INTERNATIONAL DOCTORAL
SCHOOL OF THE USC

Sandra
Viz Lasheras

PhD Thesis

Host-Based Multi-Omics for the Diagnosis and Management of Respiratory Infections

Santiago de Compostela, 2024

Doctoral Programme in Molecular Medicine

ESCOLA DE DOUTORAMENTO
INTERNACIONAL DA USC

DOCTORAL THESIS

HOST-BASED MULTI-OMICS FOR THE DIAGNOSIS AND MANAGEMENT OF RESPIRATORY INFECTIONS

Sandra Viz Lasheras

Supervisors:

Alberto Gómez Carballa

Antonio Salas Ellacuriaga

Federico Martinón Torres

Tutor: Antonio Salas Ellacuriaga

PHD PROGRAMME IN MOLECULAR MEDICINE

SANTIAGO DE COMPOSTELA / LUGO

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

Conflicto de intereses

A doutoranda Sandra Viz Lasheras declara que é titular de dúas patentes europeas solicitadas con numero **EP24382240** e **EP24382675** e desenvolvidas no contexto desta tese de doutoramento:

1. Sandra Viz Lasheras, Alberto Gómez Carballa, Antonio Salas Ellacuriaga, Federico Martinón Torres. **BACMO. EP24382240**. 5 de Marzo 2024
2. Sandra Viz Lasheras, Alberto Gómez Carballa, Antonio Salas Ellacuriaga, Federico Martinón Torres. **MYCOTEST**. **EP24382675.** 21 Xuño 2024**.**

Santiago de Compostela, a 17 de Xaneiro de 2025

USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

A Jone e Tino

AGRADECEMENTOS

Aos meus directores, Antonio, Alberto e Federico, por todo o apoio e esforzo durante estes anos, porque sen vos estas páxinas non serían posibles. A Toño, por abrirme as portas deste mundo e confiar en min dende a primeira reunión. Pola enorme sorte que teño de estar preto da túa mente excepcional e por esa parte tan humana que te caracteriza. Grazas por regalarme as doses de optimismo que ás veces me faltan a min. Para Alberto, infinitas grazas por todo o esforzo e paciencia, por confiar en min, acollerme e ensinarme dende cero e sen medo, cousas que nunca pensei que chegaría a facer. Es un exemplo a seguir. Grazas Federico, pola túa xenerosidade, polos retos e por todas as oportunidades.

Ao mellor grupo de investigación no que puiden caer para comezar esta aventura: GENVIP/GenPop. Grazas a Alba, Isa, Juli, Lara, Laura, Lúa, Majo, Miriam, Mirian, Narmeen, Nour, Patri e Xabi. A Sara e Jaco, polas súas valiosas contribucións e a axuda incondicional sempre que a precisei. E porque ao noso lado *The Office* ás veces quédase curta... facedes os meus longos días máis divertidos.

To Myrsini and her team at Imperial College for wellcome me as part of their group for three exciting months. I am truly grateful for the opportunity to learn from all of you. Special thanks to Ortensia, Becca, and Claire for always being there to help whenever needed.

Ao meu universo de *soles e estrelas*, porque cada momento con vos é un tesouro. Unha mención especial a Clara e Ada pola súa infinita xenerosidade, tanto en asuntos científicos como persoais. A Paula, Yaiza e Sabela, por acompañarme dende o inicio neste carrusel de emocións e contradicións que inevitablemente rodean a bioloxía, facéndoo todo máis sinxelo e livián. Nerea, non perdoo que me convenceras para mergullarme nisto, pero tampouco teño tempo vital para agradecer que esteas ao meu carón, do xeito no que o fas, ata o final. A esas persoas *Bravas* que a música puxo neste camiño, por acollerme e por facelo infinitamente máis pleno e feliz.

A toda a familia Viz/Lasheras, por ser o mellor refuxio para desfrutar e desconectar que unha doutoranda podería pedir. A Alberto e Andrea, grazas especiais. E por suposto aos meus pais, por terme transmitido unha milésima parte da súa curiosidade e afán por aprender, cada un ao seu xeito.

Por último, a Diego, por estar e apoiar dende o primeiro minuto e sempre.

USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

SCIENTIFIC PRODUCTION OF THE THESIS

Five articles are presented here as the foundational core of the current doctoral thesis. All of them have been published in first-quartile (Q1) and decile (D1) journals according to the JCR ranking, (see Results section and Annex I). The listed articles are:

1. Gómez-Carballa A, Rivero-Calle I, Pardo-Seco J, Gómez-Rial J, Rivero-Velasco C, Rodríguez-Núñez N, Barbeito-Castiñeiras G, Pérez-Freixo H, Cebey-López M, Barral-Arca R, Rodriguez-Tenreiro C, Dacosta-Urbieta A, Bello X, Pischedda S, Currás-Tuala MJ, Viz-Lasheras S, Martinón-Torres F, Salas A; GEN-COVID study group. A multi-tissue study of immune gene expression profiling highlights the key role of the nasal epithelium in COVID-19 severity. *Environmental Research*. 2022;7(210):112890. (**First JCR Quartile and Decile**)

2. Gómez-Carballa A, Pardo-Seco J, Pischedda S, Rivero-Calle I, Butler-Laporte G, Richards JB, Viz-Lasheras S, Martinón-Torres F, Salas A; GENDRES network. Sex-biased expression of the TLR7 gene in severe COVID-19 patients: Insights from transcriptomics and epigenomics. *Environmental Research*. 2022;12(215 Pt 2):114288. (**First JCR Quartile and Decile**)

3. Pardo-Seco J, Viz-Lasheras S, Bello X, Gómez-Carballa A, Camino-Mera A, Pischedda S, Currás-Tuala MJ, Rivero-Calle I, Dacosta A, Caamaño-Viña F, Rodríguez-Tenreiro C, Cifuentes I, Méndez C, Khor CC, Martinón-Torres F, Salas A. Whole exome sequencing identifies new susceptibility candidates underlying community-acquired pneumonia. *Genes & Diseases*. 2023;10 (5): 1248-1257. (**First JCR Quartile and Decile)**

4. Viz-Lasheras S, Gómez-Carballa A, Pardo-Seco J, Bello X, Rivero-Calle I, Dacosta AI, Kaforou M, Habgood-Coote D, Cunnington AJ, Emonts M, Herberg JA, Wright VJ, Carrol ED, Paulus SC, Zenz W, Kohlfürst DS, Van der Flier M, de Groot R, Schlapbach LJ, Agyeman P, Pollard AJ, Fink C, Kuijpers TT, Anderson S, Calvo C, Martínez-Padilla MC, Pérez-Aragón A, Gómez-Sánchez E, Valencia-Ramos J, Giménez-Sánchez F, Alonso-Quintela P, Moreno-Galarraga L, von Both U, Pokorn M, Zavadska D, Tsolia M, Vermont CL, Moll HA, Levin M, Martinón-Torres F, Salas A; EUCLIDS, DIAMONDS,

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

GENDRES, and PERFORM consortia. A 5-transcript signature for discriminating viral and bacterial etiology in pediatric pneumonia. *iScience*: 2025. (**First JCR Quartile**)

5. Viz-Lasheras S, Gómez-Carballa A, Bello X, Rivero-Calle I, Dacosta AI, Kaforou M, Habgood-Coote D, Cunnington AJ, Emonts M, Herberg JA, Wright VJ, Carrol ED, Paulus SC, Zenz W, Kohlfürst DS, Schweintzger N, Van der Flier M, de Groot R, Schlapbach LJ, Agyeman P, Pollard AJ, Fink C, Kuijpers TT, Anderson S, von Both U, Pokorn M, Zavadska D, Tsolia M, Moll HA, Vermont C, Levin M, Martinón-Torres F, Salas A; EUCLIDS, PERFORM, and DIAMONDS consortia. A diagnostic host-specific transcriptome response for *Mycoplasma pneumoniae* pneumonia to guide pediatric patient treatment. 2025. *Nature Communications*: 2025; *16*(1), 673. (**First JCR Quartile and Decile**)

USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

TABLE OF CONTENTS

ABBREVIATION LIST 11
FIGURE LIST 13
TABLE LIST 13
RESUMO 15
ABSTRACT 25
1. INTRODUCTION 35
1.1. LOWER RESPIRATORY TRACT INFECTIONS 35
1.1.1. Public health impact, implications and prevention strategies 35
1.1.2. Etiology and Pathogenesis of LRTIs 37
1.1.3. Diagnosis and treatment of LRTIs 39
1.2. PNEUMONIA DISEASE 40
1.2.1. Types of pneumonia 41
1.2.2. Challenges in the diagnosis and treatment of pneumonia 42
1.3. SARS-COV-2 OUTBREAK AND COVID-19 43
1.4. THE STUDY OF HOST RESPONSE TO INFECTIOUS DISEASES: A CHANGE OF PARADIGM 45
1.4.1. Host Immune response to infectious diseases 45
1.4.2. Sex differences in the immune response to infectious diseases 48
1.4.3. Omic approaches to the study of infectious diseases 51
1.4.3.1. Host transcriptomics for biomarker discovery in infectious diseases 52
1.4.3.2. Study of tissue-specific host response to infections 53
1.4.3.3. Machine Learning strategies to enhance diagnosis and prognosis in infectious diseases 54
2. HYPOTHESIS 57
3. OBJECTIVES 61
3.1. GENERAL AND SPECIFIC OBJECTIVES 61
4. METHODS 65
4.1. STUDY DESIGN AND SAMPLE SELECTION 65
4.2. TRANSCRIPTOMIC ANALYSIS 68
4.2.1. Sample collection and processing 68
4.2.2. Gene expression platforms 69
4.2.2.1. Microarray data processing and analysis 69
4.2.2.2. RNA-seq data processing and analysis 70

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

4.2.2.3. NanoString nCounter data processing and analysis 71
4.2.3. Unsupervised learning methods 73
4.2.3.1. Principal component analysis 74
4.2.3.2. Hierarchical clustering analysis and Heatmap 75
4.2.4. Supervised Learning 75
4.2.4.1. Machine learning for biomarker and signatures discovery 75
4.2.5. Graphical Representation 79
4.2.6. Co-expression analysis 79
4.2.7. Pathways analysis 80
4.3. EPIGENOMIC ANALYSIS 82
4.4. GENOMIC ANALYSIS 82
5. RESULTS 87
5.1. ARTICLE 1: A MULTI-TISSUE STUDY OF IMMUNE GENE EXPRESSION PROFILING HIGHLIGHTS THE KEY ROLE OF THE NASAL EPITHELIUM IN COVID-19 SEVERITY 87
5.2. ARTICLE 2: SEX-BIASED EXPRESSION OF THE TLR7 GENE IN SEVERE COVID-PATIENTS: INSIGHTS FROM TRANSCRIPTOMICS AND EPIGENOMICS 90
5.3. ARTICLE 3: WHOLE EXOME SEQUENCING IDENTIFIES NEW SUSCEPTIBILITY CANDIDATES UNDERLYING COMMUNITY-ACQUIRED PNEUMONIA 92
5.4. ARTICLE 4: A 5-TRANSCRIPT SIGNATURE FOR DISCRIMINATING VIRAL AND BACTERIAL ETIOLOGY IN PEDIATRIC PNEUMONIA 94
5.5. ARTICLE 5: A DIAGNOSTIC HOST-SPECIFIC TRANSCRIPTOME RESPONSE FOR MYCOPLASMA PNEUMONIAE PNEUMONIA TO GUIDE PEDIATRIC PATIENT TREATMENT 98
6. DISCUSSION 105
6.1. STUDY OF THE HOST RESPONSE TO COVID-19: IMPLICATIONS FOR DISEASE SEVERITY 106
6.2. CHALLENGES IN PNEUMONIA DIAGNOSIS, TREATMENT, AND SUSCEPTIBILITY. THE RESPONSE LIES IN THE HOST 111
7. CONCLUSIONS 121
8. REFERENCES 125
ANNEX I: Published or Accepted Works 141
ANNEX II: Ethical Approval 145

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

ABBREVIATION LIST

AUC: Area Under the Curve
ARDS: Acute Respiratory Distress Syndrome
CAP: Community Acquired Pneumonia
COVID-19: coronavirus disease 2019
DC: Dendritic Cell
DMPs: Differentially Methylated Positions
FS: Feature selection
GEO: Gene Expression Omnibus
GSEA: Gene Set Expression Analysis
HIC: High-Income Countries
HK: House Keeping
IFN: Interferon
IFN-I: Type I Interferons
IFN-II: Type II Interferons
ICU: Intensive Care Unit
ILCs: Innate lymphoid cells
LMIC: Low- and Middle-Income Countries
LRTI: Lower Respiratory Tract Infection
LASSO: Least Absolute Shrinkage and Selection Operator
MAF: Minor Allele Frequency
ML: Machine Learning
MyD88: Myeloid Differentiation Primary Response 88
PC: Principal Components
PCR: Polymerase Chain Reaction
RSV: Respiratory Syncytial Virus
RT-PCR: Reverse Transcription PCR
ROC: Receiver Operating Characteristic
SARS-CoV-2: Severe Acute Respiratory Syndrome Coronavirus 2
SNPs: Single Nucleotide Polymorphisms

USC
DE SANTIAGO
DE COMPOSTELA

TLRs: Toll-like receptors
TOM: Topological Overlap Matrix
WHO: World Health Organization
WES: Whole Exome Sequencing
XCI: X-Chromosome Inactivation

USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

FIGURE LIST

Figure 1. Schematic representation of the Immune System. 46
Figure 2 Simplified representation of TLR7 activation and signalling. 49
Figure 3 Human X-Chromosome Inactivation scheme. 50
Figure 4 Multi -omic approaches and its application in the context of human diseases. 51
Figure 5. Graphical representation of the study design of the thesis. 65
Figure 6 General workflow for Illumina Sequencing. 70
Figure 7 Nanostring technology pipeline. 72
Figure 8 Simplified example of a PCA visualization in 2 dimensions before and after batch correction of the data. 74
Figure 9. Example of a pipeline for biomarker or signature discovery. Self-made figure created with BioRender 77

TABLE LIST

Table 1. Main causal pathogens of LRTIs globally. 38
Table 2. Public databases used for the analysis 67
Table 3. New generated expression data used for the analysis 68
Table 4. New generated genomic data used for the analysis. 83

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

RESUMO

As infeccións respiratorias, especialmente as do tracto respiratorio inferior (LRTIs, polo seu acrónimo inglés de *Lower Respiratory Tract Infetions*), representan un problema significativo de saúde pública sendo causantes de millóns de mortes anuais en todo o mundo. Segundo a Organización Mundial da Saúde, son a quinta causa principal de morte a nivel global, cun maior impacto en poboacións vulnerables, coma nenos e persoas maiores ou inmuno-comprometidas. Por conseguinte, o impacto económico das LRTI é tamén considerable, abranguendo tanto custos directos relacionados co tratamento e hospitalización, como custos indirectos derivados da perda de produtividade e discapacidades a longo prazo. Malia que durante a pandemia do SARS-CoV-2 se produciu unha redución na incidencia das LRTIs propiciada por medidas como o distanciamento social e o uso de máscaras, estas infeccións seguen a representar un desafío global. Actualmente, a súa prevención céntrase en campañas de vacinación integradas, a redución da contaminación do aire e a promoción de boas prácticas de hixiene.

Entre as formas de LRTIs, a pneumonía destaca como unha das máis graves e letais, sendo a principal causa de morte en nenos menores de cinco anos. A súa etioloxía é diversa, podendo ser causada por bacterias, virus ou fungos e a sintomatoloxía que produce é a miúdo específica, asemellándose á de outras doenzas respiratorias. A pneumonía adoita clasificarse segundo o ambiente no que se adquiriu, distinguindo entre pneumonía adquirida na comunidade e pneumonía hospitalaria. Ademais, segundo a súa etioloxía pode dividirse tamén en pneumonía típica e atípica. As pneumonías típicas son causadas principalmente por patóxenos como *Streptococcus pneumoniae*, *Haemophilus influenzae* e *Staphylococcus aureus*, así como por virus como o respiratorio sincitial e a influenza. Pola contra, as pneumonías atípicas, que inclúen patóxenos como *Mycoplasma pneumoniae e Chlamydia pneumoniae*, que presentan un desafío diagnóstico engadido debido aos seus ciclos epidemiolóxicos menos constantes e á dificultade para detectalos mediante métodos estándar.

Por todo isto, o diagnóstico e tratamento da pneumonía presentan retos únicos que se derivan tanto da diversidade de axentes causais como da variabilidade e inespecificidade na súa presentación clínica, os atrasos na obtención de resultados

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

diagnósticos fiables e a crecente ameaza da resistencia antimicrobiana. Esta diversidade reflicte na complexidade do diagnóstico e tratamento da enfermidade, que require métodos específicos para identificar os patóxenos implicados e establecer a intervención máis axeitada. Entre os indicadores diagnósticos máis comúns destacan as probas de raios X e parámetros como os niveis de procalcitonina, proteína C-reactiva, lactato e marcadores de coagulación que poden aportar información sobre o proceso de infección activo e a súa etioloxía. Porén, estas estratexias non sempre poden excluír a presencia dunha infección bacteriana, e en ocasións resulta difícil integrar os resultados co resto de parámetros clínicos para extraer conclusións. Ademais, os marcadores descritos e utilizados para a poboación adulta, moitas veces resultan ineficaces na poboación pediátrica. O desenvolvemento de ensaios baseados na reacción en cadea da polimerasa, aumentou notablemente a nosa capacidade para diagnosticar con precisión infeccións virais coñecidas e emerxentes, ademais de posibilitar a detección de múltiples patóxenos nunha soa proba. Cómpre destacar, con todo, que os ensaios moleculares son menos eficaces na detección de infeccións bacterianas, e en moitos centros de saúde a súa detección sigue baseándose en cultivos bacterianos ou seroloxías menos específicas. Recentemente, describíronse novos biomarcadores para diferentes infeccións respiratorias empregando tecnoloxías de nova xeración e métodos avanzados de detección, que emerxen como ferramentas clave para mellorar a precisión diagnóstica. Non obstante, a día de hoxe, a maioría dos casos de pneumonía son diagnosticados unicamente baseándose no cadro clínico do paciente, e unha gran proporción son diagnosticadas incorrectamente. Ademais, o tratamento antibiótico adminístrase antes de ter unha confirmación dos resultados de laboratorio que confirmen a información sobre a etioloxía da infección (ex. viral ou bacteriana), contribuíndo ao aumento da resistencia antibiótica de certos patóxenos.

Coa recente pandemia da COVID-19 (polo seu acrónimo inglés de *coronavirus disease 2019*) causada polo SARS-CoV-2, a comunidade científica entregouse na procura de respostas sobre a enfermidade e o desenvolvemento de vacinas a unha velocidade sen precedentes. Esta doenza, caracterizada por unha ampla variedade de manifestacións clínicas que van dende casos asintomáticos ata cadros graves de pneumonía e fallo multiorgánico, puxo de manifesto a necesidade dun estudo máis profundo a nivel molecular para comprender os seus mecanismos de acción, os factores de risco asociados e as diferenzas biolóxicas implicadas na gravidade dos síntomas. A resposta inmunitaria fronte ao SARS-CoV-2 é un factor clave na evolución clínica da enfermidade, xa que determina en gran medida se o virus pode ser controlado de forma eficaz ou se, pola contra, se desencadean complicacións graves. A activación temperá e eficiente da resposta innata, principalmente mediada polos interferóns de tipo I, é esencial para limitar a replicación viral nos estadios iniciais da infección. Porén, en casos graves, esta resposta

antiviral adoita estar comprometida, o que permite a diseminación do virus e contribúe á activación excesiva da inflamación sistémica.

Por outra banda, factores como o sexo, a idade e a presenza de comorbilidades teñen un impacto significativo na dinámica desta resposta inmunitaria, condicionando tanto a susceptibilidade á infección como a gravidade da enfermidade. En particular, estudos recentes apuntan a diferenzas inmunolóxicas significativas entre homes e mulleres, atribuíbles a interaccións entre factores hormonais, xenéticos e epixenéticos, que poden influír na capacidade de recoñecer e responder ao virus. Entre os factores asociados propostos atópase o xen *TLR7*, que desempeña un papel fundamental na detección do ARN viral e na activación da resposta inmunitaria innata, promovendo a produción de interferóns tipo I e outras citocinas proinflamatorias. Este xen está localizado no cromosoma X, e diversos estudos suxiren que pola súa localización podería evitar o silenciamento característico nos xenes dun dos cromosoma X nas mulleres, xerando unha dobre dose da súa proteína, o que podería explicar parte das diferenzas observadas na gravidade da COVID-19 entre sexos.

Nos últimos anos, a investigación das enfermidades infecciosas sufriu un cambio de paradigma, cambiando o foco cara o estudo da resposta do hóspede fronte á infección. Este cambio de perspectiva é fundamental para comprender como os diferentes patóxenos interactúan co sistema inmunitario e permite identificar factores que inflúen na progresión ou resolución da doenza. Tradicionalmente, as enfermidades infecciosas analizábanse principalmente a través das características dos patóxenos e os mecanismos de infección e a sintomatoloxía que producían. Na actualidade, o enfoque na resposta específica do hóspede revelou como a variabilidade individual, incluíndo factores xenéticos, epixenéticos e inmunitarios, condiciona a susceptibilidade, a gravidade e a evolución das infeccións, así como a eficacia dos tratamentos.

O desenvolvemento recente de novas tecnoloxías no ámbito das ómicas permiten unha aproximación máis completa e integradora para estudar as interaccións entre patóxeno-hospede, facilitando a identificación de rutas clave, biomarcadores diagnósticos ou pronósticos e posibles dianas terapéuticas. Estas tecnoloxías producen unha enorme cantidade de datos complexos e multidimensionais, cuxa interpretación require o uso de ferramentas computacionais avanzadas. Neste contexto, as técnicas de aprendizaxe automática emerxen como elementos fundamentais para a análise e integración eficiente destes datos, facilitando a identificación de patróns relevantes, relacións entre variables biolóxicas e a creación de modelos preditivos que poden ser aplicados no diagnóstico, prognóstico e manexo das infeccións respiratorias. A combinación destas ferramentas coas aproximacións ómicas representa unha revolución no estudo das enfermidades infecciosas, permitindo unha comprensión máis profunda dos mecanismos moleculares

subxacentes e promovendo o desenvolvemento de estratexias de intervención máis precisas e personalizadas.

