAGE.VEGAN>.
- [176] RVAideMemoire: Testing and Plotting Procedures for Biostatistics version 0.9-83-7 from CRAN n.d. <https://rdr.io/cran/RVAideMemoire/> (accessed March 10, 2025).
- [177] Chang CM, Feng PH, Wu TH, Alachkar H, Lee KY, Chang WC. Profiling of T Cell Repertoire in SARS-CoV-2-Infected COVID-19 Patients Between Mild Disease and Pneumonia. *J Clin Immunol* 2021;41:1131–45. <https://doi.org/10.1007/s10875-021-01045-z>.
- [178] Li Y, Hu J, Wang Y, Liu D, Shi Y, Zhang J, et al. T-Cell Repertoire Characteristics of Asymptomatic and Re-Detectable Positive COVID-19 Patients. *Front Immunol* 2022;12. <https://doi.org/10.3389/fimmu.2021.769442>.
- [179] Schultheiß C, Paschold L, Simnica D, Mohme M, Willscher E, von Wenserski L, et al. Next-Generation Sequencing of T and B Cell Receptor Repertoires from COVID-19 Patients Showed Signatures Associated with Severity of Disease. *Immunity* 2020;53:442-455.e4. <https://doi.org/10.1016/j.immuni.2020.06.024>.
- [180] Niu X, Li S, Li P, Pan W, Wang Q, Feng Y, et al. Longitudinal Analysis of T and B Cell Receptor Repertoire Transcripts Reveal Dynamic Immune Response in COVID-19 Patients. *Front Immunol* 2020;11. <https://doi.org/10.3389/fimmu.2020.582010>.
- [181] Park JJ, Lee KAV, Lam SZ, Moon KS, Fang Z, Chen S. Machine learning identifies T cell receptor repertoire signatures associated with

- COVID-19 severity. *Commun Biol* 2023;6.
<https://doi.org/10.1038/s42003-023-04447-4>.
- [182] Bartleson JM, Radenkovic D, Covarrubias AJ, Furman D, Winer DA, Verdin E. SARS-CoV-2, COVID-19 and the aging immune system. *Nat Aging* 2021;1:769–82. <https://doi.org/10.1038/s43587-021-00114-7>.
- [183] Hou X, Wang G, Fan W, Chen X, Mo C, Wang Y, et al. T-cell receptor repertoires as potential diagnostic markers for patients with COVID-19. *International Journal of Infectious Diseases* 2021;113:308–17.
<https://doi.org/10.1016/j.ijid.2021.10.033>.
- [184] Hongying Cheng M, Zhang S, Porritt RA, Noval Rivas M, Paschold L, Willscher E, et al. Superantigenic character of an insert unique to SARS-CoV-2 spike supported by skewed TCR repertoire in patients with hyperinflammation n.d. <https://doi.org/10.1073/pnas.2010722117/-/DCSupplemental>.
- [185] Wang M, Gao P, Ren L, Duan J, Yang S, Wang H, et al. Profiling the peripheral blood T cell receptor repertoires of gastric cancer patients. *Front Immunol* 2022;13. <https://doi.org/10.3389/fimmu.2022.848113>.
- [186] Shomuradova AS, Vagida MS, Sheetikov SA, Zornikova K V., Kiryukhin D, Titov A, et al. SARS-CoV-2 Epitopes Are Recognized by a Public and Diverse Repertoire of Human T Cell Receptors. *Immunity* 2020;53:1245-1257.e5. <https://doi.org/10.1016/j.immuni.2020.11.004>.
- [187] Simnica D, Schultheiß C, Mohme M, Paschold L, Willscher E, Fitzek A, et al. Landscape of T-cell repertoires with public COVID-19-associated T-cell receptors in pre-pandemic risk cohorts. *Clin Transl Immunology* 2021;10. <https://doi.org/10.1002/cti2.1340>.
- [188] Sidhom JW, Baras AS. Deep learning identifies antigenic determinants of severe SARS-CoV-2 infection within T-cell repertoires. *Sci Rep* 2021;11. <https://doi.org/10.1038/s41598-021-93608-8>.
- [189] Dannebaum R, Suwalski P, Asgharian H, Du Zhipei G, Lin H, Weiner J, et al. Highly multiplexed immune repertoire sequencing links multiple lymphocyte classes with severity of response to COVID-19. *EClinicalMedicine* 2022;48:101438. <https://doi.org/10.1016/j.>
- [190] Fang X, Li X, Bian Y, Ji X, Lu J. Relationship between clinical types and radiological subgroups defined by latent class analysis in 2019 novel coronavirus pneumonia caused by SARS-CoV-2-novel coronavirus BMI Body mass index CI Confidence interval CRP C-reactive protein GGO Ground glass opacity HRCT High-resolution CT LCA Latent class