Obxectivo

Esta tese ten como obxectivo investigar as dinámicas complexas das infeccións respiratorias dende unha perspectiva centrada no hóspede. Este enfoque abrangue o estudo de aspectos clave como i) as interaccións hóspede-patóxeno e os factores que inflúen na progresión e gravidade da enfermidade ii) o papel das diferenzas de sexo na resposta inmune, incluíndo mecanismos xenéticos e epixenéticos, iii) a identificación de xenes de susceptibilidade e biomarcadores asociados á evolución clínica das infeccións e iv) o desenvolvemento de ferramentas diagnósticas innovadoras baseadas en expresión xénica empregando algoritmos de aprendizaxe automática, para de mellorar a precisión e rapidez dos métodos diagnósticos actuais. Con estes fins, propuxéronse os seguintes obxectivos específicos:

i) Analizar as causas da resposta inmune diferencial á COVID-19 mediante o estudo do transcriptoma de pacientes con COVID-19 estratificados segundo a gravidade da enfermidade.

ii) Investigar os mecanismos inmunitarios específicos de tecidos que poden subxacer á gravidade da COVID-19 mediante o estudo dos perfís de expresión xénica en diferentes mostras biolóxicas (saliva, epitelio nasal e sangue).

iii) Estudar a expresión diferencial por sexo e a regulación epixenética do xene ligado ao cromosoma X *TLR7* en pacientes con COVID-19 estratificados segundo a gravidade.

iv) Identificar variantes xenéticas asociadas coa susceptibilidade á pneumonía adquirida na comunidade (CAP) mediante a secuenciación completa do exoma.

v) Avaliar a correlación entre as variantes xenéticas identificadas e os perfís de expresión xénica en pacientes con pneumonía, avaliando o seu impacto funcional potencial na resposta inmune e na progresión da enfermidade.

vi) Desenvolver e validar unha sinatura transcriptómica de sangue total que permita diferenciar especificamente entre pneumonía viral e bacteriana en nenos.

vii) Desenvolver e validar unha sinatura transcriptómica de sangue total que permita diferenciar a pneumonía causada por *Mycoplasma pneumoniae* doutras pneumonías en nenos.

USC DE SANTIAGO DE COMPOSTELA

Métodos

A metodoloxía do traballo integra unha abordaxe multi-ómica (xenómica, transcriptómica e epixenómica) das enfermedades respiratorias utilizando diferentes tecnoloxías e cohortes para estudar estas condicións dende una perspectiva centrada no hóspede. A investigación inclúe a análise de mostras de sangue, epitelio nasal e saliva tanto de pacientes con enfermidades respiratorias como controis sans, empregando tecnoloxías como RNA-seq, microarrays e a plataforma NanoString nCounter para abordar preguntas específicas relacionadas coa resposta inmune e os mecanismos subxacentes que determinan a gravidade da enfermidade. Ademais, realizáronse estudos xenómicos e epixenómicos co obxectivo de identificar variantes xenéticas e patróns de metilación que inflúen na susceptibilidade e progresión destas enfermidades. Tamén se aplicaron algoritmos de aprendizaxe automática, como LASSO, para integrar e analizar os datos biolóxicos, desenvolvendo modelos preditivos que optimizan o diagnóstico e a clasificación de mostras segundo a súa etioloxía, na procura de mellorar a precisión das probas diagnósticas actuáis.

A tecnoloxía RNA-seq utilizouse para avaliar a expresión xénica global en sangue que permitiu analizar a actividade transcriptómica en diferentes contextos de gravidade clínica e estudar como varían as respostas inmunes nas infeccións respiratorias. O sistema NanoString permitiu unha análise enfocada dun conxunto seleccionado de xenes inmunolóxicos clave. O seu deseño empregouse para estudar as respostas inmunes específicas dos tecidos e validar as observacións dos análises transcriptómicos realizados con otras tecnolíxáis como RNA-Seq ou microarrays. O análise de metilación do ADN empregouse para estudar como as modificacións epixenéticas afectan a regulación da expresión xénica en pacientes con diferentes gravidades de infeccións respiratorias e a comparación destes niveis de metilación en xenes diferencialmente expresando segundo o sexo biolóxico. Por último, a secuenciación do exoma completo implementouse para identificar variantes xenéticas relacionadas coa susceptibilidade a pneumonía adquirida na comunidade.

Estas tecnoloxías permiten xerar unha gran cantidade de datos coa finalidade de proporcionar un enfoque holístico que abranguese desde a expresión xénica ata os mecanismos reguladores e xenéticos subxacentes. Este deseño permitiu abordar de maneira sistemática as interaccións entre factores xenéticos, epixenéticos e transcriptómicos no estudo das infeccións respiratorias, con especial foco na pneumonía e a COVID-19.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Resultados

As infeccións respiratorias, e en particular a pneumonía, continúan sendo un desafío significativo para a saúde pública global, destacándose como unha das principais causas de mortalidade no mundo. Esta tese explorou a resposta do hóspede fronte ás infeccións respiratorias a través de enfoques multi-ómicos, xerando novos coñecementos que poderían mellorar o diagnóstico, prognóstico e tratamento destas doenzas.

Resposta do hóspede ante a COVID-19: implicacións na severidade

Un dos enfoques centrais foi o estudo da resposta inmune ao SARS-CoV-2 en mostras de sangue, epitelio nasal e saliva, revelando que a resposta inmune é específica segundo o tecido e está directamente relacionada coa gravidade da enfermidade. Observouse que as respostas locais robustas nos tecidos das vías respiratorias superiores, especialmente mediadas polos IFN, son determinantes para limitar a progresión da infección. Nos casos leves, os tecidos nasais mostraron unha forte activación antiviral, con expresión elevada de xenes inducidos por IFN o cal favorece o control local do virus. Pola contra, nos casos graves, detectouse unha resposta antiviral insuficiente no epitelio nasal e unha maior activación inmune sistémica, caracterizada por unha resposta excesiva dos neutrófilos e unha activación anormal de vías inflamatorias e de coagulación no sangue. En particular, as mostras de saliva dos pacientes graves revelaron signos de propagación viral cara ao tracto respiratorio inferior, cunha sobre-expresión de xenes implicados na evasión inmune e na diseminación do virus. Estes achados suxiren que tanto o epitelio nasal como a mucosa bucal teñen un papel crucial na gravidade da COVID-19, abrindo oportunidades para novos biomarcadores diagnósticos e terapias específicas.

Por outra banda, estudouse o impacto das diferenzas biolóxicas entre sexos na resposta inmune ao COVID-19, co foco na expresión diferencial do xene *TLR7* en diferentes tecidos. Este xene, situado no cromosoma X, amosou unha expresión máis elevada en mulleres que podería estar proporcionando unha vantaxe significativa na resposta á infección en comparación cos homes. A menor expresión de *TLR7* nos casos graves masculinos foi correlacionada con respostas insuficientes de IFN-I, un compoñente esencial para o control viral temperán. Ademais, observáronse diferenzas nos patróns de metilación de *TLR7* entre homes e mulleres, o que suxire que factores os factores epixenéticos tamén contribúen a estas disparidades. Estes achados subliñan a necesidade de incorporar perspectivas de xénero na investigación e tratamento das infeccións respiratorias, permitindo a personalización de estratexias terapéuticas.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Retos no diagnóstico, tratamento e susceptibilidade á pneumonía: a resposta radica no hóspede.

A análise xenómica revelou novas variantes asociadas á susceptibilidade á pneumonía adquirida na comunidade. En particular, detectouse unha asociación significativa entre a gravidade da pneumonía e variantes no xene *FAM118A*, cuxa expresión diminuída foi observada en pacientes con pneumonía grave. Mediante os tests de asociación de colapso SKAT-O identificáronse outros candidatos como *COL13A1*, *C20orf96, HEXIM1*, *NPAS4* e *STRC*. É salientable que ningún destes xenes fora previamente asociado coa pneumonía, o que destaca o valor da secuenciación do exoma completo no descubrimento de factores xenéticos previamente descoñecidos. Ademais, a avaliación dos patróns de expresión xénica para estes seis xenes, revelou unha expresión diferencial significativa en pacientes con pneumonía en comparación cos controis, suxerindo un potencial destes xenes para fins diagnósticos ou prognósticos e abrindo una vía para o seu estudo.

En relación á pneumonía pediátrica, neste estudo identificouse unha firma transcriptómica de cinco xenes (*FAM20A, BAG3, TDRD9, MXRA7 e KLF14*) que permite diferenciar entre pneumonía bacteriana e viral con alta precisión (AUC > 0.92). Esta firma, validada con éxito nunha cohorte independente, destaca polo seu potencial para mellorar o diagnóstico actual e reducir o uso innecesario de antibióticos, contribuíndo así a combater a resistencia antimicrobiana e mellorar o manexo clínico destas infeccións.

No caso das pneumonías causada por *Mycoplasma pneumoniae*, os análises transcriptómicos mostraron unha resposta moi específica fronte a este patóxeno, o que permitiu describir varias firmas transcriptómicas que discriminan con elevada precisión esta etioloxía doutros tipos de pneumonía como a viral ou a causada por outras bacterias que non son mycoplasmas. Estas firmas destacan pola súa alta sensibilidade e especificidade, sendo particularmente útiles para guiar tratamentos con antibióticos macrólidos no canto de tratamento con beta-lactámicos que son o antibiótico de primeira liña administrado por defecto.

Estes resultados resaltan a importancia dun enfoque centrado no hóspede así como do uso de tecnoloxías multi-ómicas no estudo das enfermidades infecciosas e subliñan o potencial clínico das firmas transcriptómicas e dos biomarcadores identificados. Os achados desta tese abren o camiño cara á implementación de ferramentas de diagnóstico máis rápidas, precisas e accesibles, favorecendo un enfoque máis personalizado e efectivo no manexo das infeccións respiratorias. Coa continua evolución das tecnoloxías de secuenciación e análise ómica, así como o desenvolvemento de novos dispositivos de detección transferibles aos puntos de atención, prevese que estas innovacións transformen

USC DE SANTIAGO DE COMPOSTELA

o manexo clínico das infeccións respiratorias, reducindo a carga destas enfermidades e mellorando o curso das mesmas en pacientes de todo o mundo. Ademais o uso de diferentes métodos de estudo da expresión xénica, como RNA-Seq, microarrays e a tecnoloxía nCounter, dan robustez ás achegas deste traballo.

Conclusións

Os resultados obtidos a partir das análises realizadas durante esta tese doutoral permitiron acadar as seguintes conclusións:

i) A gravidade da enfermidade a nivel sistémico pode depender en gran medida da resposta inmune local ao COVID-19. A evidencia destaca que a resposta inmune na cavidade nasal, particularmente o equilibrio entre as respostas antivirais mediadas por interferón e as vías inflamatorias, pode determinar se o virus queda contido localmente ou se disemina de forma sistémica, influíndo así na progresión e gravidade da enfermidade.

ii) A cavidade bucal pode desempeñar un papel clave na diseminación sistémica da infección por SARS-CoV-2. As análises transcriptómicas revelan que os casos graves de COVID-19 amosan replicación viral activa e disrupción nas funcións celulares nos tecidos orais, suxerindo que a cavidade bucal funciona non só como un lugar de entrada para o virus, senón tamén como un punto que contribúe á súa diseminación sistémica.

iii) As diferenzas específicas por sexo na expresión de *TLR7* e nos patróns de metilación indican unha influencia na gravidade do COVID-19. Os homes, especialmente aqueles con enfermidade grave, mostraron consistentemente unha menor expresión de *TLR7* en comparación coas mulleres. Esta redución, impulsada por mecanismos xenéticos e epixenéticos vinculados ao cromosoma X, contribúe a respostas antivirais deficientes e a un maior risco de desenlaces graves, destacando o papel clave de *TLR7* na modulación da gravidade da enfermidade.

iv) Unha firma transcriptómica de 5 xenes pode diferenciar con precisión pneumonía bacteriana de pneumonía viral, e foi validada nunha cohorte independente. Este método ten potencial para ser traducido a dispositivos de diagnóstico no punto de atención, proporcionando unha ferramenta rápida e fiable para mellorar o manexo clínico da pneumonía pediátrica e reducir a prescrición innecesaria de antibióticos para infeccións virais.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

v) A pneumonía por *Mycoplasma pneumoniae* altera a expresión xénica do hóspede dun xeito altamente específico, distinguible mediante firmas transcriptómicas. Combinacións de 8 transcritos de 17 xenes permiten diferenciar con precisión a pneumonía por M. pneumoniae doutras pneumonías bacterianas e virais. Dado que *M. pneumoniae* é resistente aos antibióticos beta-lactámicos, aplicar estas firmas na práctica clínica podería mellorar a toma de decisións no tratamento e reducir o uso inadecuado de antibióticos.

vi) Identificáronse variantes xenéticas asociadas coa susceptibilidade á pneumonía adquirida na comunidade, observándose novas asociacións en xenes como *FAM118A*, *NPAS4*, *C20orf96*, *COL13A1* e *STRC*.

vii) A expresión xénica dos xenes candidatos identificados está significativamente desregulada na pneumonía non-COVID-19 e na pneumonía grave por COVID-19. Isto suxire a súa implicación en interaccións críticas hóspede-patóxeno e a súa posible contribución á gravidade da CAP, apoiando a súa investigación como posibles biomarcadores.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

ABSTRACT

Respiratory infections, especially lower respiratory tract infections (LRTIs), represent a significant public health problem, causing millions of deaths annually worldwide. According to the World Health Organization, they are the fifth leading cause of death globally, with a greater impact on vulnerable populations, such as children and elderly or immuno-compromised people. Consequently, the economic impact of LRTI is also considerable, encompassing both direct costs related to treatment and hospitalization, as well as indirect costs derived from lost productivity and long-term disabilities. Although during the SARS-CoV-2 pandemic there was a reduction in the incidence of LRTIs due to measures such as social distancing and the use of masks, these infections continue to represent a global challenge. Currently, their prevention is focused on integrated vaccination campaigns, reduction of air pollution and promotion of good hygiene practices.

Among the forms of LRTIs, pneumonia stands out as one of the most serious and lethal, being the main cause of death in children under five years of age. Its etiology is diverse, it can be caused by bacteria, viruses or fungi and the symptoms it produces are very specific, similar to those of other respiratory diseases. Pneumonia should be classified according to the environment in which it was acquired, distinguishing between community-acquired pneumonia and hospital-acquired pneumonia. Furthermore, according to its etiology it can also be divided into typical and atypical pneumonia. Typical pneumonias are mainly caused by pathogens such as Streptococcus pneumoniae, *Haemophilus influenzae* and *Staphylococcus aureus*, as well as by viruses such as respiratory syncytial and influenza. On the other hand, atypical pneumonias, which include pathogens such as *Mycoplasma pneumoniae* and *Chlamydia pneumoniae*, present an added diagnostic challenge due to their less constant epidemiological cycles and the difficulty to detect them by standard methods.

For all these reasons, the diagnosis and treatment of pneumonia present unique challenges arising from the diversity of causative agents, the variability and non-specificity of its clinical presentation, the delays in obtaining reliable diagnostic results, and the growing threat of antimicrobial resistance. This diversity reflects the complexity of the diagnosis and treatment of the disease, which requires specific methods to identify

the pathogens involved and to establish the most appropriate intervention. Among the most common diagnostic indicators are X-ray tests and parameters such as levels of procalcitonin, C-reactive protein, lactate and coagulation markers that can provide information about the active infection process and its etiology. However, these strategies cannot always exclude the presence of a bacterial infection, and sometimes it is difficult to integrate the results with other clinical parameters to draw conclusions. In addition, the markers described and used for the adult population are often ineffective in the pediatric population. The development of tests based on the polymerase chain reaction has significantly increased our capacity to accurately diagnose known and emerging viral infections, in addition to enabling the detection of multiple pathogens in a single test. It should be noted, however, that molecular tests are less effective in the detection of bacterial infections, and in many health centers their detection is still based on bacterial cultures or less specific serologies. Recently, new biomarkers for different respiratory infections have been described using new generation technologies and advanced detection methods, which emerge as key tools to improve diagnostic accuracy. However, today, most cases of pneumonia are diagnosed only on the basis of the patient's clinical record, and a large proportion are diagnosed incorrectly. In addition, antibiotic treatment is administered before laboratory results confirming information on the etiology of the infection (e.g. viral or bacterial), contributing to the increase in antibiotic resistance of certain pathogens.

With the recent pandemic of COVID-19 (coronavirus disease 2019) caused by SARS-CoV-2, the scientific community has been searching for answers about the disease and the development of vaccines at an unprecedented speed. This disease, characterized by a wide variety of clinical manifestations ranging from asymptomatic cases to severe pneumonia and multiorgan failure, highlighted the need for a more in-depth study at the molecular level to understand its mechanisms of action, the associated risk factors and the biological differences involved in the severity of symptoms. The immune response to SARS-CoV-2 is a key factor in the clinical course of the disease, as it largely determines whether the virus can be effectively controlled or whether, on the contrary, severe complications are triggered. A timely and efficient activation of the innate response, mainly mediated by type I interferons, is essential to limit viral replication in the initial stages of infection. However, in severe cases, this antiviral response must be compromised, which allows the virus to spread and contributes to the excessive activation of systemic inflammation.

On the other hand, factors such as sex, age and the presence of comorbidities have a significant impact on the dynamics of this immune response, conditioning both the susceptibility to infection and the severity of the disease. In particular, recent studies point

USC UNIVERSITAT DE SANTIAGO

to significant immunologic differences between men and women, attributable to interactions between hormonal, xenogeneic and epigenetic factors, which may influence the ability to recognize and respond to the virus. Among the proposed associated factors is the TLR7 gene, which plays a fundamental role in the detection of viral RNA and in the activation of the innate immune response, promoting the production of type I interferons and other proinflammatory cytokines. This xen is located on the X chromosome, and several studies suggest that its location could prevent the characteristic silencing in one X chromosome genes in women, resulting in a low dose of its protein, which could explain part of the differences observed in the severity of COVID-19 between sexes.

In recent years, research on infectious diseases has undergone a paradigm shift, changing the focus to the study of the host response to infection. This change of perspective is fundamental to understand how the different pathogens interact with the immune system and allows us to identify factors that influence the progression or resolution of the disease. Traditionally, infectious diseases were analyzed mainly through the characteristics of the pathogens and the mechanisms of infection and the symptoms they produced. Nowadays, the focus on the specific response of the host has revealed how individual variability, including xenogeneic, epigenetic and immune factors, conditions the susceptibility, severity and evolution of infections, as well as the efficacy of treatments.

The recent development of new technologies in the field of omics allows a more complete and integrative approach to study pathogen-host interactions, facilitating the identification of key pathways, diagnostic or prognostic biomarkers and possible therapeutic targets. These technologies produce an enormous amount of complex and multidimensional data, whose interpretation requires the use of advanced computational tools. In this context, machine learning techniques emerge as fundamental elements for the efficient analysis and integration of these data, facilitating the identification of relevant patterns, relationships between biological variables and the creation of predictive models that can be applied in the diagnosis, prognosis and management of respiratory infections. The combination of these tools with omics approaches represents a revolution in the study of infectious diseases, allowing a deeper understanding of the underlying molecular mechanisms and promoting the development of more precise and personalized intervention strategies.

Objective

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

The aim of this study is to investigate the complex dynamics of respiratory infections from a host-centered perspective. This approach encompasses the study of key

aspects such as i) host-pathogen interactions and factors influencing disease progression and severity ii) the role of sex differences in immune response, including xenogeneic and epigenetic mechanisms, iii) the identification of susceptibility genes and biomarkers associated with the clinical course of infection, and iv) the development of innovative diagnostic tools based on gene expression using automatic learning algorithms to improve the accuracy and speed of current diagnostic methods. To this end, the following specific objectives were proposed:

(i) To analyze the causes of the differential immune response to COVID-19 by studying the transcriptome of patients with COVID-19 stratified according to the severity of the disease.

ii) To investigate the tissue-specific immune mechanisms that may underlie the severity of COVID-19 by studying the gene expression profiles in different biological samples (saliva, nasal epithelium and blood).

iii) To study the differential expression by sex and the epigenetic regulation of X-linked xenograft TLR7 in patients with COVID-19 stratified according to severity.

iv) To identify genetic variants associated with susceptibility to community-acquired pneumonia (CAP) by complete exome sequencing.

v) To evaluate the correlation between the identified genetic variants and the gene expression profiles in patients with pneumonia, assessing their potential functional impact on immune response and disease progression.

vi) To develop and validate a whole blood transcriptomic synthesis that allows to differentiate specifically between viral and bacterial pneumonia in children.

vii) To develop and validate a whole blood transcriptomic synthesis that allows differentiation between pneumonia caused by Mycoplasma pneumoniae and other pneumonia in children.

Methods

The methodology of the work integrates a multi-omics approach (genomics, transcriptomics and epigenomics) of respiratory diseases using different technologies and cohorts to study these conditions from a host-centered perspective. The research includes the analysis of blood, nasal epithelium and saliva samples from both patients with respiratory diseases and healthy controls, using technologies such as RNA-seq, microarrays and the NanoString nCounter platform to address specific questions related to immune response and the underlying mechanisms that determine the severity of the

disease. In addition, genomic and epigenomic studies were performed with the aim of identifying genetic variants and methylation patterns that influence the susceptibility and progression of these diseases. Machine learning algorithms, such as LASSO, were also applied to integrate and analyze the biological data, developing predictive models that optimize the diagnosis and classification of samples according to their etiology, in order to improve the accuracy of the current diagnostic tests.

RNA-seq technology was used to evaluate global gene expression in blood, which allowed us to analyze transcriptomic activity in different contexts of clinical severity and to study how immune responses vary in respiratory infections. The NanoString system allowed a focused analysis of a selected set of key immunological genes. Its design was used to study tissue-specific immune responses and to validate the observations of transcriptomic analyses performed with other technologies such as RNA-Seq or microarrays. DNA methylation analysis was used to study how epigenetic modifications affect the regulation of gene expression in patients with different severities of respiratory infections and to compare these methylation levels in differentially expressed genes according to biological sex. Finally, the whole exome sequencing was implemented to identify genetic variants related to susceptibility to community-acquired pneumonia.

These technologies allow us to generate a large amount of data in order to provide a holistic approach ranging from gene expression to underlying regulatory and transcriptomic mechanisms. This design allowed us to systematically address the interactions between genomic, epigenetic and transcriptomic factors in the study of respiratory infections, with special focus on pneumonia and COVID-19.

Results

Respiratory infections, and in particular pneumonia, continue to be a significant challenge for global public health, standing out as one of the main causes of mortality worldwide. This study explored the host response to respiratory infections through multi-omic approaches, generating new knowledge that could improve the diagnosis, prognosis and treatment of these diseases.

Host response to COVID-19: implications in severity

One of the main approaches was the study of the immune response to SARS-CoV-2 in blood, nasal epithelium and saliva samples, revealing that the immune response is tissue-specific and directly related to the severity of the disease. It was observed that robust local responses in the upper respiratory tract tissues, especially mediated by IFNs, are determinant in limiting the progression of the infection. In mild cases, the nasa tissues showed a strong antiviral activation, with high expression of IFN-induced genes, which

favors the local control of the virus. On the other hand, in severe cases, there was an insufficient antiviral response in the nasal epithelium and a higher systemic immune activation, characterized by an excessive neutrophil response and an abnormal activation of inflammatory and coagulation pathways in blood. In particular, saliva samples from severe patients revealed signs of viral spread to the lower respiratory tract, with an over-expression of genes involved in immune evasion and virus dissemination. These findings suggest that both nasal epithelium and buccal mucosa play a crucial role in the severity of COVID-19, opening opportunities for new diagnostic biomarkers and targeted therapies.

On the other hand, we studied the impact of biological differences between sexes on the immune response to COVID-19, focusing on the differential expression of the TLR7 xenon in different tissues. This xene, located on the X chromosome, showed a higher expression in women that could be providing a significant advantage in the response to infection compared to men. The lower expression of TLR7 in severe male cases was correlated with insufficient IFN-I responses, an essential component for temperate viral control. In addition, differences in TLR7 methylation patterns were observed between men and women, suggesting that epixenetic factors also contribute to these disparities. These findings underline the need to incorporate gender perspectives in the research and treatment of respiratory infections, allowing the personalization of therapeutic strategies.

Challenges in diagnosis, treatment and susceptibility to pneumonia: the answer lies in the host.

Genomic analysis revealed new variants associated with susceptibility to community-acquired pneumonia. In particular, a significant association was detected between pneumonia severity and variants in the FAM118A gene, whose decreased expression was observed in patients with severe pneumonia. Other candidates such as COL13A1, C20orf96, HEXIM1, NPAS4 and STRC were identified by SKAT-O collapse association tests. It is noteworthy that none of these genes were previously associated with pneumonia, which highlights the value of whole exome sequencing in the discovery of previously unknown genetic factors. In addition, the evaluation of the gene expression patterns for these six genes revealed a significant differential expression in patients with pneumonia compared to controls, suggesting a potential of these genes for diagnostic or prognostic purposes and opening an avenue for their study.

In relation to pediatric pneumonia, in this study we identified a transcriptomic signature of five genes (FAM20A, BAG3, TDRD9, MXRA7 and KLF14) that allows differentiating between bacterial and viral pneumonia with high accuracy (AUC > 0.92).

This signature, successfully validated in an independent cohort, stands out for its potential to improve the current diagnosis and reduce the unnecessary use of antibiotics, thus contributing to combat antimicrobial resistance and improve the clinical management of these infections.

In the case of pneumonia caused by Mycoplasma pneumoniae, transcriptomic analysis showed a very specific response to this pathogen, which allowed us to describe several transcriptomic signatures that discriminate with high precision this etiology from other types of pneumonia such as viral or caused by other bacteria that are not mycoplasmas. These signatures stand out for their high sensitivity and specificity, being particularly useful to guide treatment with macrolide antibiotics when beta-lactams are the first-line antibiotic administered by default.

These results highlight the importance of a host-centered approach and the use of multi-omics technologies in the study of infectious diseases and underline the clinical potential of the transcriptomic signatures and biomarkers identified. The findings of this study open the way to the implementation of faster, more accurate and accessible diagnostic tools, favoring a more personalized and effective approach to the management of respiratory infections. With the continuous evolution of sequencing and omic analysis technologies, as well as the development of new point-of-care transferable detection devices, these innovations are expected to transform the clinical management of respiratory infections, reducing the burden of these diseases and improving the course of these diseases in patients around the world. In addition, the use of different methods of gene expression study, such as RNA-Seq, microarrays and nCounter technology, give robustness to the findings of this work.

Conclusions

The results obtained from the analyses carried out during this thesis allowed the following conclusions to be reached:

(i) The severity of the disease at a systemic level may depend largely on the local immune response to COVID-19. The evidence highlights that the immune response in the nasal cavity, particularly the balance between interferon-mediated antiviral responses and inflammatory pathways, may determine whether the virus remains locally contained or spreads systemically, thus influencing the progression and severity of the disease.

ii) The oral cavity may play a key role in the systemic spread of SARS-CoV-2 infection. Transcriptomic analyses reveal that severe cases of COVID-19 show active viral replication and disruption of cellular functions in oral tissues, suggesting that the

oral cavity functions not only as an entry site for the virus, but also as a site that contributes to its systemic spread.

iii) Sex-specific differences in TLR7 expression and methylation patterns indicate an influence on the severity of COVID-19. Men, especially those with severe disease, consistently showed lower TLR7 expression compared to women. This reduction, driven by X-linked xenogeneic and epigenetic mechanisms, contributes to poor antiviral responses and a higher risk of severe outcomes, highlighting the key role of TLR7 in modulating disease severity.

iv) A 5-gene transcriptomic signature can accurately differentiate bacterial pneumonia from viral pneumonia and was validated in an independent cohort. This method has the potential to be translated to point-of-care diagnostic devices, providing a rapid and reliable tool to improve the clinical management of pediatric pneumonia and reduce the unnecessary prescription of antibiotics for viral infections.

v) Mycoplasma pneumoniae pneumonia alters host gene expression in a highly specific way, distinguishable by transcriptomic signatures. Combinations of 8 transcripts of 17 genes allow precise differentiation of M. pneumoniae pneumonia from other bacterial and viral pneumonias. Since M. pneumoniae is resistant to beta-lactam antibiotics, applying these signatures in clinical practice could improve treatment decision making and reduce the inappropriate use of antibiotics.

vi) Genetic variants associated with susceptibility to community-acquired pneumonia were identified, and new associations were observed in genes such as FAM118A, NPAS4, C20orf96, COL13A1 and STRC.

vii) Gene expression of the identified candidate genes is significantly deregulated in non-COVID-19 pneumonia and in severe COVID-19 pneumonia. This suggests their involvement in critical host-pathogen interactions and their possible contribution to the severity of CAP, supporting their investigation as possible biomarkers.

USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

1. INTRODUCTION

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

1. INTRODUCTION

The present thesis is framed within the theoretical context of human infectious diseases, that represent a global challenge to public health, with a particular focus on the impact of lower respiratory tract infections in the host. This work, constitutes a comprehensive analysis of this type of respiratory infections from a host-centric perspective using multi-omic approaches, examining the biology and immune response involved, the intricate host-pathogen interactions, the differences across demographic groups and biological tissues and providing useful tools to improve diagnostics and management of the disease. The structure follows a coherent and unified strategy throughout the chapters and articles presented that aims to expand the understanding of respiratory infections, with particular emphasis on pneumonia and coronavirus disease 2019 (COVID-19).

1.1. LOWER RESPIRATORY TRACT INFECTIONS

Respiratory tract infections include any infectious disease that affects the upper or lower respiratory tract. According to the World Health Organization (WHO) the majority of deaths from respiratory infections are caused by lower respiratory infections (LRTIs), predominantly pneumonia (1) that account for more than 2 million fatalities each year (2). These infections encompass a variety of conditions that primarily affect the lower parts of the respiratory system, such as the bronchi, bronchioles and alveoli. LRTIs include acute bronchitis, bronchiolitis and pneumonia (3, 4). They can be caused by a wild range of pathogens including bacteria, viruses, and some fungal species. Clinically, these infections imply significant challenges due to their potential severity, rapid progression, and the possibility of complications, such as respiratory failure, sepsis, and chronic lung disease. LRTIs are thus an important public health problem and a leading cause of illness and death in people of all ages, particularly affecting vulnerable populations such as young children, the elderly, and immunocompromised individuals.

1.1.1. Public health impact, implications and prevention strategies

Respiratory infections represent the single most common infections globally, contributing significantly to morbidity and frequently leading to the inappropriate use of antibiotics (5). They represent the fifth leading cause of death around the world and remain the most common

USC UNIVERSIDADE

cause of mortality in children under 5 years (1, 6). The epidemiological impact of respiratory infections is underscored by their high incidence and prevalence, notably higher in areas characterized by low-income status, inadequate vaccination coverage and malnourished population. It varies significantly between low- and middle-income countries (LMICs) and high-income countries (HICs), driven mostly by differences in healthcare infrastructure, disease burden, and access to medical care (7). Conversely , in HICs, chronic liver and renal conditions, previous hospitalization, low body mass index, male sex, and regular contact with children are considered main risk factors that increase the likelihood of acquiring respiratory infections (5). According to the last global burden disease study, in 2021, 344 million people were affected globally by LRTIs, with approximately 2.18 million deaths. This data is slightly lower than the estimates before the Coronavirus Disease 2019 (COVID-19) pandemic but this reduction was mostly due to COVID-19 associated interventions that not only reduce the transmission of Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2), but also the transmission of other pathogens that cause LRTIs (2, 8). However, despite this transient reduction of infections, the overall burden of LRTIs remains high, especially in LMICs. Consequently, LRTIs have a significantly economic impact, that includes a direct cost related to medical treatment, hospital stays, and medication, as well as indirect costs such as loss of productivity, long-term disability or disability-adjusted life-years, and premature death (9-12).

Surveillance and epidemiological studies highlight and agreed on the need for both the systematic collection of epidemiological and risk factors data as well as the enhancement of robust and coordinated public health strategies to monitor and control the spread of LRTIs in order to reduce the global burden of disease (13, 14) (15). Important preventive actions to control the impact of LRTIs include boosting integrated vaccination campaigns, lowering indoor air pollution, and educating people about good hygiene practices, all of which are particularly crucial for vulnerable groups (6). At the moment, vaccination against *Streptococcus pneumoniae*, *Haemophilus influenzae* type b, influenza, SARS-CoV-2 and monoclonal antibodies against respiratory syncytial virus (RSV) are the most widely promoted methods for preventing LRTIs (6). The pneumococcal conjugate vaccine (PCV) has proven to be highly effective in reducing the global burden of pneumococcal disease, both in morbidity and mortality following widespread vaccination campaigns (16, 17). The *Haemophilus influenzae* type b (Hib) vaccine has been also crucial in reducing severe infections caused by this pathogen, particularly in young children (18). Before the introduction of the Hib vaccine in the 1980 decade, this pathogen was a leading cause of childhood meningitis and pneumonia worldwide, especially in children under five (19). The influenza (flu) vaccines (that include inactivated vaccines, adjuvanted vaccines, intranasal vaccines, etc.) have been used for decades to protect against the seasonal flu. According to the WHO influenza vaccines reduce the risk of severe illness, hospitalization, and death, particularly in high-risk groups such as children, the elderly,

and individuals with chronic illnesses (20). The SARS-CoV-2 vaccines, such as the mRNA vaccines (Pfizer-BioNTech and Moderna) and viral vector vaccines (AstraZeneca, Johnson & Johnson), have had a transformative impact on global health since their emergency use authorization in 2020. The global impact of COVID-19 vaccination, particularly with booster doses, has been substantial in reducing the severity of disease and enhancing immunity against evolving variants of the virus (21-23). The rapid development and commercialization of these vaccines marks an unprecedented milestone of cooperation and efficiency within the international vaccine field. Recently, novel strategies aimed at mitigating the impact of respiratory syncytial virus (RSV) have been approved and implemented across various countries. Presently, the maternal RSV vaccine, administered to pregnant women between 32 and 36 weeks of gestation alongside newly developed monoclonal antibodies, such as Nirsevimab, constitute the primary interventions for preventing RSV infection (24, 25). Preliminary data from regions where these strategies have been deployed indicate a significant reduction in hospital admissions for RSV-related bronchiolitis and pneumonia, particularly among infants and adults over 60 years of age. Additionally, these interventions have shown promising results in reducing hospitalization rates and mortality (26-28). Further development and implementation of these and other preventive measures is expected to greatly reduce the burden of RSV-associated LRTIs, especially in young children (27, 29).

All of the mentioned vaccines and strategies have demonstrated their impact in reducing incidence, severity and health care cost associated with LRTIs, particularly in vulnerable populations, such as children, the elderly, and those with underlying health conditions (30), representing a hopeful strategy for the management of respiratory diseases.

1.1.2. Etiology and Pathogenesis of LRTIs

As mentioned before, respiratory tract infections or diseases can be caused by a wide variety of microorganisms including viruses, bacteria and fungi (31, 32). They can vary depending on clinical illness and epidemiological situation; the usual distribution of common agents for LRTIs in the past decades is summarized in **Table 1**.

The most important etiologic agents of severe lower respiratory illness include bacteria such as *Streptococcus pneumoniae* and *Haemophilus influenzae*, and viruses, such as RSV and influenza virus. Although the proportion of LRTIs deaths attributable to the individual etiologies differed by age group, globally *Streptococcus pneumoniae* account for most of the deaths across all age groups. Viruses are more common in mild upper and middle respiratory tract infections as well as in bronchiolitis in children, while bacteria are the main cause of pneumonia, especially in adults (33). Viral and bacterial syndromes can overlap and recent

USC DE COMPOSTELA

evidence suggests that some bacterial pneumonias may be secondary to viral respiratory tract infections (10, 33, 34).

Table 1. Main causal pathogens of LRTIs globally.

LRTI	Bacterial	Virus	Fungi
Pneumonia	Streptococcus pneumoniae *Haemophilus influenzae* *Mycoplasma pneumoniae* *Chlamydia pneumoniae* *Bordetella pertussis* *Legionella pneumophila* *Klebsiella pneumoniae* *Staphylococcus aureus* *Streptococcus pyogenes* *Coxiella burnetiid* *Gram-negative enteric bacilli*	Respiratory syncytial virus (RSV) Influenza virus (A/B) Rhinovirus Coronavirus (including SARS-CoV-2) Adenovirus Human Metapneumovirus (hMPV)	*Histoplasma capsulatum* *Coccidioides spp* *Aspergillus spp.* *Pneumocystis jirovecii* *Blastomyces dermatitidis* *Cryptococcus neoformans* *Paracoccidioides brasiliensis*
Bronchitis and Bronchiolitis	*Rare*	Respiratory syncytial virus (RSV) Influenza virus (A/B) Coronavirus (including SARS-CoV-2) Adenovirus Human Metapneumovirus (hMPV) Polyomaviruses Bocavirus	Rare

The pathogenesis of respiratory infections varies depending on the pathogen, the host's immune response, and environmental factors. Severe LRTIs, such as pneumonia and acute respiratory distress syndrome (ARDS), often involve pathogens that evade the immune system, leading to inflammation and tissue damage (32). Bacteria like *Streptococcus pneumoniae* produce virulence factors, such as capsular polysaccharides and specific toxins, which allow them to evade immune detection and cause extensive inflammation in the lungs (35). Viral infections, such as those caused by RSV, influenza or SARS-CoV-2, can also lead to severe immune responses. RSV induces inflammation and airway obstruction, leading to respiratory failure in young children (36), while viruses such as influenza or SARS-CoV-2 can cause cytokine storms, leading to lung damage and worsening symptoms, especially in adults (37).

The severity of respiratory infections depends on the pathogen's ability to induce inflammation but also heavily on the host's immune response, which is particularly weak in immunocompromised individuals and the elderly. Bacterial infections often cause extensive tissue damage and prolonged recovery, whereas viral infections are usually self-limiting but can lead to severe disease in immunocompromised individuals. In cases of viral-bacterial co-

infection, the initial viral infection weakens the immune system, facilitating bacterial colonization and worsening the clinical outcome, which highlights the complex nature of respiratory diseases pathogenesis.

1.1.3. Diagnosis and treatment of LRTIs

Significant progress has been made in the diagnosis of LRTIs over the past decades. However, diagnosis and identification of the causal pathogen is still an ongoing challenge, requiring significant improvements. Identifying the specific cause of LRTIs is often difficult when relying only on clinical evaluation, as symptoms may overlap with other respiratory conditions. The appropriate collection of samples to detect the causative pathogen is also important, considering both the testing method and the possibility of obtaining the necessary specimens in specific populations, such as children. Specimens can be collected through invasive methods (such as blood tests, thoracentesis, transthoracic needle aspiration, bronchoalveolar lavage via bronchoscopy, or protected specimen brush) or non-invasive methods (such as induced or expectorated sputum, nasopharyngeal or oropharyngeal swabs, or urine) (38).

Diagnosis has traditionally relied on microbial culture, antigen or antibody immunological methods, polymerase chain reaction (PCR) based tests and chest X-rays (39). The most common types of specimens in traditional microbial culture methods are blood, bronchoalveolar fluid and sputum (40). Main limitations of current diagnostic techniques include their time-consuming nature, low detection rates, invasiveness, and the challenge of obtaining difficult-to-access specimens, particularly in the pediatric population (39, 41). Additionally, it is often very challenging to detect atypical pathogens such as *Legionella*, *Mycoplasma*, and *Chlamydia* using traditional methods. Sputum samples, driven by the collection practice, are easily contaminated by oropharyngeal colonizing bacteria making difficult to determine the accuracy of the results (42). The development and improvement of PCR assays have noticeably increased our capability of accurately detect viral infections as it allows the interrogation of multiple viruses in a single test (43, 44). However, the detection of pathogens by antigen and/or antibody immunology and PCR are based on the genetic sequence of known pathogens, and unknown or not common pathogens cannot be detected in the common clinical practice in many centres (44). Furthermore, these methods primarily rely on detecting respiratory pathogens in specimens from sites distant to the infection, such as upper respiratory tract (URT) samples, blood, and urine. In these cases, they might be detecting nucleic acid from a recent illness or an asymptomatic colonization, especially in children (45), and, therefore, not the primary causative pathogen. Specific techniques for the diagnosis of respiratory infections

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

can vary depending on the suspected causal pathogens, the available resources and the type of respiratory infection and are described in more detail in the following sections.

Due to the risks of severe bacterial infectious diseases and the above-mentioned limitations in the diagnosis of LRTI, many guidelines recommend administrating antibiotic as a preventive measure in cases where there are diagnostic limitations, or other associated risks to the patient (45). Nowadays, antimicrobial resistance is a growing problem that causes millions of deaths globally every year. The misuse of antimicrobials, such as overprescribing, improper dosing, or selecting treatments that do not align with established guidelines, can significantly increase the risk of developing antimicrobial resistance. Most antibiotics are prescribed in outpatient clinics and acute LRTIs, including acute bronchitis, accounting for the majority of unnecessary prescriptions (46). It is estimated that by 2050, the annual mortality rate from *antibiotic resistance* will reach 10 million deaths, exceeding that of cancer and other major causes of death like diabetes or road traffic accidents (47).

For all these reasons, enhancing the diagnosis of respiratory diseases is essential, not only for improving disease diagnosis but also for enabling more accurate and targeted treatment. This advancement will help slow the rise of multidrug-resistant bacteria, mitigate their serious consequences, and reduce the costs associated with hospitalizations, diagnostic procedures, prolonged illness, and the overall burden on healthcare systems.

1.2. PNEUMONIA DISEASE

Pneumonia is a prevalent acute respiratory infection affecting the alveoli and distal airways of the lungs. The most common signs and symptoms are dyspnea, cough, fever, and emerging focal chest signs (48). It remains one of the most lethal LRTIs worldwide and represents a significant health concern, being a leading cause of morbidity and mortality across all age groups and sexes. It disproportionately affects vulnerable populations, including young children (<5 years), older adults (>65 years), and individuals with underlying health conditions (49). In particular, it is the leading cause of death in children under the age of 5. The incidence is higher in males than in females. Among patients who do not require hospital admission, the mortality rate is less than 1%. In contrast, short-term mortality (including deaths during hospitalization and within 30 days of admission) for hospitalized patients varies between 4% and 18%, and for patients in intensive care units mortality rates can escalate to as high as 50% (48). The epidemiology of pneumonia is continually evolving, driven by advancements in molecular diagnostics, novel treatments and vaccines, and public health interventions. Since 2,000, a series of respiratory infectious diseases pandemics, including influenza and COVID-19, have emerged as global health threats, significantly impacting epidemiological patterns

(49). The expenses associated to pneumonia are significant, and only a few strategies, such as shortening hospital stays, optimizing antibiotic use, and implementing vaccination programs, have effectively helped to reduce these costs (48).

1.2.1. Types of pneumonia

Pneumonia disease is generally divided into two different types: community-acquired pneumonia (CAP) or hospital-acquired pneumonia (HAP), which includes ventilation-associated pneumonia (VAP). The differences between CAP and HAP are influenced by the environment in which they are acquired (community *vs.* healthcare) and patient-specific risk factors, such as colonization in the respiratory tract and other pre-existing conditions (49). The causative microorganisms of CAP and HAP are also significantly different. CAP is often caused by pathogens like *Streptococcus pneumoniae*, respiratory viruses, and *Haemophilus influenzae*. While HAP is primarily caused by *Staphylococcus aureus*, including Methicillin-resistant Staphylococcus (MRSA), Enterobacter and gram-negative bacteria like *Pseudomonas aeruginosa*. These pathogens are typically acquired during hospital stays, where patients may be exposed to more resistant strains due to invasive procedures or prolonged antibiotic use (50). Consequently, management and treatment also vary between these two types of pneumonia. CAP cases are usually treated with broader-spectrum antibiotics such as amoxicillin and the treatment HAP often requires the use of potent or combination antibiotics, such as piperacillin-tazobactam, meropenem, cefepime, or vancomycin, to cover the wide range of resistant pathogens in severe cases (51, 52).

Regarding the etiology of pneumonia, the disease is commonly divided into “Typical” and “Atypical” pneumonia, although this designation is more a simplified framework rather than a reflection of prevalence differences. Typical pneumonia is caused by common bacteria (for instance *Streptococcus pneumoniae, Haemophilus influenzae*, and *Staphylococcus aureus*) whereas atypical pneumonia is caused by “atypical” pathogens (such as *Mycoplasma pneumoniae [M .pneumoniae]*, *Chlamydia pneumoniae*, and *Legionella species*) that are not detectable with Gram stain and cannot be cultured using standard methods. However, the prevalence and importance of these pathogens’ changes based on epidemiological cycles. *M. pneumoniae* is the most common pathogen of this type of pneumonia and the epidemiological cycles are represented by outbreaks each approximately 4-5 years (53). *M. pneumoniae* can cause from asymptomatic to severe pneumonia. What makes *M. pneumoniae* particularly fascinating is its remarkably small genome, containing just over 800 kb. This compact genetic information presents many challenges in understanding its biology. Continued research into the pathogenesis of *M. pneumoniae* has revealed that its interactions with the human immune system play a critical role in their pathogenesis and the development of disease (54). However,

USC DE COMPOSTELA

significant gaps remain unknown about how this pathogen synthesizes the specific molecules necessary for its survival and virulence. Finally, the most relevant characteristic influencing the treatment of *M. pneumoniae* infection is the fact that it lacks a cell wall, which makes it resistant to the most commonly used type of antibiotics, the beta-lactams (55).

1.2.2. Challenges in the diagnosis and treatment of pneumonia

The clinical presentation of pneumonia is diverse and may mimic other respiratory conditions such as asthma, bronchiolitis, pertussis, lung abscess, bronchiectasis, or malaria (49, 56-61). Besides, the clinical evaluation of pneumonia generally does not differentiate between bacterial, viral, or mixed etiologies and it is based only on clinical symptoms, making it difficult for healthcare providers to determine the most appropriate treatment without further diagnostic support. This is particularly important because the management of bacterial pneumonia typically involves antibiotic prescription and in most cases the treatment is initiated before the final diagnostic contributing to the improper use of antibiotics and promoting bacterial resistance. In pediatric pneumonia, it is common that all patients with clinical findings suggesting bacterial pneumonia received empirical antibiotics without further test (62). Despite efforts to enhance international guidelines (56, 57, 63), the variability in causative microorganisms and symptoms imply a significant challenge for clinicians in healthcare centres (61, 64-67).