- analysis MERS Middle East respiratory syndrome SARS Severe acute respiratory syndrome WHO World Health Organization 2019. <https://doi.org/10.1007/s00330-020-06973-9>/Published.
- [191] Bowyer RCE, Huggins C, Toms R, Shaw RJ, Hou B, Thompson EJ, et al. Characterising patterns of COVID-19 and long COVID symptoms: evidence from nine UK longitudinal studies. *Eur J Epidemiol* 2023;38:199–210. <https://doi.org/10.1007/s10654-022-00962-6>.
- [192] Vivaldi G, Pfeffer PE, Talaei M, Basera TJ, Shaheen SO, Martineau AR. Long-term symptom profiles after COVID-19 vs other acute respiratory infections: an analysis of data from the COVIDENCE UK study. 2023.
- [193] Van Smeden M, Naaktgeboren CA, Reitsma JB, Moons KGM, De Groot JAH. Latent Class Models in Diagnostic Studies When There is No Reference Standard - A Systematic Review. *Am J Epidemiol* 2014;179:423–31. <https://doi.org/10.1093/aje/kwt286>.
- [194] Gatti A, Radrizzani D, Viganò P, Mazzone A, Brando B. Decrease of Non-Classical and Intermediate Monocyte Subsets in Severe Acute SARS-CoV-2 Infection. *Cytometry Part A* 2020;97:887–90. <https://doi.org/10.1002/cyto.a.24188>.
- [195] Etzerodt A, Moestrup SK. CD163 and inflammation: Biological, diagnostic, and therapeutic aspects. *Antioxid Redox Signal* 2013;18:2352–63. <https://doi.org/10.1089/ars.2012.4834>.
- [196] Zhang D, Guo R, Lei L, Liu H, Wang Y, Wang Y, et al. Frontline Science: COVID-19 infection induces readily detectable morphologic and inflammation-related phenotypic changes in peripheral blood monocytes. *J Leukoc Biol* 2021;109:13–22. <https://doi.org/10.1002/JLB.4HI0720-470R>.
- [197] Rajamanickam A, Kumar NP, Pandiarajan AN, Selvaraj N, Munisankar S, Renji RM, et al. Dynamic alterations in monocyte numbers, subset frequencies and activation markers in acute and convalescent COVID-19 individuals. *Sci Rep* 2021;11. <https://doi.org/10.1038/s41598-021-99705-y>.
- [198] Ruiz Ortega M, Pogorelyy M V., Minervina AA, Thomas PG, Mora T, Walczak AM. Learning predictive signatures of HLA type from T-cell repertoires. *PLoS Comput Biol* 2025;21:e1012724. <https://doi.org/10.1371/journal.pcbi.1012724>.
- [199] Mazzotti L, Gaimari A, Bravaccini S, Maltoni R, Cerchione C, Juan M, et al. T-Cell Receptor Repertoire Sequencing and Its Applications: Focus

- on Infectious Diseases and Cancer. *Int J Mol Sci* 2022;23.
<https://doi.org/10.3390/ijms23158590>.
- [200] Barenes P, Quiniou V, Shugay M, Egorov ES, Davydov AN, Chudakov DM, et al. Benchmarking of T cell receptor repertoire profiling methods reveals large systematic biases. *Nat Biotechnol* 2021;39:236–45.
<https://doi.org/10.1038/s41587-020-0656-3>.



UNIVERSIDAD
DE GRANADA