Particularly, the diagnosis of pneumonia typically involves a combination of clinical assessment, imaging, and microbiological testing. One of the most widely used diagnostic method is chest X-ray. However, they may not always detect early or mild cases of pneumonia, leading to potential delays in treatment, especially in resource-limited settings. Additionally, chest X-ray are limited in their ability to differentiate the etiology of the disease (i.e., viral, bacterial) and other common test such as pneumococcal antigen in urine, PCR in respiratory swabs, or serological tests are usually time-consuming and produce delays in obtaining results (68). Some effort has been made to look for specific biomarkers that could help in the diagnosis and prognosis of pneumonia, including the differentiation between viral and bacterial infections. Blood biomarkers such as C-reactive protein (CRP) and procalcitonin (PCT) have been described to be effective in differentiating bacterial from viral infections in adult populations (56, 69).

Regarding pediatric pneumonia, there are additional limitations. One major challenge is the difficulty in obtaining representative samples such as sputum from children, which in some cases obstructs accurate diagnosis. Additionally, many biomarkers promising for diagnostic and monitoring purposes in adults are less effective in children and cannot reliably differentiate

USC DE SANTIAGO DE COMPOSTELA

viral from bacterial pneumonia (70, 71). Thus, biomarkers levels in adults may not correlate in the same way with the disease severity or the presence of infection in the pediatric population. Furthermore, the immune response in children differs from that of adults, which can affect the effectiveness of diagnostic markers in each case, underscoring the necessity for targeted research focused on the pediatric population to enhance diagnostic precision and ensure the implementation of appropriate clinical interventions.

In the particular case of *M. pneumoniae*, that cause the atypical pneumonia, cultures are usually the gold-standard for bacterial diagnosis but its diagnosis suppose an additional challenge due to the challenge in obtaining a representative sample, the slow growth and the low sensitivity (72). Serology for *M. pneumoniae* is another common approach (72), but tests show low specificity, and Immunoglobulin M (IgM) can remain positive for months after the infection (73). There has been a gradual shift towards the replacement of traditional methods by PCR for the diagnosis of *M. pneumoniae* pneumonia, especially after COVID-19 emergency (74). PCR is a rapid method which offers high sensitivity and specificity compared to serology and culture, and is currently widely-used in HICs. Specific diagnosis of *M. pneumoniae* infection is essential as this pathogen lacks a cell wall making it inherently resistant to beta-lactam antibiotics such as amoxicillin, which target bacterial cell wall synthesis and are the first line antibiotics most commonly used in the clinical practice. As a result, standard treatments for typical CAP are ineffective against *M. pneumoniae*, necessitating the use of alternative antibiotics, such as macrolides, tetracyclines, or fluoroquinolones. Thus, clinicians usually start with empirical antibiotic treatment and adjust it later based on microbiological test results (75) but they still not fast and accurate enough. Continued research into the genetic and molecular characteristics of *M. pneumoniae* is crucial for developing novel approaches for diagnosis, treatment, and prevention.

1.3. SARS-COV-2 OUTBREAK AND COVID-19

SARS-CoV-2 is the novel coronavirus responsible for the global pandemic known as COVID-19. It belongs to the broader coronavirus family, which includes other well-known viruses such as those responsible for the common cold, Severe Acute Respiratory Syndrome (SARS), and Middle East Respiratory Syndrome (MERS). SARS-CoV-2 first emerged in late 2019 in Wuhan, China and its rapid transmission marked the beginning of a global health crisis (76). It is characterized by its highly contagious nature, primarily transmitted through respiratory droplets, which facilitated its rapid spread worldwide. The pandemic lead to a range of public health measures, including lockdowns, travel restrictions, and social distancing. Sadly, due to its ability to cause severe respiratory illness, the virus led to millions of

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

hospitalizations and deaths overwhelming healthcare systems globally and changing the daily life for individuals across the globe.

The unprecedent development and availability of different vaccines in a really short period of time coupled with rigorous public health measures implemented across thousands of countries has played a crucial role in rapidly controlling the COVID-19 outbreak (77). Public health initiatives, such as social distancing, mask mandates, and extensive testing, have further contributed to lowering transmission rates also reducing the circulation of other common pathogens. In addition, the rapid development of molecular tests and diagnostic methods has been instrumental in the notable decrease in hospitalizations and fatalities worldwide. The integration of SARS-CoV-2 detection into PCR tests allowed for highly accurate and early detection, facilitating the identification of infected individuals even in the absence of symptoms, which helped isolate cases before further spread. Other significant advancement was the development of antigen tests, including lateral flow assays (LFAs), that even if they are not as sensible as PCR test, provided a faster and cheaper alternative for large-scale testing without the need for healthcare facilities (78). These adaptations in diagnostic technology, coupled with coordinated data-sharing platforms and global collaborations, have been key to limiting the spread of COVID-19, enabling quicker identification of infected individuals, controlling outbreaks, facilitating timely interventions, and ultimately reducing the pressure on healthcare systems worldwide.

COVID-19 typically produce influenza-like symptoms, but its clinical presentation can vary significantly ranging from being completely asymptomatic to causing severe conditions, including pneumonia, extensive lung injury, and progressive respiratory failure (79, 80). Unlike other viruses that have caused global pandemics (such as influenza A H1N1), SARS-CoV-2 disproportionately affects adults more frequently than children (81, 82).

A global effort has also been made to decipher the pathogenesis of SARS-CoV-2 and the reasons behind the varying outcomes and severities of COVID-19. It is known that this virus enters human cells primarily through the binding of its spike protein to the ACE2 receptor located on the host's surface of cells (83). The host immune response varies significantly between individuals and the imbalance between innate and adaptive immune responses is one of the main factors responsible for severe disease course. Severe cases are characterized by showing a reduced number of lymphocyte T subpopulations (CD4+ and CD8+). Usually, immune response involves the activation of type I interferons (IFN-I)**,** which are crucial for controlling viral replication in the early stages. However, in severe COVID-19 cases, there is often an impairment of the type I IFN response, leading to delayed viral clearance and allowing the virus to spread further (84). In this case, as the infection progresses, the immune system

USC DE COMPOSTELA

becomes increasingly activated leading to an overproduction of pro-inflammatory cytokines, a phenomenon known as "cytokine storm" (37). This excessive inflammatory response contributes to tissue damage, especially in the lungs, and is associated with severe disease outcomes, including ARDS, or multi-organ failure leading to high mortality rates in the severe cases (37). In this particular field, transcriptomic studies have significantly contributed to our understanding of the disease, the pathogenesis and the host specific immune response to SARS-CoV-2 (85-88). However, there are still important aspects that remain poorly understood regarding the virus and its effects on the host. Importantly, there are derived or associated conditions, such as long COVID-19 and Multisystem Inflammatory Syndrome in Children (MIS-C), that highlight the complex consequences of COVID-19 and required further investigation and understanding.

1.4. THE STUDY OF HOST RESPONSE TO INFECTIOUS DISEASES: A CHANGE OF PARADIGM

The study of host responses to infectious diseases is pivotal to understanding how different pathogens interact with the immune system and to identify factors that contribute to disease progression or resolution. Historically, infectious diseases were examined primarily from the pathogen perspective, mechanisms of infection, and the presentation of the diseases they caused, but a shift towards the host study has revealed critical insights about how individual variability, such as genetic, epigenetic, and immune differences, shapes disease outcomes and how host's immune response often determines the susceptibility, severity and progression of infections, as well as the effectiveness of treatments (89). The recent development of different omic technologies in the past decades has enable a more balanced and holistic approach to investigate how pathogens interact with the immune system and other biological functions of the organism, promoting at the time the identification of key pathways, biomarkers, and potential therapeutic targets (90).

1.4.1. Host Immune response to infectious diseases

The human immune response to defense against antigenic molecules is primarily mediated through two distinct defense mechanisms: the innate and adaptive immune systems. The innate immune system is characterized by the physical and mucosal barriers present in the body's surfaces such as the respiratory, gastrointestinal, and reproductive tracts. Key components include a variety of cells that are rapidly recruited to the site of infection upon recognition of pathogens, providing an immediate but non-specific response to prevent the spread of the infection. Thus, the innate immune system consists of all the immune defenses that lack immunologic memory. A traditionally accepted characteristic of innate responses was that they remain unchanged regardless of how often the antigen is encountered. However, more

recent research has revealed that certain innate immune cells have the ability to "remember" past encounters with pathogens, thereby enhancing their responsiveness upon re-exposure. This phenomenon, known as trained immunity, represents a form of immune memory within the innate immune system (91). Innate immunity primarily involves macrophages, innate lymphoid cells (ILCs), natural killer (NK) cells, dendritic cells (DC), and granulocytes (a diverse group of white blood cells encompassing neutrophils, basophils, and eosinophils), **Figure 1** (92).

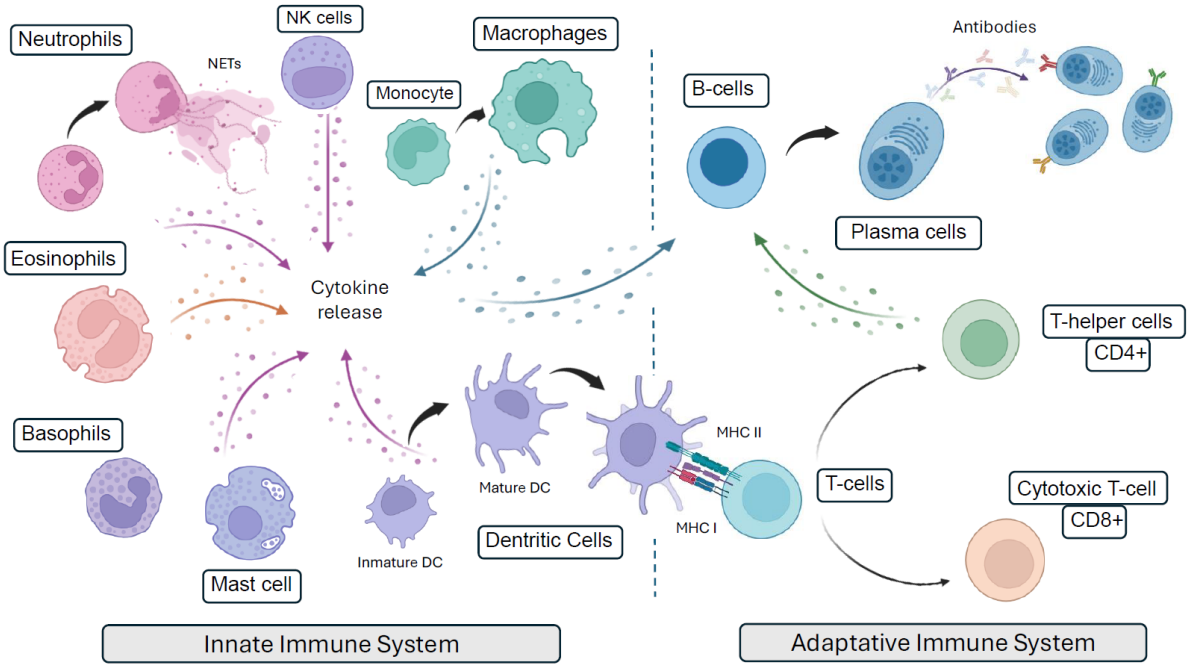


Figure 1. Schematic representation of the Immune System.
Self-made figure using BioRender.

Macrophages are among the first responders activated upon the recognition of pathogens. These cells detect foreign microorganisms through specific surface receptors called Pattern Recognition Receptors (PPRs) such as Toll-like receptors (TLRs), that trigger their activation and the release of cytokines, initiating the inflammatory response. Neutrophils, also rapidly migrate to the site of infection, where they phagocytize pathogens and release pro-inflammatory mediators, enzymes, reactive oxygen species and neutrophil extracellular traps (NETs) to destroy pathogens. Beyond their role in pathogen elimination, neutrophils are involved in the regulation of inflammation and immune homeostasis (93). DC mature upon pathogen recognition and act as a bridge between the innate and adaptive immune systems by capturing antigens and presenting them to T cells, thus initiating adaptive immune responses. Among the most important cytokines released after antigen recognition by these innate immune cells are

interferons of the IFN-I family, particularly IFN-α and IFN-β. The IFN-I pathway induces the expression of IFN-stimulated genes, which play key roles in blocking viral replication, enhancing immune cell recruitment, regulating the inflammatory response, and triggering downstream cascades that help contain the virus before adaptive immunity becomes fully engaged (92). NK cells, on the other hand, do not rely on antigen presentation; instead, they recognize and kill infected or abnormal cells, such as those infected by viruses or cancerous cells, by releasing cytotoxic substances. Finally, ILCs play an essential role in mucosal immunity, particularly in the defense against infections in epithelial barriers, like the gut, by producing cytokines that regulate the immune response, maintaining tissue homeostasis and initiating responses without direct pathogen recognition (92, 94).

The adaptative immune system is partially activated by the innate immune response and is characterized by its specificity. This system generally confers long-lasting protective immunity against reinfections by the same pathogen, which is known as immunological memory. Key components of the adaptative immune system are the lymphocytes, which provide lifelong immunity following exposure to the antigen either through infection or vaccination. There are two main types of lymphocytes: B cells and T cells. B cells are primarily responsible for the production of antibodies (immunoglobulins), which recognize and neutralize pathogens such as bacteria and viruses. These antibodies can also form immune complexes with antigens, facilitating their clearance through processes like phagocytosis and activation of the complement system (92, 95). T cells are further subdivided into two main types: cytotoxic T cells (CD8+) and helper T cells (CD4+). Cytotoxic T cells directly kill infected host cells by recognizing antigens presented on the surface of these cells. Helper T cells, on the other hand, activate and coordinate the immune response by releasing cytokines that stimulate other immune cells, such as B cells, macrophages, and cytotoxic T cells. The recognition of antigens by T cells is mediated through the major histocompatibility complex (MHC) molecules, which present fragments of the pathogen to T cell receptors. MHC molecules are classified into two types: MHC class I, which presents antigens to CD8+ cytotoxic T cells, and MHC class II, which presents antigens to CD4+ helper T cells (92) (**Figure 1**). Dysregulation or depletion of T cells during the infectious process, especially of T helper cells, can lead to immune system failure and conditions like lymphopenia, a state characterized by a reduced number of T cells in the bloodstream (85).

Thus, the immune response to an infection involves various well-coordinated steps, encompassing both innate and adaptive immunity for initiating a targeted and coordinated immune response while also ensuring immunological memory for long-term protection (95).

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

1.4.2. Sex differences in the immune response to infectious diseases

There is substantial evidence of differences in the immune response of human males and females against bacteria, virus, parasites, allergens and self-antigen. Even if males and females possess the same immune cells, it has been reported that females develop higher innate, humoral and cellular immune responses to infections and vaccines (96, 97). This phenomenon has been attributed to a variety of biological factors, including hormonal, genetic, and epigenetic differences between sexes. For instance, females have higher levels of immunoglobulins and B cells activity, allowing them to produce faster and more effective antibody responses against pathogens (97). This immunological advantage also extends to vaccine responses, where females tend to develop higher antibody titters and greater immunological memory after vaccination compared to males. However, it also implies higher susceptibility to develop, for example, autoimmune diseases and more adverse reactions to vaccination (98). Thus, diseases such as rheumatoid arthritis, systemic lupus erythematosus, and multiple sclerosis, are significantly more common in females than in males (99). Particularly, one of the genes proposed to contribute to this susceptibility to autoimmune diseases is the Toll-Like Receptor 7 (*TLR7*) gene (100). TLR7, along with other TLR proteins participates in the recognition of extracellular nucleic acids that are internalized by immune cells ending up in an endosome. Endosomal TLRs like TLR7 may then detect the internalized single strand RNA (ssRNA) of the virus leading to the activation of signaling pathways via myeloid differentiation primary response 88 (MyD88); **Figure 2**. This signaling cascade subsequently promotes the expression of inflammatory genes, such as genes encoding for cytokines and chemokines, through the activation of NF-κB (nuclear factor kappa-light-chain-enhancer of activated B cells) and also activates interferon regulatory factors (IRFs), leading to the production and release of type I interferons (IFN-I), such as IFN-α (101); **Figure 2.**

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

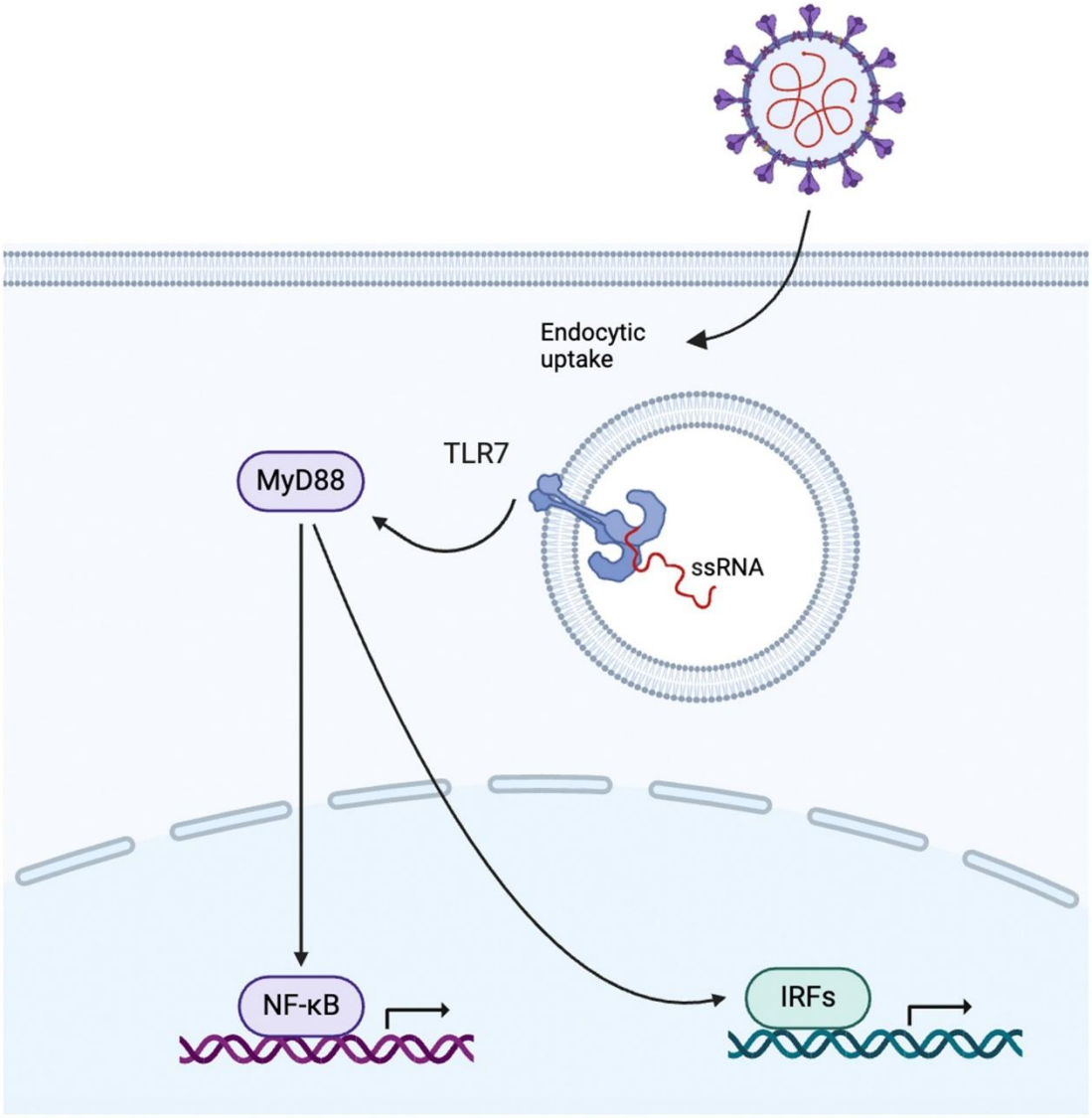


Figure 2 Simplified representation of TLR7 activation and signalling.
Self-made figure created with BioRender.

Interestingly, deficiencies in *TLR7* and other errors in the IFN-I pathway, have also been found to be associated with susceptibilities to specific infectious diseases (102). Genetic studies were the first approach that suggested the role of *TLR7* in the differential sex incidence of COVID-19 (102, 103). One particularly large study showed that carrying a rare predicted loss-of-function or missense variant in the *TLR7* gene was a risk factor for developing severe disease (104).

However, the correlation between these genetic findings and the expression of *TLR7* in response to infectious disease, as well as potential epigenetic modifications, remains poorly understood. *TLR7* is linked to the X chromosome (**Figure 3**). To prevent the double dosage of

X-linked genes in females compared to males, one of the X chromosomes in females undergoes random inactivation during the early stages of embryonic development. This process of X-chromosome inactivation (XCI), results in a mosaic pattern of gene expression, where some cells express genes from one X chromosome while others express genes from the other (105). However, certain X-linked genes, particularly those located in the pseudo autosomal regions (PARs) that are involved in pairing between the X and Y chromosomes, may escape this inactivation. This phenomenon of XCI is not absolute, has inter-individual variability, and can also differ between tissues (106) such as in specific immune cells (100). As the *TLR7* gene is located in a highly probable XCI escape region (107) (**Figure 3)**, it is hypothesized that it could have a biallelic product in females and a monoallelic in males.

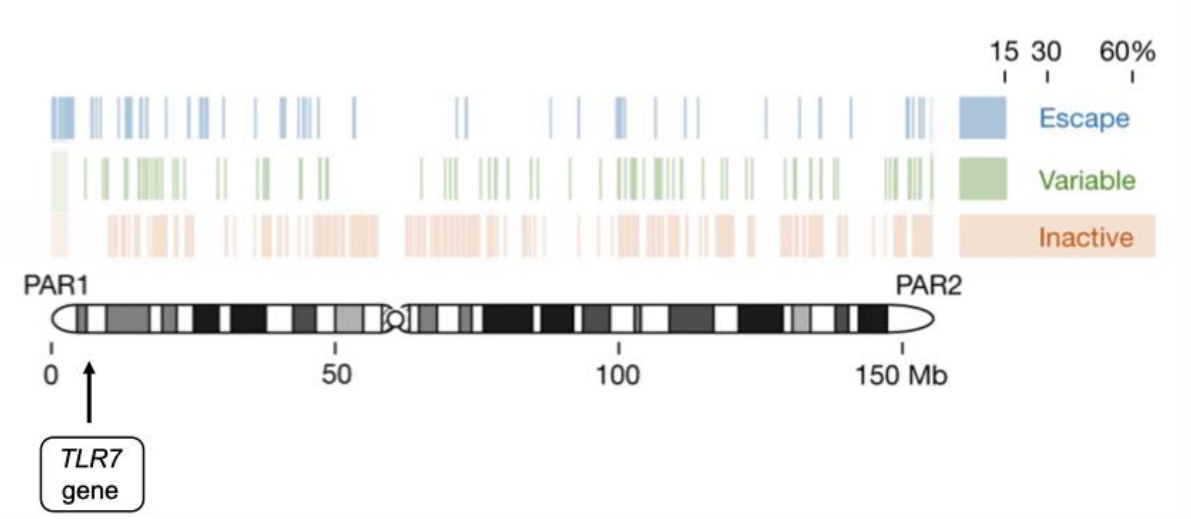


Figure 3 Human X-Chromosome Inactivation scheme.
Modified from Tukiainien et al.(107)

In COVID-19 disease, disparities in outcomes and severity between sexes have been also evident since early stages of the pandemic, with higher mortality and hospitalization rates in male COVID-19 patients compared to female (108). Some of the proposed hypotheses for these differences include: *i)* changes in *ACE2* expression between sexes; *ii)* varieties in immune pathways like TLR7 and IFN-1 (102, 109); and *iii)* the more exaggerated cytokine response in males that can potentially lead to a "cytokine storm" and severe organ damage (110). Given the proposed differences in TLR7 dosage, its relationship with the interferon pathway and inflammatory processes, along with its X-chromosome location of the gene, *TLR7* expression emerges as an important candidate to further investigate differences in COVID-19 disease severity between females and males.

USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

1.4.3. Omic approaches to the study of infectious diseases

Currently, there is a wide range of "omics" fields that represent the high-throughput study of different biological molecules or systems. The development of large-scale sequencing and other high-throughput technologies enables to quantify the highly dimensional and interactive molecular layer systems that are implicated in biological processes (111). Each type of omic layer focuses on studying a different aspect of genetic material reflecting different stages of genetic evolution or transformation (DNA, RNA, proteins, etc.). Most wildly studied omics include genomics, transcriptomics, proteomics, metabolomics and epigenomics (**Figure 4**).

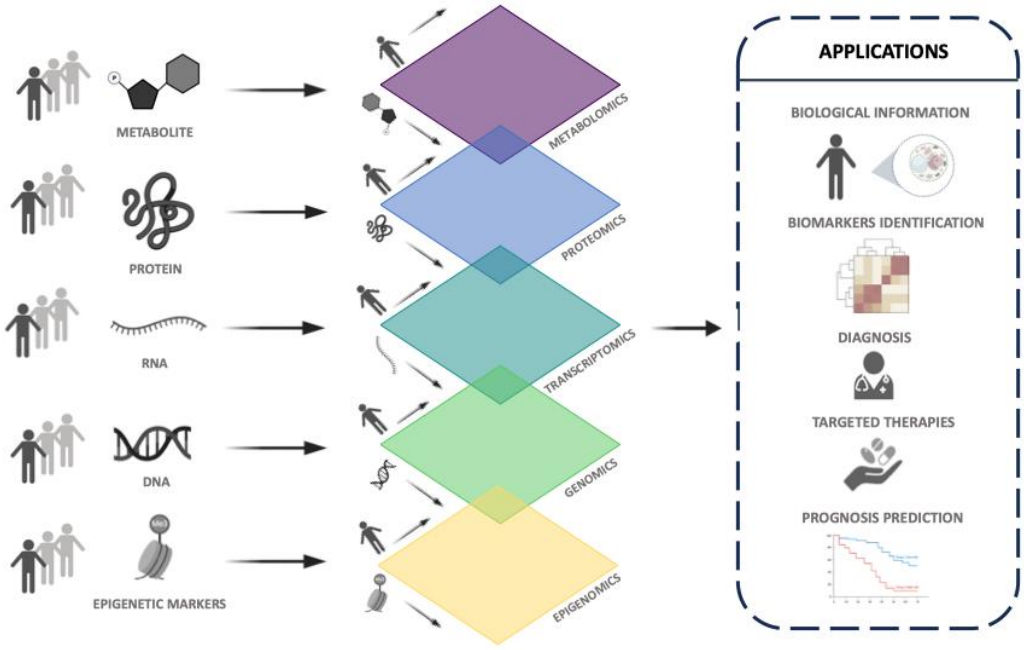


Figure 4 Multi -omic approaches and its application in the context of human diseases. Self-made figure created with BioRender.

Genomics focuses on the "original" DNA sequence, epigenomics deal with the structural and chemical modifications to the DNA molecule, such as DNA methylation and histone modifications as well as other mechanisms that regulate gene expression including chromatin remodeling, microRNAs (miRNAs) and non-coding RNAs (ncRNAs). Transcriptomics studies different types of RNA molecules in a specific cell type or tissue at a certain developmental stage or under a specific physiological condition to decipher the expression patterns of genes. Finally, proteomics studies the proteins and their interactions, while metabolomics focuses on the determination and quantification of metabolites in a biological system (112). All of them have become powerful tools that enhance our understanding of different infectious diseases and

their associated molecular processes, as well as offering potential for the development improved diagnostic, prognostic, and therapeutic solutions (**Figure 4**).

Next-generation sequencing (NGS) has revolutionized the analysis of genetic material, providing comprehensive insights into both DNA and RNA. This technology allows for rapid sequencing of entire genomes and transcriptomes, allowing researchers to explore genetic variations, gene expression patterns, and epigenetic regulation processes at an unprecedented scale and speed, providing valuable insights into human biology and disease.

1.4.3.1. Host transcriptomics for biomarker discovery in infectious diseases

Gene expression studies allow us to deep into the link between the genotype and the phenotype of an organism. There are different methods and approaches that can be employed to investigate gene expression depending on different factors that include available budget, research goal, urgency and the RNA type of interest (mRNA, miRNA, ncRNA, etc.). Gene expression technologies enable the study of human biology through two main approaches: targeted or hypothesis-free strategies. Targeted strategies are employed when specific genes of interest or genes implicated in a particular biological mechanism are already identified. This is particularly useful for testing or validating targets in specific contexts. For such purposes, technologies like RT-qPCR, nCounter, or RNA microarrays could be the more suitable option (113). On the other hand, untargeted approaches, such as Bulk-RNA sequencing (RNA-seq) and single-cell RNA-seq, allow for the study of almost the entire transcriptome of an organism or cell, providing a very detailed view of gene expression under specific conditions. RNA-seq is gaining increasing popularity and is rapidly replacing microarrays due to its ability to identify novel transcripts, isoforms, analyze complex alternative splicing events, study allele-specific expression, and enable meta-transcriptomic studies, allowing the sequencing of RNA molecules from different species coexisting in the same ecosystem simultaneously (114, 115). It is the preferred option for the identification of biomarkers or genes that may not have been previously considered (116). RNA-seq is especially useful in discovery-driven research, often referred to as hypothesis-free analysis, as it offers an unbiased approach that allows to explore the "complete picture" before formulating more specific research hypotheses.

The decrease in sequencing costs over the past few years, along with the reduction in microarray prices, has made RNA expression profiling more accessible. This has opened up new opportunities for studying a wide range of clinical conditions, with RNA-seq emerging as a promising tool for the study of human diseases, including respiratory tract infections. The exploration of host transcriptomics in infectious disease research has gained significant traction in recent years, improving our understanding of host-pathogen interactions and facilitating the

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

creation of valuable tools for disease stratification, diagnosis and prognosis based on the expression of specific transcripts in response to an infection. Numerous studies have focused on identifying host gene expression biomarkers and signatures associated with various infectious diseases, such as tuberculosis (117-119), H1N1 (120), RSV (121), rotavirus (122), and malaria (123) as well as biomarkers to differentiate between bacterial and viral infections in febrile patients (116, 124-128). However, many challenges remain in diagnosis of specific syndromes such as pneumonia, or the identification of less common pathogens that are particularly difficult to detect and are often diagnosed late, such as *M. pneumoniae*.

Despite significant advancements in transcriptomic research and the growing understanding of host-pathogen interactions, there remains a gap between the knowledge generation in this field and its translation into clinical practice. While the potential for using gene expression profiling in disease diagnosis and prognosis is clear, the development of practical and accessible tools that can implement this knowledge in real-world clinical settings has been slower. Thus, there is an urgent need for the development of point-of-care (POC) devices that can simplify diagnostic procedures, particularly crucial in resource-limited environments, where access to large, expensive equipment and specialized personnel is often restricted. Filling this gap would improve not only disease management but also enhance the global access to advanced diagnostic technologies, making them more widely available for clinicians and patients in everyday scenarios (116).

1.4.3.2. Study of tissue-specific host response to infections

Each type of tissue or specimen provides unique advantages and disadvantages that can significantly influence the outcomes of research. Analyzing multiple tissue types can offer a more comprehensive understanding of the interplay between different tissue responses and disease mechanisms, facilitating the identification of more effective biomarkers for diagnosis, prognosis and treatment in infectious diseases.

Whole blood samples are often used in clinical settings due to their accessibility. Blood specimens offer a systemic overview of the disease, containing extensive information, as includes circulating immune cells, cytokines, and other soluble factors that can reflect systemic inflammatory responses (129). For instance, peripheral blood cells share transcriptomic patterns with other major organs, and it is estimated that more than 80% of the transcriptome is shared with brain, colon, heart, kidney, liver, prostate, spleen or stomach (130). However, while blood can provide insights into the general immune status of an individual, it may not capture localized responses to respiratory infections that occur in a specific manner in other tissues, such as the lungs or airway mucosa.

USC DE SANTIAGO DE COMPOSTELA

Saliva represents a promising and valuable tissue type for analysis. It is particularly advantageous because it can be collected non-invasively, making it a convenient option for frequent sampling. Saliva contains various biomolecules, including proteins, antibodies, and nucleic acids, which can be indicative of local and systemic health. In the context of respiratory diseases, saliva can provide insights into the inflammatory processes occurring in the oral and throat regions (131). However, it may be less informative regarding deep lung pathologies due to dilution and the complex oral microbiome, which can complicate biomarker detection. In addition, saliva is a highly complex sample because it contains various enzymes, proteins, and other biological substances that can affect the accuracy of results. These components, such as salivary amylase**,** proteases**,** and ribozymes, can potentially interfere with the analysis if not handled properly (111).

Nasopharyngeal samples are crucial for understanding respiratory diseases, especially those affecting the upper respiratory tract. These samples allow for direct assessment of the mucosal immune response and can provide valuable information about the presence of pathogens, inflammatory markers, and immune mediators (132). The advantage of nasopharyngeal sampling lies in its proximity to the respiratory system, which can yield more relevant data regarding local responses to infections or other respiratory conditions.

Despite the importance of the previously mentioned it is important to note that all these samples contain a complex mixture of mRNA expressed by a variety of cell types and for certain research questions it would be necessary to examine the transcriptome at the single-cell level in order to avoid diluting specific responses from individual cell types due to the presence of mRNA from other cell populations (confounding effects of cellular heterogeneity) (133).

1.4.3.3. Machine Learning strategies to enhance diagnosis and prognosis in infectious diseases

Although omic data hold great potential, using them to investigate phenotypic diversity is not always a straightforward task. Omic datasets are intricate, highly dimensional and often affected by experimental noise and various technical or biological confounding factors, which can hide relevant signals within unrelated data. A widely used approach to separate these signals and extract meaningful patterns is through the application of machine learning (ML) techniques.

ML algorithms have the ability to learn from omic data, capture generalizable patterns, and reduce the complexity of the original measurements which makes them very useful tools to deal with all the biological information and complex datasets generated by next-generation sequencing technologies, such as transcriptomic data. By analyzing gene expression profiles,

USC DE COMPOSTELA

ML models can predict disease phenotypes, identify biomarkers, and generate diagnostic and prognostic signatures that may guide clinical decision-making. (134). For instance, well-established ML applications are already widely applied in transcriptomic data for the identification of diagnostic and prognostic signatures for conditions such as cancers and infectious diseases, where transcript-based signatures have proven very useful for patient stratification and outcome prediction (88, 126, 135, 136). The ML techniques are generally divided in two major categories: supervised and unsupervised learning algorithms. Supervised learning, often used in biomedical contexts, identifies patterns based on labelled examples of past data while unsupervised learning algorithms aim to identify patterns without needing already previous information about the real phenotype or outcome. Supervised learning is the preferred option for biomarker discovery and transcriptomic signatures development, especially in the medical context.

In the context of omic data for classification and prediction, a major challenge is the high dimensionality of the data, where the number of variables (genes or transcripts) far exceeds the number of samples. This imbalance necessitates a careful approach to variable selection, as retaining only the most informative features is essential to prevent overfitting and improve model interpretability. Feature selection and regularization techniques have become central tools in transcriptomics for reducing dimensionality, enabling more robust and clinically applicable models (see methods section for more detail). These techniques allow for the development of models and biomarker signatures that can generalize effectively, capturing essential patterns while minimizing noise.

Given the inherent complexity of omic data, effective ML pipeline may help to integrate data preprocessing, normalization, and rigorous validation methods, ensuring that models capture relevant biological signals and remain generalizable.

USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

2. HYPOTHESIS

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

2. HYPOTHESIS

Studying respiratory infections from the host's perspective is crucial for understanding how human biology influences their pathogenesis and severity. Variations in host genetic factors and molecular responses to lower respiratory tract infections offer valuable insights into the nature of these infection, their progression, variability in clinical outcomes among individuals, and susceptibility to the disease. In this context, several hypotheses are proposed to explore how host-specific responses, whether through gene modulation, identification of blood biomarkers, or the influence of genetic and epigenetic factors, can offer new tools for accurate diagnosis, differentiation of infection etiology, and guidance in more targeted treatments and prevention strategies in LRTIs.

A host-based, multi-omic approach hold promise for addressing these critical aspects, contributing to the development of innovative tools and strategies to improve the clinical management and patient outcomes.

General hypotheses:

i) The modulation of host gene expression in response to lower respiratory tract infections varies across different tissues. This variability may influence disease outcomes and provide information about the infection status; **Addressed in the first article (section 5.1 of results)**

ii) The gene expression and epigenetic regulation of the X-linked *TLR7* gene may drive differential immune responses to LRTIs between females and males, potentially contributing to disease severity; **Addressed in the second article (section 5.2 of results).**

iii) Genetic variation is associated with susceptibility to respiratory infections. The implicated genes may influence disease severity, and their analysis could provide valuable insights into the genetic basis of infection outcomes; **Addressed in the third article (section 5.3 of results**).

iv) Syndrome-specific host-transcriptomic responses in whole blood offer a more accurate method for distinguishing viral from bacterial pneumonia compared to

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

other biomarkers and previously published gene expression-based signatures; **Addressed in the fourth article (section 5.4 of results**).

v) Pathogen-specific host transcriptomic responses in whole blood can aid in diagnosing and guiding therapeutic decisions for atypical pneumonia caused by *M. pneumoniae*; **Addressed in the fifth article (section 5.5 of results**).

USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

3. OBJECTIVES

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

3. OBJECTIVES

This thesis aims to investigate the complex dynamics of lower respiratory tract infections from a host-centered, multi-omic perspective. It focused on key areas, including host-pathogen interactions, the influence of sex on infection response, the identification of susceptibility genes, epigenomics as modulator agent of gene expression, and the development of innovative diagnostic tools using transcriptomics and machine learning-based approaches

3.1. GENERAL AND SPECIFIC OBJECTIVES

i) To study the variability of gene expression response of immune-related genes in different human tissues of COVID-19 patients and its relevance to disease outcomes; **Addressed in the first article (section 5.1 of results).**

- To study the causes of the differential immune response to COVID-19 analyzing the gene expression of immune-related genes in COVID-19 patients stratified by disease severities.

- To investigate tissue-specific immune mechanisms that may be underlying severity of COVID-19 studying gene expression profiles from different biological samples (saliva, nasal epithelium and whole blood).

ii) To investigate sex-differences in the immune response of COVID-19 patients from transcriptomic and epigenomic perspectives; **Addressed in the second article in section 5.2 of results**).

- To investigate the sex-biased expression and epigenetic regulation of the X-linked *TLR7* gene in COVID-19 patients stratified by severity.

iii) To explore genetic factors underlying susceptibility to community-acquired pneumonia (CAP) and their potential role in the disease's pathogenesis; **Addressed in the third article (section 5.3 of results**).

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

- To identify genetic variants associated with susceptibility to CAP using whole exome sequencing.

- To assess the correlation between identified genetic variants and gene expression patterns in pneumonia patients, assessing the potential functional impact on the immune response and disease progression.

iv) To develop and validate whole blood transcriptomic signatures that improve the diagnosis and treatment of pediatric pneumonia allowing to identify the etiology of the infection; **Addressed in the fourth and fifth articles (section 5.4 and 5.5 of results**).

- To develop and validate in an independent cohort a whole blood transcriptomic signature that allows to specifically differentiate between viral and bacterial pneumonia in children.

- To develop and validate in an independent cohort a whole blood transcriptomic signature that allows to specifically differentiate *Mycoplasma pneumoniae* pneumonia from all other pneumonias in children.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

4. METHODS

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

4. METHODS

4.1. STUDY DESIGN AND SAMPLE SELECTION

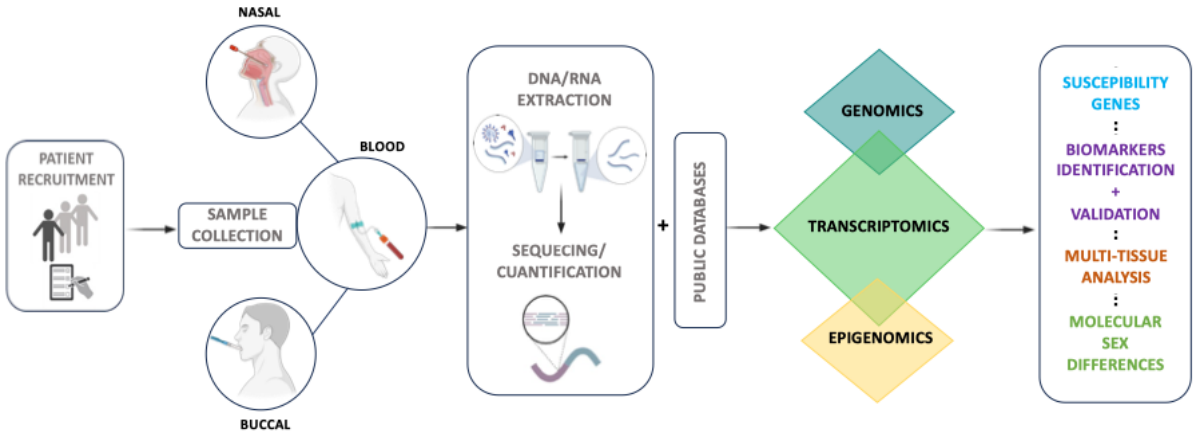


Figure 5. Graphical representation of the study design of the thesis.
Self-made figure created with BioRender.

All new generated data from human patients were obtained thought harmonized procedures for patient recruitment, classification, clinical data and sample collection, processing and storage across the participating centers. For every patient, a prospective and standardized case report form was used to record patient demographic characteristics, presence of comorbidities, and clinical signs and symptoms at the time of the initial presentation at the healthcare center. Informed consent forms were obtained from all participants, or their guardians, prior to sample collection. All required ethical approvals were secured by each participating country's Ethics Committee.

All data used in the different studies of this thesis were collected under the umbrella of national and international consortiums or downloaded from the NCBI Gene Expression Omnibus (GEO) or ArrayExpress public repositories. In particular, COVID-19 adult samples from blood, nasal epithelium and saliva (**Figure 5**) used in the analysis of article 1 and 2 of this thesis were recruited through the GENCOVID study group (www.gencovid.eu) at Hospital

Clínico Universitario de Santiago de Compostela (Galicia, Spain) between March 2020 and May 2020. COVID-19 was defined according to the national guidelines in Spain (https://www.mscbs.gob.es/profesionales/saludPublica/ccayes/alertasActual/nCov/documentos.htm). Main Inclusion criteria was positive PCR for SARS-CoV-2 and severities were stratified in mild (when managed as outpatient), moderate (when patients were hospitalized) and severe (when patients were admitted to intensive care unit [ICU]). In addition, other transcriptomic data from convalescence COVID-19 patients were downloaded from Lynn Laboratory BitBucket (https://bitbucket.org/lynnlab/covid-sa) (137) and Jackson et al. (87), and microarray transcriptomic data from viral and bacterial infections used for comparisons were downloaded from GEO repository (**Table 2**). Epigenomic DNA methylation data from COVID-19 patients were downloaded from GSE179325 GEO dataset derived from Barturen et al. study (138).

Adult saliva samples from patients with CAP were collected through the CAPPRIC study, a multi-centre active surveillance study aiming at evaluating the main characteristics of CAP in patients older than 18 years (139) . Other transcriptomic data used in article 5 were downloaded from GEO and ArrayExpress (https://www.ebi.ac.uk/biostudies/arrayexpress) public repositories (**Table 2**). The inclusion criteria for the CAPPRIC study selected patients older than 18 years with confirmed diagnosis of pneumonia by chest radiography at one of the 24 Spanish primary care centers.

Pediatric blood samples used in the studies focused on the discovery of biomarkers and transcriptomic signature for the diagnosis of viral/bacterial pneumonias and *M. pneumoniae* pneumonia (articles 3 and 4) were either recruited through European Consortiums or downloaded from GEO repository (**Table 2**):

- European Union Childhood Life-threatening Infectious Diseases Study (EUCLIDS-https://www.euclids-project.eu/); GA279185.
- PErsonalised Risk assessment in Febrile illness to Optimise Real-life Management across the European Union (PERFORM-https://www.perform2020.org/); GA668303.
- Diagnosis and Management of Febrile Illness using RNA Personalised Molecular Signature Diagnosis (DIAMONDS- https://www.diamonds2020.eu); GA848196.

For these pediatric studies, pneumonia disease was defined for patients that match the following criteria: i) inflammation of one or both lungs: lobar or segmental or multilobar consolidation on chest X-ray; ii) clinical symptoms compatible with acute respiratory infection and following radiological findings of consolidation/pleural effusion: alveolar consolidation (dense or fluffy opacity that occupies a portion or whole of a lobe or of the entire lung that may or may not contain air-bronchograms) or pleural effusion (fluid in the lateral pleural space and not just in the minor or oblique fissure) that was spatially associated with a pulmonary

parenchymal infiltrate (including other infiltrate) or obliterated enough of the hemithorax to obscure an opacity.

Table 2. Public databases used for the analysis

Identifier	Repository	Technology and Platform	Samples	Study
MTAB-10926	ArrayExpress	Illumina Novaseq 6000	COVID-19 adult patients + Healthy Controls	**ARTICLE 1,2 & 3**
COVID-SA	Bitbucket	Illumina Novaseq 6000	Convalescent COVID-19	**ARTICLE 1**
GSE64456	GEO	Illumina HumanHT-12 V4.0 expression beadchip	Viral and bacterial infections + Healthy Controls	**ARTICLE 1**
GSE40012	GEO	Illumina HumanHT-12 V3.0 expression beadchip	Adult CAP + Healthy Controls	**ARTICLE 1 & 3**
GSE42026	GEO	Illumina HumanHT-12 V3.0 expression beadchip	Viral and bacterial infections + Healthy Controls	**ARTICLE 1**
GSE60244	GEO	Illumina HumanHT-12 V4.0 expression beadchip	Viral and bacterial infections + Healthy Controls	**ARTICLE 1**
GSE72829	GEO	Illumina HumanHT-12 V3.0 expression beadchip	Viral and Bacterial infections + Healthy Controls	**ARTICLE 1**
GSE69529	GEO	Illumina HiSeq 2500	Viral and Bacterial infections + Healthy Controls	**ARTICLE 1**
GSE103119	GEO	Illumina HumanHT-12 V4.0 expression beadchip	Pediatric CAP + Healthy Controls	**ARTICLE 3,4 & 5**
GSE42834	GEO	Illumina HumanHT-12 V4.0 expression beadchip	Adult CAP + Healthy Controls	**ARTICLE 3**
GSE65682	GEO	Affymetrix Human Genome U219 Array	CAP adult samples + Healthy Controls	**ARTICLE 3**
GSE179325	GEO	Methylation EPIC BeadChip	COVID-19 adult samples + Healthy Controls	**ARTICLE 2**

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

4.2. TRANSCRIPTOMIC ANALYSIS

4.2.1. Sample collection and processing.

Table 3. New generated expression data used for the analysis

Study	Source	Type of Sample	Platform	Sample size	
ARTICLE 1 ARTICLE 2 & ARTICLE 3	GENCOVID study	Nasal epithelium	nCounter	n= 49	n=38 (COVID-19 patients)
					n=11 (Healthy Controls)
		Saliva	nCounter	n= 53	n= 41 (COVID-19 patients)
					n=12 (Healthy Controls)
		Whole blood	nCounter	n= 54	n=41 (COVID-19 patients)
					n=13 (Healthy Controls)
ARTICLE 4	EUCLIDS, PERFORM & DIAMONDS consortia networks	Whole blood	HiSeq 4000 (discovery cohort)	n= 192	n= 154 (Pneumonia patients) / n= 40 (Bacterial)
					n=9 (Viral)
					n= 38 (Healthy Controls)
			NovaSeq 6000 (Validation cohort)	n= 32	n= 22 (Bacterial)
					n= 10 (Viral)
ARTICLE 5	EUCLIDS, PERFORM & DIAMONDS consortia networks	Whole blood	HiSeq 4000	n= 81	n= 8 (Controls with asthma)
					n= 73 (pneumonias w/o pathogen)
			NovaSeq 6000	n= 41	n= 9 (*M. pneumoniae pneumonia*)
					n= 22 (Bacterial pneumonia)
					n= 10 (Viral pneumonia)
					n= 35 (Healthy Controls)

For whole blood samples recruited within the three European projects (EUCLIDS, PERFORM, DIAMONDS) RNA-seq with Illumina technology was performed. Whole blood was collected from patients and healthy controls into PAXgene blood RNA tubes (PreAnalytiX) or Tempus™ RNA tubes, stored at −80 °C and RNA was isolated according to each manufacturer's instructions (**Table 3**).

For the analysis including new patients with COVID-19 and matched healthy controls (**Table 3**), the nCounter technology (Nanostring) was employed to investigate gene expression patterns of immune-related genes. Each biological sample was collected using the appropriate buffers and tubes: *i*) blood samples were collected using PAXgene Blood RNA Tubes (PreAnalytiX) and RNA was isolated using PAXgene blood miRNA extraction kit (Qiagen); *ii*) saliva and nasal epithelium samples were collected in Oragene CP-190 kit (DNAGenotek); RNA from saliva and nasal epithelium specimens was isolated using the RNeasy microkit

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

(Qiagen); *iii*) nasopharyngeal samples were collected in Universal Transport Medium tubes (UTM) tubes to check for the presence as well as the viral load of SARS-CoV-2 at the same time as collection of blood, saliva, and nasal epithelium samples. Molecular detection of viral particles was performed through a multiplex real-time PCR using the Allplex™SARS-CoV-2 Assay (Seegene).

4.2.2. Gene expression platforms

As described above, different technologies were utilized to infer gene expression information from the biological samples of the selected individuals (microarray chips, bulk RNA-seq and nCounter technology), the interpretation and normalization of these data imply small variants for each case that are described following.

4.2.2.1. Microarray data processing and analysis

Microarray expression data is obtained from a chip containing predefined synthetic nucleic acid molecules (probes) that hybridize with the DNA or cDNA of the sample of interest which is labelled with a fluorescent dye (140). Raw fluorescence intensity values are generated for each probe on the microarray chip and represents the amount of labelled cDNA that was bound to each probe, which correlates with the expression level of the corresponding gene or transcript.

Raw data from each dataset was pre-processed, annotated and normalized using the adequate R package for each specific platform and chip: *illuminaHumanv4.db* and *illuminaHumanv3.db* and *lumi* (141-143) for Illumina® microarrays data and the package *Oligo* (144) for Affymetrix® datasets.

From the three adult datasets of pneumonia samples (article 5), the R package COCONUT (COmbat CO-Normalization Using conTrols) was used to combine all microarray datasets for the meta-analysis and reduce batch effects as it use the healthy controls to pre-normalize de data (145). COCONUT is an unbiased co-normalization method that assumes that all healthy controls across studies come from the same statistical distribution, estimating first correction factors from each dataset's healthy control samples, and then applying them to the disease samples in each dataset. This procedure removes technical differences while still retaining within-dataset differences between groups. In all the analysis implying microarray data, R package *limma* was used to infer differentially expressed genes between groups (146) and moderated t-statistic was applied. The linear models were fitted considering the age and sex as categorical covariates when necessary.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

4.2.2.2. RNA-seq data processing and analysis

Bulk RNA-sequencing is a high-throughput sequencing technology that directly detects the nucleotide sequences of cDNA molecules. Library preparation for RNA-seq typically begins with total RNA sample preparation followed by a ribosome removal step in most of the cases. The total RNA sample is then converted to cDNA and fragmented. Adaptors are added to these fragments for subsequent amplification. After PCR amplification, the cDNA library is ready for undergo the following sequencing cycles. Illumina technology employed in the present analysis use sequencing-by-synthesis (SBS) chemistry, where fluorescently labelled terminator nucleotides are incorporated one base at a time (in each cycle of sequencing). High-resolution optics capture images of each cycle, enabling accurate base calling across millions of DNA fragments simultaneously (147) (**Figure 6**).

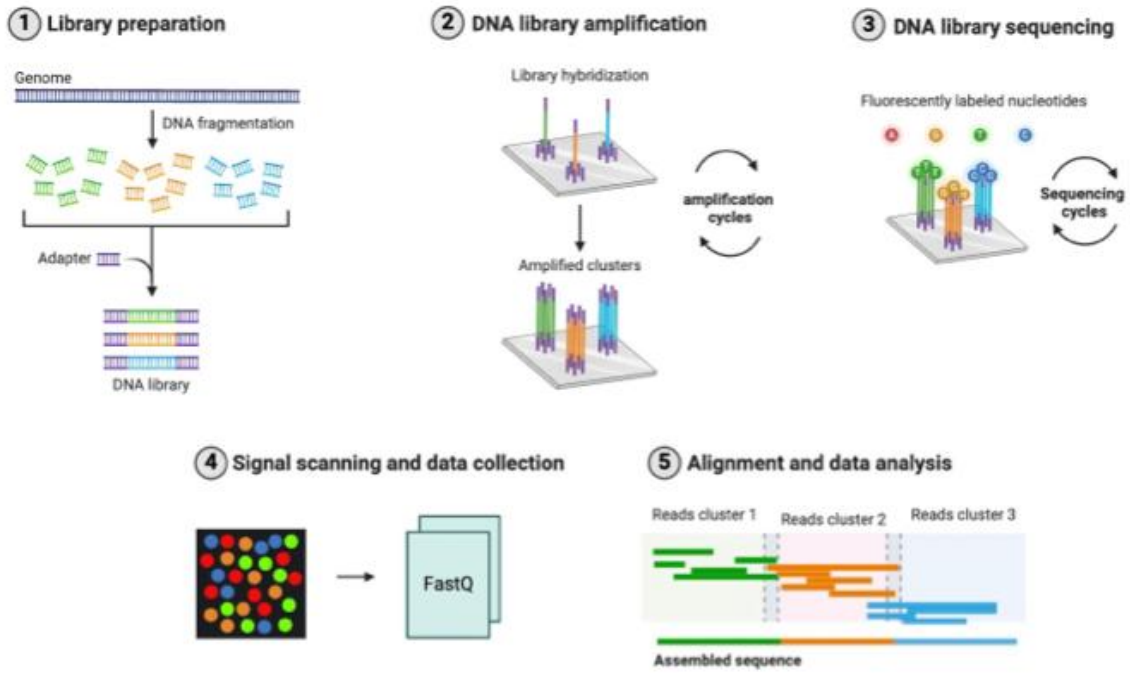


Figure 6 General workflow for Illumina Sequencing.
Created using BioRender

Raw data is presented as unmapped reads in FastQ format, which consist of a text file that contains the nucleotide sequence with the quality information of each nucleotide base in addition to other technical data useful to detect quality of the sequencing and possible outliers. These sequences need to be aligned with the corresponding reference genome (i.e. GRCh38.p14) to perform the quantification of the gene expression data and subsequent normalization. Normalization between samples is a necessary step because in RNA-seq the number of read counts is associated with several factors including sequencing depth (148), gene length and mRNA expression level and the GC-content (149), between others.

In the present thesis, RNA sequencing was conducted using two different Illumina protocols and platforms. For the EUCLIDS cohort, RNA sequencing was conducted on a HiSeq 4000 platform, with library preparation and sequencing of 30 million 75 or 100 bp paired-end reads. The Illumina's TruSeq RNA Sample Preparation Kit was used for library preparation, and ribosomal and globin RNA depletion was performed using the Illumina® Ribo-Zero Gold kit. For the PERFORM/DIAMONDS cohort, after RNA isolation, an additional DNAse treatment was carried out with the RNA clean & concentrator kit (Zymo Research) prior to sequencing. After quantification and a normalization step, a strand specific library preparation was completed using NEBNext® Ultra™ II mRNA kit (New England Biolabs) and NEB rRNA/globin depletion probes following manufacturer's recommendations. Individual libraries were normalized using Qubit, pooled together and diluted. The sequencing was performed using a 150 paired-end configuration in a Novaseq6000 platform. For all RNA-seq data, quality control of raw data was carried out using *FastQC* (150), alignment and read counting were performed using *STAR* (151), alignment filtering was done with *SAMtools* (152) and read counting was carried out using *FeatureCounts* (153).

RNAseq data was processed for batch correction using control samples and *COMBAT-Seq* R package. The *RUVg* method implemented in the *RUVSeq* package of R (154) was employed to identified other factors of unwanted variation in the dataset. Normalization differential expression (DE) assessment and covariate correction were carried out using *DESeq2* R package (154, 155). In addition, the models were fitted including known covariates such as sex, batch, and type of collection tube, to correct the differences in gene expression related to these factors. Differentially expressed genes (DEGs) were identified through the negative binomial distribution implemented in the *DESeq2* package (155). A generalized linear model was fitted, and a *t*-statistic was calculated for each gene. *P*-values were corrected for multiple testing using the Benjamini-Hochberg False Discovery Rate (FDR).

4.2.2.3. NanoString nCounter data processing and analysis

The nCounter system is a direct digital molecular detection technology developed by Nanostring. This multiplex system directly measures the actual quantity of target molecules present in the sample. The nCounter system can simultaneously measure hundreds of targets in a single reaction, providing high sensitivity and specificity, even with low expression levels or low quantity in the input sample. Additionally, it does not require a prior enzymatic reaction such as PCR or RT-PCR, minimizing the likelihood of false positives and technical biases in the results (**Figure 7**).

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

The digital detection uses a pair of unique probes for each molecule of interest, consisting of a reporter probe with an specific 6-color code (Reporter CodeSet) at its 5' end, and a capture probe with biotin at its 3' end (Capture ProbeSet), which are hybridised to the target RNA and it is used to bind the hybridization complex (streptavidin-biotin affinity) to the cartridge allowing the reading of the colour codes for each target molecule on the nCounter device (**Figure 7**) . Each colour combination uniquely identifies a target molecule, allowing for the identification of up to 800 targets per panel.

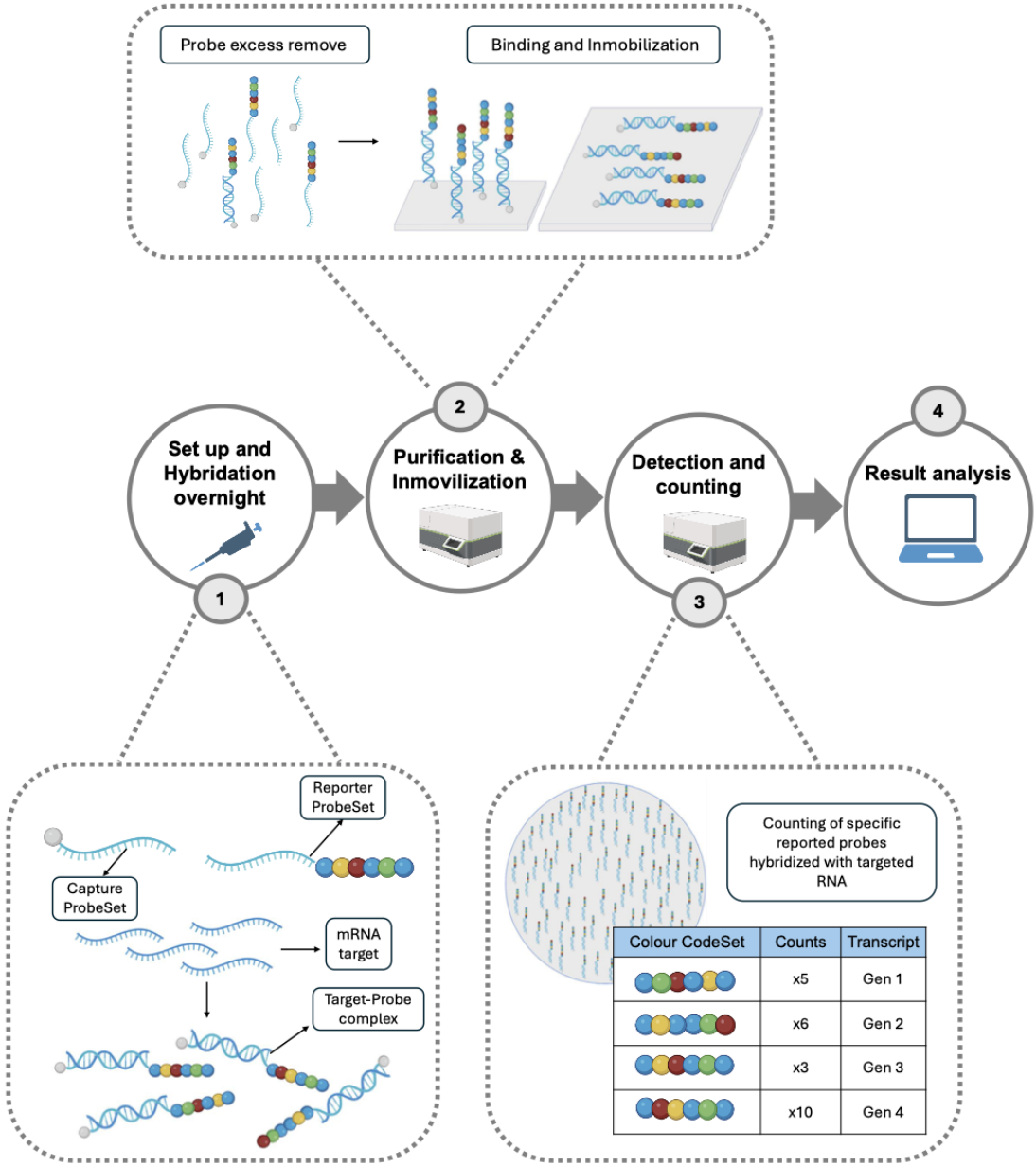


Figure 7 Nanostring technology pipeline.
Self-made figure created with BioRender

In the present study, the Immunology Panel was selected to analyze 594 genes involved in the main immunological pathways (including major classes of cytokines and their receptors, chemokine ligands and receptors, IFNs and their receptors, the TNF-receptor superfamily, and the KIR family genes). Gene expression patterns of different tissues from COVID-19 patients and controls were evaluated using this panel through the *SPRINT* nCounter system from NanoString Technologies.

The standard gene expression protocol for 12x RNA hybridization with 5ul of RNA as input and a hybridization time of 18h was used for the assay. Control and COVID-19 samples were mixed in the plates to avoid technical sample bias. After a quality control check of the raw data analyzing different parameters recommended by the manufacturer, data were normalized using the optimal House Keeping (HK) genes selected through the *GeNorm* algorithm (156) implemented in the R package *CrtlGene* (157) in order to correct for quantity of the RNA loaded into the reaction. We excluded HK candidate genes with less than 50 raw counts in any of the samples compared from the reference selection analysis. Data normalization was performed through a combined approach using both *DESeq2* and *RUVSeq* as described Bhattacharya et al. (158). Genes with counts below the background (mean + 2 standard deviations [SD] of the negative control spikes in the code set) were not considered for the differential expression (DE) analysis. Differential expression analysis was carried out with the *DESeq2* R package. After data normalization and background thresholding, lower expressed genes were eliminated from the list of DEG. Finally, age and sex were also added as known covariables to correct the model.

We used the *CIBERSORTX* (159) tool to input immune cell types abundance from nCounter data in different patient categories, including mild and healthy controls given that real cell counts were not available from clinical data. We employed a leukocyte gene signature matrix (160) obtained from microarray data as reference and applied a B-mode batch correction to account for cross-platform variation and 500 permutations for the analysis. We carried out a Wilcoxon test to assess statistical significance between categories.

4.2.3. Unsupervised learning methods

Unsupervised learning encompasses a range of statistical methods designed to explore and uncover patterns in multivariate without predefined labels or outcomes. Unsupervised learning discovers hidden patterns, structures, and relationships within datasets without the need to create predictive models. This approach allows for the visualization of relationships between variables and the identification of groups of similar observations or individuals. Among the most widely used unsupervised learning techniques are principal component analysis (PCA) and clustering analysis (161).

USC DE SANTIAGO DE COMPOSTELA

4.2.3.1. Principal component analysis

Principal Component Analysis (PCA) is a statistical technique commonly used for data pre-processing, feature extraction and dimensionality reduction as it can extract the most informative features from large and multivariate datasets while preserving the most relevant information from them. PCA analysis summarizes the information content of large datasets into a smaller set of uncorrelated variables known as principal components (PC). These principal components are linear combinations of the original variables that have the maximum variance compared to other linear combinations. In transcriptomic datasets, when performing PCA, the axes represent linear combinations of expression values (i.e., counts). The projection of the samples onto the first axis maximizes the variance across all possible axes. The second axis maximizes the variance of all axes that are orthogonal to the first component, and this continues for the subsequent components. This ensures that the first component explains the greatest amount of variability. Ideally, the condition of interest (i.e. case/control) should be the one showing the biggest proportion of variance, and therefore represented by the PC1. Even if an elevate number of different PCs are calculated, most of the variation is usually retained in the first 1-10 PC being significantly higher in the first two components (PC1 and PC2) that are the represented ones in the analysis described in this thesis (**Figure 8**).

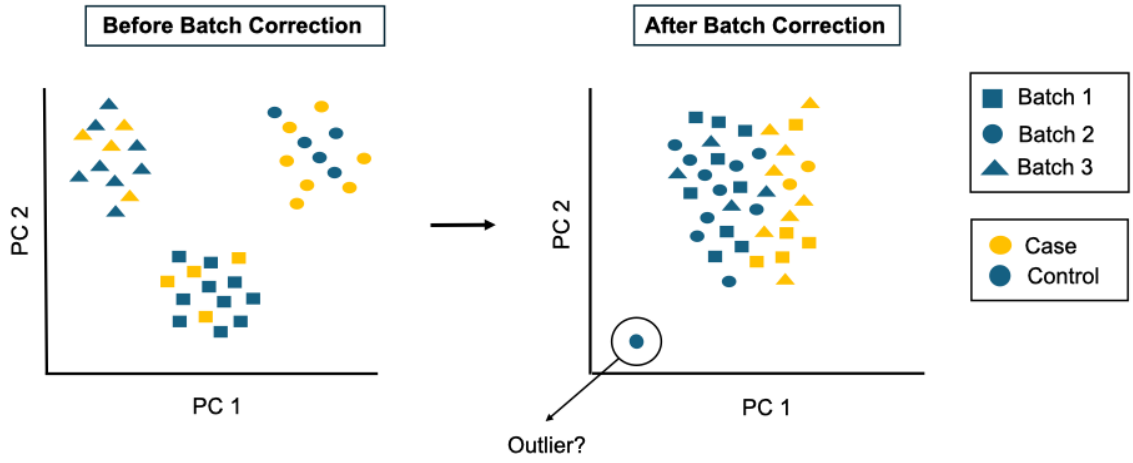


Figure 8 Simplified example of a PCA visualization in 2 dimensions before and after batch correction of the data. Self-made figure

In these studies, PCA was used to visualize the global transcriptome patterns of both RNA-seq, microarray and nCounter data in order to verify the quality of the raw data, the normalization and the batch effect correction processes as well as the detection of outlier samples. Additionally, by focusing on the most variable transcripts (*i.e.* the 500 most variable), PCA can help to highlight the biological factors driving the observed patterns. PCA analysis was conducted using *PCAtools* package of R.

4.2.3.2. Hierarchical clustering analysis and Heatmap

Hierarchical clustering is other common approach to interpret and visualize transcriptomic data by grouping genes and samples based on their expression patterns. Unlike dimensionality reduction techniques like PCA, hierarchical clustering heatmaps directly display the expression levels of selected genes across samples, enabling the identification of relevant patterns and clusters.

In the present analyses, gene clusters were generated first by filtering transcripts based on their log_2 fold change (Log_2FC) and adjusted p-value adjusted from the DE analysis, retaining only those with significant differential expression. Hierarchical clustering was then performed on the normalized expression values to group genes and samples according to their similarity. Heatmaps were created using the *ComplexHeatmap* package in R (162), which provides advanced customization options, such as detailed annotations, gradient color schemes, and dendrograms to illustrate clustering relationships facilitating data interpretation.

4.2.4. Supervised Learning

Unlike unsupervised learning, supervised learning relies on predefined classes, which are determined by humans and represent a finite set of labels assigned to the data. The algorithm's primary goal is to analyze the labeled data, uncover patterns, and build mathematical models that can make accurate predictions on new unseen data (161). In biological and medical research this approach is widely used for classification and regression, where predicting specific outcomes is essential.

4.2.4.1. Machine learning for biomarker and signatures discovery

ML algorithms can identify patterns to learn from omic data and reduce the complexity of the datasets with the information derived from biological tissues. The original datasets are usually composed by tens of thousands of measurements producing highly dimensional information that require sophisticated bioinformatics pipelines and mathematical algorithms to be processed and adapted for each specific aim of study.

In the context of modelling omic data using supervised learning for classification or prediction objectives, such as diagnostic applications, it is essential to reduce the number of variables to the minimum necessary. Usually, the main challenge in this type of analysis comes from the fact that the number of variables (genes, transcripts, proteins, etc.) is much larger than the number of subjects or biological samples (163). The reduction of variables that do not bring

USC DE COMPOSTELA

important information to the analysis is essential to allow the development of useful and precise models for predictions in a real scenario. Furthermore, in the investigation of diagnostic signatures, it is it is often found that models utilizing fewer variables can yield comparable performance to more complex models and this reduction not only enhances interpretability but also simplifies the translation of findings into a rapid and manageable diagnostic method. By minimizing the number of variables required for accurate predictions, the resulting models become more user-friendly and facilitate quicker clinical application.

In order to develop any prediction model or identifying diagnostic biomarkers within transcriptomic data there are some crucial points that should be carefully considered to simplify the outcome and prevent overfitting. First, and as explained in the previous sections, data should be normalized and corrected in order to eliminate the variability that is not related with the phenotype or condition that we are interested in, such as technical bias that may be present in the data to avoid that they have a negative impact on the final results. Secondly, variable subset selection, also called Feature Selection (FS), is a key step to reduce the dimensionality, reduce noise, improve model performance, and increase model interpretability. There are several methods to perform FS that are commonly classified in filters, wrappers, embedded methods or ensemble techniques (164). Filter methods rank features based on statistical criteria (variance between samples, correlation measures, etc.) without involving a predictive model, while wrappers iteratively evaluate subsets of features using model performance metrics. Embedded methods incorporate feature selection into the model training process, and ensemble techniques combine multiple models to improve feature ranking and robustness (165). The main reasons of the importance of apply a feature selection process include *i)* the simplification of the data to make it more suitable for traditional statistical methods, *ii)* the overfitting prevention, as too many variables can lead a model to learn from noise in that specific data rather than the underlying patterns *iii)* the result of models more interpretable *iv)* the reduction of computational demands, speeding up model training and making it feasible to apply more complex algorithms or techniques that may be computationally intensive, *v)* minimization of noise in the data by selecting a robust subset of transcripts**.**

The overfitting is one of the main problems when developing ML models. It has been described that when the number of variables is higher than the number of samples it is relatively straightforward to find genes that perform accurately on the data set used to train the model but fail absolutely when tested ins external validations. That is usual when the model captures noise or random fluctuations in the training data generating overfitting and subsequently bad predictions on future observations not included in the training set (161). For this reason, and to be able to evaluate the real performance of the models, the initial data should be divided at least in two different groups, one of them called “training set” that is the one used to develop

and refine the model allowing it to learn patterns in the data, and the "test set" that remains untouched during the training process and is only used to assess the model's ability to generalize to new data. A truly generalizable model that avoids overfitting the training data should also demonstrate strong performance on the test dataset. Conversely, if the model is overfitted to the training data, it is unlikely to perform well on new data in the test set. Thus, evaluating the model on the test subset allows us to assess its performance on unseen data, providing a more realistic measurement of its effectiveness (**Figure 9**).

Along with the feature selection process, there are additional strategies that allow to avoid overfitting. One of them is the regularization process. Regularization techniques prevent overfitting in machine learning models by adding a penalty based on the complexity of the model´s variables. There are two types of regularization: L1 (used by LASSO model) and L2 (used by Ridge model). There are also models that combine L1 and L2 penalties (used by Elastic Net model). These penalties reduce the values of the coefficients of the less informative variables (transcripts) towards zero (L2) or equal zero (L1). This shrinkage needs the selection of a regulation parameter called lambda (λ), that fixes the quantity of penalty to the model coefficients (166). A higher λ value increases the penalty, resulting in more coefficients being shrunk to zero and a simpler model.

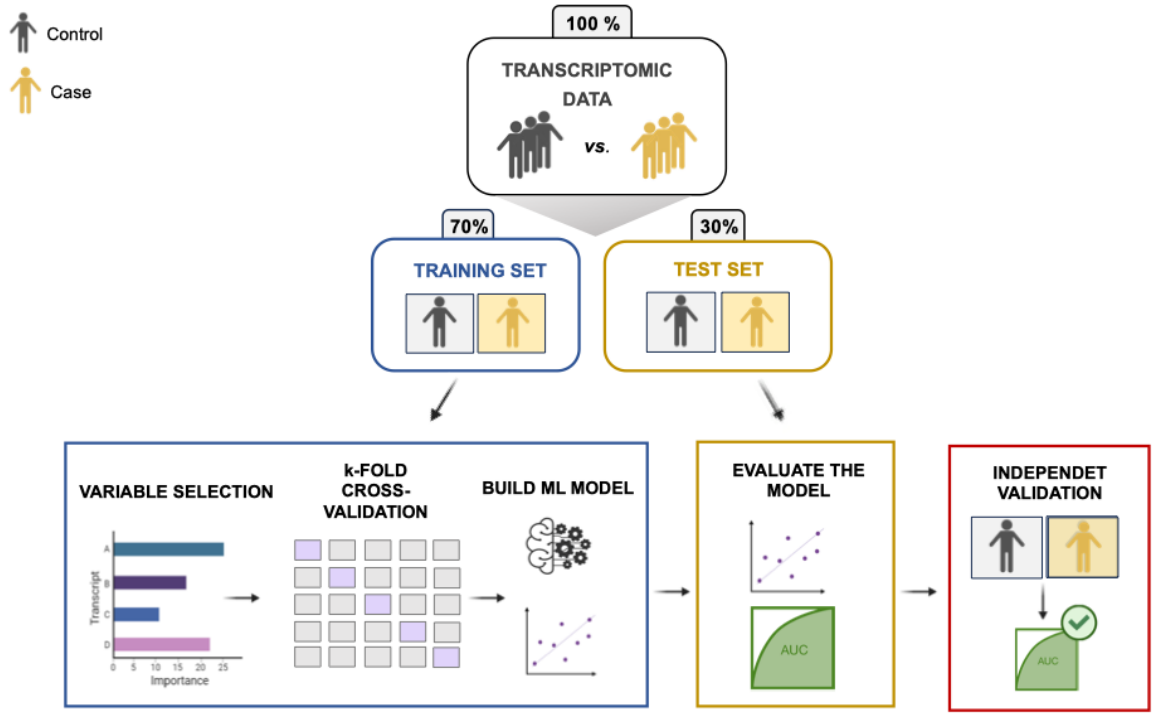


Figure 9. Example of a pipeline for biomarker or signature discovery. Self-made figure created with BioRender

In addition, especially when the training set is not as extensive as desired, it is also helpful to perform a cross-validation approach within the training set. Cross-validation involves dividing the training data into multiple subsets, or "folds," and then training the model on some folds while validating it on the remaining fold This process is repeated across all folds, ensuring that each data point in the training set is used both for training and validation at least once (167). Cross-validation helps prevent the model from becoming too specific to a single partition of data, reducing the risk of overfitting by allowing the model to learn from varied subsets of the data (**Figure 9**).

Furthermore, external validation with data from an entirely separate source can enhance confidence in the model's robustness. This involves testing the model on a dataset collected independently of the training and test sets, ideally under different conditions or including different populations.

To generate all transcriptomic signatures presented in this thesis through the analysis of pneumonia samples, a combined approach was adopted by performing a first subset of variables using DEGs extracted from previous test between viral and bacterial pneumonia or mycoplasma and viral pneumonias respectively. Then some filters were applied to these DEGs to ensure that the selected transcripts that enter the model are well expressed (and thus will be easily detected when necessary) and also that are between the top DEGs (adjusted *p*-value <0.01 and LFC > 2). In the fourth article, a bootstrapping approach was employed to randomly divide the complete dataset into 1,000 independent subsets accounting for 70% of the samples (Training sets; all of them with different combinations of *M. pneumonia* pneumonia and viral pneumonia) and 1.000 combinations with the 30% of the samples not used in the corresponding training subsets (Test sets).

Given the robustness and ease of interpretability of the results, we chose to use the Least Absolute Shrinkage and Selection Operator (LASSO) with a 10-fold cross validation. LASSO is a form of penalized regression (L1) that reduces the regression coefficients to zero by penalizing the regression model with the sum of the absolute coefficients, effectively removing those variables from the model and consequently performing an implicit variable selection. The advantage of the LASSO compared to other penalized regressions such as ridge is that the penalty causes that genes with a minimum contribution to the model have a coefficient equal to zero. Which means that LASSO allows on the one hand to shrink the model coefficients and on the other hand to perform variable selection. As a consequence, LASSO generates simpler and easier to interpret models, that incorporate only a fraction of the transcripts. λ value was selected using the cv.glmet function from *glmet R* package (168) that perform a cross-validation approach to select the optimal value that bring simplicity to the model maintaining accuracy.

USC DE COMPOSTELA

After the application of LASSO regularization, the coefficients of the selected transcripts were refined by performing a general linear model (GLM) analysis. This step enables a more accurate estimation of the relationship between the selected features (transcripts) and the target variable.

Finally, Receiver Operating Characteristic (ROC) curves and Area Under the Curve (AUC) values were calculated to evaluate the performance and predictive accuracy of the transcriptomic signatures. ROC curves represent the balance between sensitivity (true positive rate) and 1-specificity (false positive rate). AUC values quantify the overall performance of the model measuring the area under that ROC curve. An AUC value of 1 indicates perfect discrimination between conditions, while an AUC of 0.5 reflects the worse discrimination as it is equivalent to random chance. *ROCR* and *pROC* R packages were employed to calculate and represent ROC curves and AUC values (169, 170). *OptimalCutPoints* R package was used to select the cut-off values to differentiate categories of interest (171).

4.2.5. Graphical Representation

All graphics were generated using R (www.r-project.org) software using the *ggplot2 ComplexHeatmap and EnhancedVolcano* packages (162, 172, 173).

4.2.6. Co-expression analysis

In order to find co-expression patterns related with the desired phenotype of study, Weighted Gene Correlation network analysis (WGCNA) from gene expression data were performed using *WGCNA* R package (174). The objective is to identify clusters of genes, known as modules, based on their co-expression patterns across samples. These modules are groups of genes that exhibit highly correlated expression, suggesting they may work together in related biological processes. They are then correlated to specific traits or conditions to determine their relevance to the biological context of the study, and finally associated with known biological pathways using the followed mentioned approaches.

The first step in constructing a weighted co-expression network involves calculating Pearson correlations for all gene pairs in the dataset. In this network, nodes represent gene expressions, while edges reflect the strength of pairwise Pearson correlations between them. Modules are identified through hierarchical clustering based on gene expression data. Average-linkage hierarchical clustering combined with a gene dissimilarity measure is employed to build a dendrogram (cluster tree) representing the network structure. After constructing the dendrogram, a height cutoff is selected to define gene clusters. Modules correspond to distinct branches of the dendrogram, representing groups of genes with similar expression patterns.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Normalized expression data from genes showing the most variant expression values between samples (top 50% with higher variance) were used as input for the analysis. The similarity matrix between each pair of genes across all samples was calculated using correlation values and transformed into an adjacency matrix. Subsequently, the topological overlap matrix (TOM) and the corresponding dissimilarity (1–TOM) values were computed. The resulting topological overlap is a biologically meaningful measure of gene similarity based on the co-expression relationship between two genes. The parameters for co-expression module detection were set as a minimum module size of 30, medium sensitivity for cluster splitting, and a dendrogram cut height threshold of 0.2 for module merging. The correlation between module eigengenes and clinical traits was analyzed to identify modules significantly associated with i) the mild-severe phenotype (article 1) and ii) viral pneumonia phenotype (article 3). Module membership (MM), as a measure of intramodular connectivity, was also calculated. The correlation between gene significance (GS) and MM was explored, and the average absolute gene significance for all genes within a module was calculated to highlight the most important modules (Hub modules).

In addition, in the fifth article, genome-wide co-expression correlations were analyzed to investigate gene functions and explore gene-gene interactions using the *CorrelationAnalyzeR* R package (175). This tool leverages immune-related datasets from over 12,000 samples under normal conditions as a reference to identify both the most divergent and the most similar co-expression correlations between genes of interest. This analysis extends beyond individual genes, offering insights at the pathway level to enhance our understanding of gene regulatory networks and biological processes.

4.2.7. Pathways analysis

To infer the biological mechanisms underlying the conditions of study, Gene Expression Enrichment Analyses were performed using gene expression data and reference databases of gene-sets associated with molecular pathways, such as Gene ontology (GO) (176), Enciclopedia de Kyoto de Genes y Genomas (KEGG) (177) and Reactome (178).

The goal of gene expression enrichment analysis is to determine whether a specific set of genes of interest is overrepresented, or "enriched," in relation to the entire gene population with respect to certain characteristics, such as biological functions, molecular pathways, or mechanisms. This involves applying statistical tests to assess whether these genes are more frequently associated with a given feature than would be expected by random chance, considering the overall gene population. A gene or gene set is considered "enriched" if the observed frequency of a particular characteristic (like involvement in a pathway or biological

USC DE SANTIAGO DE COMPOSTELA

function) in the sample is higher than the frequency expected by chance. Typically, the input for this type of analysis is a list of genes that are differentially expressed, associated with a specific phenotype, or highly expressed in a given condition. Pathways analysis can be also applied to the functional assessment of clusters of co-expressed genes that are correlated with a given phenotype.

In the analysis of this thesis, three enrichment approaches were employed:

- Over-representation analysis (ORA): a method used to determine whether specific gene sets (e.g., pathways, biological functions) are over-represented among a list of genes of interest, in this case, differentially expressed genes or clusters of co-expressed genes associated with the phenotype. ORA uses statistical tests (Fisher's exact test) to compare the proportion of genes associated with a given feature (like a biological pathway or function) in a set of interest (DE genes) against the proportion of genes associated with the same feature in the entire transcriptome. If the proportion of genes associated with a particular feature is significantly higher in the set of interest than would be expected by chance, the feature is considered enriched. ORA is useful for identifying over-represented pathways or biological functions based on predefined gene lists but requires setting arbitrary cutoffs (e.g., adjusted p-value or fold change) to define the interest gene list (179).
- Gene Set Expression Analysis (GSEA): a computational method that determines whether a predefined set of genes shows statistically significant, concordant differences between two biological states (e.g., diseased vs. healthy samples) across the entire ranked list of genes. GSEA takes all genes in a dataset and ranks them based on their expression levels or some other relevant metric (e.g., fold change or p-value). It then evaluates whether the genes in a particular gene set (e.g., a pathway) are randomly distributed throughout this ranked list or are found more often at the top or bottom of the list (180). If a gene set is consistently found at the extremes of the ranked list (either highly upregulated or downregulated), it is considered significantly enriched. GSEA does not require an arbitrary cut-off and can capture smaller but coordinated changes in gene expression within a pathway or function so it can detect coordinated changes across an entire pathway, even when individual gene changes are modest(179).
- Gene Set Variation Analysis (GSVA): GSVA is an unsupervised method that evaluates variations in the activity of biological pathways or gene sets across samples (181). It calculates a continuous activity score for each pathway in every sample, reflecting the collective expression of genes within the set. Unlike previous methods, GSVA focuses on pathway-level analysis instead of individual gene expression and does not rely on predefined thresholds or experimental groupings (e.g. “case” *vs.* “control”).

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

ClusteProfiler (182) and *GSVA (181)* R packages were employed to perform enrichment analysis and visualization of results.

4.3. EPIGENOMIC ANALYSIS

Epigenomic analyses involve the identification, quantification and study of heritable modifications that regulate gene expression without altering the underlying genetic sequence. Epigenetic mechanisms are mediated by chemical modifications of the DNA itself or by modifications of proteins. Some of the most well-characterized epigenetic modifications include DNA methylation, chromatin remodeling, histone modification, and mechanisms associated with non-coding RNA. All of them play a crucial role in development, disease, and cellular responses to environmental factors (183).

In the present analysis, methylation data from Illumina's Infinium Methylation EPIC Bead Chip were used to analyze DNA methylation levels across the genome of COVID-19 patients with different severities. DNA methylation is a process where a methyl group (CH3) is added to DNA, mainly in regions where there is a cytosine (C) nucleotide directly followed by a guanine (G) nucleotide and connected by a phosphate bond called CpG sites. This process often affects gene activity by either repressing or activating gene expression, thereby playing a critical role in regulating cellular functions and development. The Methylation EPIC BeadChip is an array platform designed to measure the methylation of cytosines at more than 850,000 specific CpG sites in regulatory and promoter regions of genes, providing detailed information on how alterations in methylation can affect gene expression. In the present thesis, raw data were downloaded from GEO under accession number GSE179325 and subsequently filtered and normalized using *minfi* (184) R package. To identify differentially methylated positions (DMPs) between the phenotype groups, the *limma* package was utilized to fit a linear model adjusted for age. The methylation levels of CpG sites within the *TLR7* gene were assessed across the different severity groups.

4.4. GENOMIC ANALYSIS

Whole Exome Sequencing (WES) is a common technique used to sequence the entire exonic portion of the genome, which represents the protein-coding regions of the DNA. This method involves capturing and sequencing the exons, which account for around 1-2% of the human genome but contain most of known disease-related mutations (185). WES enables the discovery of genetic variations such as single nucleotide polymorphisms (SNPs), insertions, deletions, and mutations that may contribute to the disease phenotype or the susceptibility of individuals to the disease.

USC DE COMPOSTELA

To investigate genetic susceptibility to community-acquired pneumonia (CAP), a case-control study was conducted using whole exome sequencing of adult CAP patients and healthy controls (**Table 4)**. DNA was extracted from saliva samples, and all samples were sequenced using Illumina instruments (**Table 4)**.

Table 4. New generated genomic data used for the analysis.

Study	Source	Type of Sample	Platform	Sample size	
ARTICLE 3	CAPRIC study	Saliva	HiSeq 400 and NovaSeq 6000	n= 738	n= 300 (CAP patients) n= 438 (Healthy Controls)

The raw sequencing data underwent quality control and preprocessing, including alignment to the hg19 reference genome, removal of duplicated reads, and identification of potential errors. Variants were detected and genotyped using the Haplotype Caller from Genome Analysis Toolkit (GATK). Quality filtering was also performed, removing indels, non-biallelic positions, monomorphic SNPs, and variants with a genotyping rate lower than 99.9%. Kinship analysis was performed analyzing Identity By Descent pairwise values (IBD) to remove related individuals, and ancestry analysis was performed using Admixture software (186) and Identity By State pairwise values (IBS) to remove the variability caused by different continental origin.

SNP association tests were carried out using different genetic models to identify potential genetic factors underlying CAP. First, single-point SNP association tests were applied to alleles, followed by genotype-based tests considering additive, dominant, and recessive models. Fisher's exact test was used for both allelic and genotype-based association analysis between the CAP cohort and healthy controls. Allelic association tests focused on variants with a Minor Allele Frequency (MAF) > 0.05 and with a Hardy-Weinberg equilibrium p-value > 0.001 in both case and control groups. For genotype-based analysis, only SNPs with greater than 5% frequency in all genotype categories were included.

In addition, a gene-based SKAT-O (Optimized Sequence Kernel Association Test) was applied to assess the combined effect of rare variants (MAF < 0.05) within selected genes. This test evaluates the cumulative impact of variants in a gene (or region) and helps to expand our understanding of the genetic contribution underlying CAP susceptibility. Only genes with at least ten variants were considered for these analyses.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Plink v1.9 software (187) was employed to curate the data and perform the single-point association test. R statistical software was used for the gene-based association test. Specifically, the SKAT R library (188) was utilized to compute SKAT-O association tests. A nominal significance level was set to 0.05. To monitor multiple tests, we used the False Discovery Rate (FDR) approach

USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

5. RESULTS

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

5. RESULTS

5.1. ARTICLE 1: A MULTI-TISSUE STUDY OF IMMUNE GENE EXPRESSION PROFILING HIGHLIGHTS THE KEY ROLE OF THE NASAL EPITHELIUM IN COVID-19 SEVERITY

The results from this chapter have already been published as Alberto Gómez-Carballa [a,b,c], Irene Rivero-Calle [a,c,d], Jacobo Pardo-Seco [a,b,c], José Gómez-Rial [a, c, e], Carmen Rivero-Velasco [f], Nuria Rodríguez-Núñez [g], Gema Barbeito-Castiñeiras [h i], Hugo Pérez-Freixo [j], Miriam Cebey-López [a b c], Ruth Barral-Arca [a b c], Carmen Rodriguez-Tenreiro [a c d], Ana Dacosta-Urbieta [a c d], Xabier Bello [a b c], Sara Pischedda [a b c], María José Currás-Tuala [a b c], Sandra Viz-Lasheras [a b c], Federico Martinón-Torres [a c d 1], Antonio Salas [a b c 1] GEN-COVID study group [2]. A multi-tissue study of immune gene expression profiling highlights the key role of the nasal epithelium in COVID-19 severity. Environmental research. 2022; 210: 112890. DOI: https://doi.org/10.1016/j.envres.2022.112890

Affiliations

a- Genetics, Vaccines and Infections Research Group (GENVIP), Instituto de Investigación Sanitaria (IDIS) de Santiago, Santiago de Compostela, Spain.

b- Unidade de Xenética, Instituto de Ciencias Forenses (INCIFOR), Facultade de Medicina, Universidade de Santiago de Compostela (USC), and GenPoB Research Group, Instituto de Investigación Sanitaria (IDIS), Hospital Clínico Universitario de Santiago (SERGAS), Galicia, Spain.

c- Centro de Investigación Biomédica en Red de Enfermedades Respiratorias (CIBERES), Madrid, Spain.

d- Translational Pediatrics and Infectious Diseases, Department of Pediatrics, Hospital Clínico Universitario de Santiago de Compostela, Santiago de Compostela, Spain.

e- Laboratorio de Inmunología. Servicio de Análisis Clínicos. Hospital Clínico Universitario (SERGAS), Galicia, Spain.

f-Intensive Medicine Department, Hospital Clìnico Universitario de Santiago de Compostela, Galicia, Spain.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

g- Pneumology Department, Hospital Clìnico Universitario de Santiago de Compostela, Galicia, Spain.

h- Clinical Microbiology Unit, Complexo Hospitalario Universitario de Santiago Santiago de Compostela, Spain.

i- Instituto de Investigación Sanitaria de Santiago, Santiago de Compostela, Spain.

j-Preventive Medicine Department, Hospital Clínico Universitario de Santiago de Compostela, Spain.

Link: https://www.sciencedirect.com/science/article/pii/S0013935122002171

Abstract

Coronavirus Disease-19 (COVID-19) symptoms range from mild to severe illness; the cause for this differential response to infection remains unknown. Unravelling the immune mechanisms acting at different levels of the colonization process might be key to understand these differences. We carried out a multi-tissue (nasal, buccal and blood; *n* = 156) gene expression analysis of immune-related genes from patients affected by different COVID- 19 severities, and healthy controls through the *nCounter* technology. Mild and asymptomatic cases showed a powerful innate antiviral response in nasal epithelium, characterized by activation of IFN pathway and downstream cascades, successfully controlling the infection at local level. In contrast, weak macrophage/ monocyte driven innate antiviral response and lack of IFN signalling activity were present in severe cases. Consequently, oral mucosa from severe patients showed signals of viral activity, cell arresting and viral dissemination to the lower respiratory tract, which ultimately could explain the exacerbated innate immune response and impaired adaptative immune responses observed at systemic level. Results from saliva transcriptome suggest that the buccal cavity might play a key role in SARS-CoV-2 infection and dissemination in patients with worse prognosis. Co-expression network analysis adds further support to these findings, by detecting modules specifically correlated with severity involved in the abovementioned biological routes; this analysis also provides new candidate genes that might be tested as biomarkers in future studies. We also found tissue specific severity-related signatures mainly represented by genes involved in the innate immune system and cytokine/chemokine signalling. Local immune response could be key to determine the course of the systemic response and thus COVID-19 severity. Our findings provide a framework to investigate severity host gene biomarkers and pathways that might be relevant to diagnosis, prognosis, and therapy.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

DOI: 10.1016/j.envres.2022.112890

JOURNAL: Environmental Research (ISSN: 1096-0953)

CATHEGORY: PUBLIC, ENVIRONMENTAL & OCCUPATIONAL HEALTH; ENVIRONMENTAL SCIENCES

RELATIVE POSSITION IN THE CATHEGORY: **16/207** in PUBLIC, ENVIRONMENTAL & OCCUPATIONAL HEALTH (**Q1 and D1**) and **32/275 in** ENVIRONMENTAL SCIENCES (**Q1**).

IMPACT FACTOR: **8.3** (2022)

OTHER QUALITY EVIDENCE: Environmental Research (ISSN: 1096-0953) is an international peer-reviewed open access journal focused on publishing high-quality research in environmental science, biomedicine, and public health. Rank by Journal Citation Indicator situate this journal in the **15/334 (D1)** position for ENVIROMENTAL SCIENCES and **21/400 (D1)** in PUBLIC, ENVIRONMENTAL & OCCUPATIONAL HEALTH. The journal had 16.997 citations in 2022.

AUTHOR CONTRIBUTION: In the present study, the doctoral student took part in the data analysis and the writing process.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Environmental Research 210 (2022) 112890

Contents lists available at ScienceDirect

Environmental Research

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

A multi-tissue study of immune gene expression profiling highlights the key role of the nasal epithelium in COVID-19 severity

Alberto Gómez-Carballa [a,b,c], Irene Rivero-Calle [a,c,d], Jacobo Pardo-Seco [a,b,c], José Gómez-Rial [a,c,e], Carmen Rivero-Velasco [f], Nuria Rodríguez-Núñez [g], Gema Barbeito-Castiñeiras [h,i], Hugo Pérez-Freixo [j], Miriam Cebey-López [a,b,c], Ruth Barral-Arca [a,b,c], Carmen Rodriguez-Tenreiro [a,c,d], Ana Dacosta-Urbieta [a,c,d], Xabier Bello [a,b,c], Sara Pischedda [a,b,c], María José Currás-Tuala [a,b,c], Sandra Viz-Lasheras [a,b,c], Federico Martinón-Torres [a,c,d,1], Antonio Salas [a,b,c,1,*], GEN-COVID study group [2]

[a] Genetics, Vaccines and Infections Research Group (GENVIP), Instituto de Investigación Sanitaria (IDIS) de Santiago, Santiago de Compostela, Spain
[b] Unidade de Xenética, Instituto de Ciencias Forenses (INCIFOR), Facultade de Medicina, Universidade de Santiago de Compostela (USC), and GenPoB Research Group, Instituto de Investigación Sanitaria (IDIS), Hospital Clínico Universitario de Santiago (SERGAS), Galicia, Spain
[c] Centro de Investigación Biomédica en Red de Enfermedades Respiratorias (CIBERES), Madrid, Spain
[d] Translational Pediatrics and Infectious Diseases, Department of Pediatrics, Hospital Clínico Universitario de Santiago de Compostela, Santiago de Compostela, Spain
[e] Laboratorio de Inmunología. Servicio de Análisis Clínicos. Hospital Clínico Universitario (SERGAS), Galicia, Spain
[f] Intensive Medicine Department, Hospital Clìnico Universitario de Santiago de Compostela, Galicia, Spain
[g] Pneumology Department, Hospital Clìnico Universitario de Santiago de Compostela, Galicia, Spain
[h] Clinical Microbiology Unit, Complexo Hospitalario Universitario de Santiago Santiago de Compostela, Spain
[i] Instituto de Investigación Sanitaria de Santiago, Santiago de Compostela, Spain
[j] Preventive Medicine Department, Hospital Clínico Universitario de Santiago de Compostela, Spain

ARTICLE INFO

Keywords:
COVID-19 severity
Gene expression
Immune response
Pathways analysis
Multi-tissue
Differential expression analysis
Co-expression analysis
SARS-CoV-2

ABSTRACT

Coronavirus Disease-19 (COVID-19) symptoms range from mild to severe illness; the cause for this differential response to infection remains unknown. Unravelling the immune mechanisms acting at different levels of the colonization process might be key to understand these differences. We carried out a multi-tissue (nasal, buccal and blood; $n = 156$) gene expression analysis of immune-related genes from patients affected by different COVID-19 severities, and healthy controls through the *nCounter* technology. Mild and asymptomatic cases showed a powerful innate antiviral response in nasal epithelium, characterized by activation of interferon (IFN) pathway and downstream cascades, successfully controlling the infection at local level. In contrast, weak macrophage/monocyte driven innate antiviral response and lack of IFN signalling activity were present in severe cases. Consequently, oral mucosa from severe patients showed signals of viral activity, cell arresting and viral dissemination to the lower respiratory tract, which ultimately could explain the exacerbated innate immune response and impaired adaptative immune responses observed at systemic level. Results from saliva transcriptome suggest that the buccal cavity might play a key role in SARS-CoV-2 infection and dissemination in patients with worse prognosis. Co-expression network analysis adds further support to these findings, by detecting modules specifically correlated with severity involved in the abovementioned biological routes; this

Abbreviations: BALF, Boncho-Alveolar Lavage Fluids; COVID-19, Coronavirus Disease 19; CRP, C-Reactive Protein; DEG, Differentially Expressed Gene; DE, Differential Expression; GEO, Gene Expression Omnibus; GO, Gene Ontology; GSEA, Gene-Set Enrichment Analysis; GS, Gene Significance; HK, House Keeping; H-B, Benjamini-Hochberg; ISGs, Interferon Stimulated Genes; IFN, Interferon; LRT, Lower Respiratory Track; MM, Module membership; NETs, Neutrophil Extracellular Traps; NP, Naso-Pharyngeal; ORA, Over-Representation; PAMPs, Pathogen-Associated Molecular Patterns; PCA, Principal Component Analysis; PRR, Pattern Recognition Receptors; QC, Quality Control; SARS-CoV-2, Evere acute Respiratory Syndrome Coronavirus 2; SD, Standard Deviation; TOM, Topological Overlap Matrix.

* Corresponding author. Unidade de Xenética, Instituto de Ciencias Forenses (INCIFOR), Facultade de Medicina, Universidade de Santiago de Compostela (USC), and GenPoB Research Group, Instituto de Investigación Sanitaria (IDIS), Hospital Clínico Universitario de Santiago (SERGAS), Galicia, Spain.
E-mail address: antonio.salas@usc.es (A. Salas).
[1] equal contribution.
[2] GEN-COVID Study Group (www.gencovid.eu), Spain

https://doi.org/10.1016/j.envres.2022.112890
Received 9 November 2021; Received in revised form 11 January 2022; Accepted 2 February 2022
Available online 22 February 2022
0013-9351/© 2022 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

analysis also provides new candidate genes that might be tested as biomarkers in future studies. We also found tissue specific severity-related signatures mainly represented by genes involved in the innate immune system and cytokine/chemokine signalling. Local immune response could be key to determine the course of the systemic response and thus COVID-19 severity. Our findings provide a framework to investigate severity host gene biomarkers and pathways that might be relevant to diagnosis, prognosis, and therapy.

1. Introduction

The severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), which causes the coronavirus disease 19 (COVID-19), was first identified in patients suffering from a respiratory disease in Wuhan (Hubei province, China) in December 2019 (Gómez-Carballa et al., 2020b; Pipes et al., 2020). Soon after, the virus quickly spread worldwide favored by two main catalyzers, namely, variants showing higher transmissibility (Davies et al., 2021) and superspreading events (Althouse et al., 2020; Gómez-Carballa et al., 2020a; Lemieux et al., 2020; Gómez-Carballa et al., 2021).

COVID-19 is mainly characterized by producing influenza-like illness (Jin et al., 2020; Lamers et al., 2020). A wide range of symptoms have been described, from absolutely no symptoms to severe pneumonia, lung damage, and progressive respiratory failure (Huang et al., 2020). Investigating the immunological processes underpinning SARS-CoV-2 pathogenesis is key for the development of efficient therapeutic strategies and vaccines against COVID-19. The immune system dysregulation described in severe patients is characterized by a reduce number of lymphocyte T subpopulations $CD4^+$ and $CD8^+$, pointing to an absence of T-cell specific response (Qin et al., 2020; Yang et al., 2020; Zheng et al., 2020b). Impairment of the adaptative immune system, together with a non-effective host immune response by innate cells, could be the cause of a higher viral replication and the activation of inflammatory processes, resulting in tissue damage and a more severe disease course (Gomez-Rial et al., 2020; Qin et al., 2020). The hyper-inflammatory response occurring during the acute phase is mainly driven by the monocyte-macrophage activation (Gómez-Rial et al., 2020), the release of pro-inflammatory cytokines (McGonagle et al., 2020; Qin et al., 2020; Zhang et al., 2020) and increase of other mediators e.g. ferritin and C-reactive protein (CRP). The infiltration of monocytes-derived macrophages from blood points to these cells as the main cause of the lung tissue damage in severe cases (Gomez-Rial et al., 2020).

Gene expression studies conducted on blood samples from COVID-19 patients indicated that genes associated with inflammatory and hypercoagulability pathways (Zhou et al., 2021) and the imbalance between innate and adaptive immune responses are the main factors responsible for severe disease course (Zheng et al., 2020a). Thair et al. (2021) found a differential transcriptomic pattern in blood that differentiated COVID-19 patients from other viral infections, suggesting that their top differentially expressed genes (DEGs) may be linked to new mechanisms of pathogen evasion from host immune response. Most recently, Aschenbrenner et al. (2021) revealed neutrophil activation signatures in severe cases, coupled with an elevated expression of genes related to coagulation and platelet function, and absence of T-cell activation.

Most of the studies to date were carried out *in vitro* or nasal-derived epithelial cells (Gamage et al., 2020; Sajuthi et al., 2020; Sungnak et al., 2020; Alfi et al., 2021). A gene expression study carried out on naso-pharyngeal (NP) swabs of infected patients pointed to different expression patterns depending on the viral load, age, and sex (Lieberman et al., 2020). Comparison of expression profiles of infected adults and children (Pierce et al., 2021) indicated an early and more efficient innate immune response in nasal mucosa from children, thus conditioning the final outcome of the disease. So far, only one study undertaken in three severe patients, has proposed candidate transcripts for severity and disease outcome in NP samples (Jain et al., 2021).

Saliva has also been proposed as a good alternative to NP swabs for SARS-CoV-2 detection (Pasomsub et al., 2020; To et al., 2020; Wyllie et al., 2020). The study of the specific host gene expression profile produced by the infection in oral mucosa has not raised comparable interest; only a recent study by Huang et al. (2021) highlighted the importance of the oral cavity in viral transmission and dissemination.

We carried out a multi-source gene expression study in blood, nasal, and saliva samples, collected on asymptomatic, mild, and severe patients. Our study aims at understanding the differential immune response that leads to different disease severities, and finding specific severity-related biomarkers to improve patient management and prognostic prediction.

2. Materials and methods

2.1. Patient recruitment

We prospectively collected blood, saliva, and nasal epithelium samples from 52 COVID-19 patients with confirmed PCR-positive diagnosis of SARS-CoV-2 infection at the Hospital Clínico Universitario of Santiago de Compostela (Galicia; Spain) from March to June 2020. We also recruited samples from healthy controls, defined as asymptomatic subjects with no diagnosis of COVID-19 disease at the time of sample collection. To avoid potential effects of ancestry on gene expression (Barral-Arca et al., 2019), we assured that patients and healthy controls were of (self-reported) South-European ancestry. In addition, we also carried out a maximum likelihood estimation of individual ancestries from multi-locus SNP data using whole exome sequencing data available from a significant proportion of the patients (35/52; data not shown). Procedures for exome data processing, quality control, variant annotation and estimation of ancestries were performed as done previously (Salas et al. 2017, 2018). All the analyzed samples were of European genomic ancestry (Supplementary Table 1) corroborating self-reported expectations.

Cases were classified at the time of sample collection by severity illness: severe (ICU admission), moderate (non-ICU but admitted to hospital) and mild (domiciliary lockdown patients with mild symptoms or asymptomatic; herein mild patients). Comparisons were carried out as follows: (*i*) whole blood from 41 patients and 13 controls, (*ii*) nasal epithelium from 38 patients and 11 controls, and (*iii*) saliva from 41 patients and 12 controls. The complete tissue sample set could be collected for 27 out of the 52 patients. Demographic and clinical data from the patients are summarized in Table 1 (see Supplementary Table 1 for more details).

The study was approved by the Ethical Committee of Clinical Investigation of Galicia (CEIC ref. 2020/178) and was conducted in accordance with the principles of the Declaration of Helsinki and all applicable Spanish legislation, namely Biomedical Research Act (Law 14/2007–3 of July), Act 41/2002 on the Autonomy of the Patient, Decree SAS/3470/2009 for Observational Studies, and 15/1999 Data Protection Act. Written informed consent was obtained for subjects before study inclusion.

2.2. Sample collection and processing

Blood, nasal epithelium, and saliva samples were sampled from patients and controls at the same timepoints. Blood samples (2.5 ml) were collected in PAXgene tubes (BD). RNA was isolated using PAXgene blood miRNA extraction kit (Qiagen) following manufacturer recommendations and carrying out the on-column DNase I treatment during the

Table 1
Demographic and clinical characteristics of the COVID-19 cohort and controls. Phenotype stratification of the patients refers to the severity at the time of sample collection. Some of the clinical parameters are only available for hospitalized cases (moderate and severe). Quantitative variables are represented by the median (IQR) and qualitative variables as percentages with absolute numbers in brackets. Asterisks indicate statistically significant differences between groups.

	Controls	Mild	Moderate	Severe
Gender				
Female	68.0% (17/25)	56.5% (13/23)	50.0% (7/14)	26.7% (4/15)
Male	32.0% (8/25)	43.5% (10/23)	50.0% (7/14)	73.3% (11/15)
Age				
	41 (34–52)	38 (29.5–46.5)	58.5 (37.75–67.75)	68 (58–71)
Final Severity				
Control	100.0% (25/25)	0.0% (0/23)	0.0% (0/14)	0.0% (0/15)
Death	0.0% (0/25)	0.0% (0/23)	0.0% (0/14)	6.7% (1/15)
Mild	0.0% (0/25)	100.0% (23/23)	0.0% (0/14)	0.0% (0/15)
Moderate	0.0% (0/25)	0.0% (0/23)	92.9% (13/14)	0.0% (0/15)
Severe	0.0% (0/25)	0.0% (0/23)	7.1% (1/14)	93.3% (14/15)
Comorbidity				
	4.0% (1/25)	26.1% (6/23)	78.6% (11/14)	86.7% (13/15)
Days of symptoms to sampling				
	8 (6.75–10.25)	13 (12–15.75)	12 (9.5–14.5)	
Admission days				
	–	–	9.5 (6.5–11)	34 (24.5–53)
Admission days to sampling				
	–	–	4 (3–5)	4 (2–6)
ICU				
ICU	–	–	7.1% (1/14)	100% (15/15)
ICU days	–	–	11 (11–11)	22.5 (10.75–28.75)
ICU days to sampling	–	–	–	3 (1.5–5)
Symptom				
Fever	–	–	92.9% (13/14)	86.7% (13/15)
Respiratory	–	–	100% (14/14)	93.3% (14/15)
Gastrointestinal	–	–	42.9% (6/14)	28.6% (4/14)
Musculoskeletal	–	–	64.3% (9/14)	60.0% (9/15)
Sensory	–	–	21.4% (3/14)	26.7% (4/15)
Headache	–	–	35.7% (5/14)	0.0% (0/15)
Xray Infiltrates				
	–	–	71.4% (10/14)	86.7% (13/15)
Xray pneumonia with consolidation				
	–	–	64.3% (9/14)	50% (7/14)
Xray Pleural Effusion				
	–	–	14.3% (2/14)	7.1% (1/14)
Treatments				
Oxygen	–	–	64.3% (9/14)	100% (14/14)
Invasive Vent	–	–	7.1% (1/14)	78.6% (11/14)
Inotropes	–	–	0.0% (0/14)	35.7% (5/14)
Haemofiltration	–	–	0.0% (0/14)	14.3% (2/14)
Corticoids before sample collection	–	–	14.3% (2/14)	66.7% (10/15)
Antibiotics	–	–	100% (14/14)	100% (15/15)
Antiviral	–	–	78.6% (11/14)	100% (15/15)
Antimalarial	–	–	100% (14/14)	100% (15/15)
Monoclonal antibody	–	–	0.0% (0/14)	46.6% (7/15)
Analytical parameters				
White Cells (10^9/L)	–	–	5.505 (4.82–6.3725)	9.41 (5.275–12.36)
Platelets (10^9/L)	–	–	273 (227.25–400.75)	203 (129–255.5)
*Neutrophils (10^9/L)**	–	–	3.095 (2.445–3.78)	6.89 (3.185–10.55)
*Lymphocytes (10^9/L) **	–	–	1.32 (1.1875–1.52)	0.47 (0.415–0.905)
Monocytes (10^9/L)	–	–	0.42 (0.29–0.58)	0.33 (0.195–0.475)
*Fibrinogen (mg/dl) **	–	–	500 (500–500)	373.5 (318–450)
Creatinine (mg/dl)	–	–	0.75 (0.58–1.03)	0.8 (0.765–0.94)
CRP (mg/L)	–	–	23.39 (15.5025–53.63)	48.255 (11.15–99.6225)
PCT (mg/ml)	–	–	0.06 (0.04–0.1375)	0.15 (0.105–0.2725)
*Ferritin (ng/dl) **	–	–	398.5 (139.75–491)	1584.5 (704–2934)
*D–Dimer (ng/ml) **	–	–	518 (402–763)	2038 (861–7832)
*Lactate (mmol/L) **	–	–	0.9 (0.85–0.95)	1.72 (1.42–2.39)
*Maximum CRP (mg/L) **	–	–	49.225 (17.795–123.78)	134.205 (71.9475–172.29)
*Maximum Neutrophils (10^9/L) **	–	–	4.56 (3.4775–6.8)	9.715 (5.85–11)
*Minimum Lymphocytes (10^9/L) **	–	–	0.87 (0.7–1.1675)	0.43 (0.345–0.4775)

extraction process. Saliva and nasal epithelium samples were collected in Oragene CP-190 kit (DNA Genotek). We used sponges (SC-2 kit from DNA Genotek) to assist saliva collection in ICU patients unable to spit. RNA from saliva and nasal epithelium specimens was isolated using 500 μl of sample and the RNeasy microkit (Qiagen). We used Oragene RNA neutralizer solution (DNA Genotek) to precipitate impurities and inhibitors and slightly modified the protocol provided by the extraction kit as recommended by the Oragene tubes supplier. After isolation, RNA amount and integrity was checked using TapeStation 4200 (Agilent), and DV200 values were calculated to both ensure that >50% of the of the RNA fragments were above 200 nt and estimate the optimal sample input for the *nCounter* analysis. Nasopharyngeal samples were collected in Universal Transport Medium tubes (UTM) tubes to check for the presence as well as the viral load of SARS-CoV-2 at the same time as collection of blood, saliva, and nasal epithelium samples. Molecular detection of viral particles was performed through a multiplex real-time

PCR using the Allplex™SARS-CoV-2 Assay (Seegene).

2.3. Nanostring nCounter expression assay

Immunological gene expression patterns of different tissues from COVID-19 patients and controls were evaluated through the *SPRINT nCounter* system (NanoString Technologies) and the Immunology Panel (594 genes involved in immunological processes). This panel covers the main immunological routes, including major classes of cytokines and their receptors, chemokine ligands and receptors, IFNs and their receptors, the TNF-receptor superfamily, and the KIR family genes. We used the standard gene expression protocol for 12x RNA hybridization with 5 μl of RNA as input, and a hybridization time of 18h for all samples. We mixed control and COVID-19 samples in all the same plates to avoid technical sample bias. We first carried out a quality control (QC) of the raw data checking some technical parameters following manufacturer recommendations to verify the absence of technical problems. Samples that did not pass technical QC, or with low number of genes detected, were excluded for downstream analysis.

2.4. Sample normalization and differential expression analysis

For the longitudinal multi-tissue experiment, normalization of the gene expression data was carried out for each tissue separately by selecting the optimal number and best reference candidate from the set of housekeeping (HK) genes included in the panel, as well as other endogenous genes that showed high levels of stability between samples. The *GeNorm* algorithm (Vandesompele et al., 2002) implemented in the R package *CrtlGene* (Zhong, 2019) was used to test for the most stable and optimal number of genes for normalization. We excluded HK candidate genes with less than 50 raw counts in any of the samples compared from the reference selection analysis. Data normalization was performed through a combined approach using both *DESeq2* (Love et al., 2014) and *RUVSeq* (Risso et al., 2014) as described in (Bhattacharya et al., 2020). Genes with counts below the background (mean + 2 standard deviations [SD] of the negative control spikes in the code set) were not considered for the differential expression (DE) analysis. After data normalization and background thresholding, we eliminated lower expressed genes from the list of DEG, retaining a total of 495 genes in blood samples, 544 genes in nasal epithelium samples and 545 genes in saliva samples for downstream analysis. DEG between cohorts were obtained using *DESeq2* (Love et al., 2014). We used age and sex as covariables to correct the model. Volcano plots and heatmaps were built with *EnhancedVolcano* (Blighe et al., 2020) and *ComplexHeatmap* (Gu et al., 2016) R packages, respectively. All graphics were generated using R (www.r-project.org) software and the *ggplot2* package (Wickham, 2016). Wilcoxon test was used to assess statistical significance between patient groups, and Spearman test for the correlation indices. We focused on highly statistically significant expression changes by considering as thresholds an adjusted P-value < 0.05, and $\log_2FC \geq |1.5|$ for downstream analysis. For the transversal multi-tissue experiment we sub-selected 21 patients for which samples from the three tissues were available (mild: $n = 11$; severe: $n = 10$). We employed a complex model design including the tissue/severity interaction to find tissue-specific severity effects. As references for the comparisons in the transversal analysis, we used blood tissue and the mild severity category. We followed the same procedure as in the longitudinal analysis but data from the three tissues were normalized together after selecting a set of adequate reference genes.

2.5. Pathway analysis

We followed both over-representation (ORA) and gene-set enrichment (GSEA) (Subramanian et al., 2005) approaches to identify biological functions, cellular components and pathways related to the differences in gene expression. For ORA we used the DEG as input (threshold: $\log_2FC \geq |1.5|$), and the genes included in the expression panel as pool for statistical calculations. For GSEA we used all available molecular measurements ($\log_2FC$) from the genes included in the DE analysis to detect coordinated changes in the expression of genes from same pathway. Analyses were carried out using the *Clusterprofiler* (Yu et al., 2012) R package. We applied the Benjamini-Hochberg (H–B) procedure for multiple test correction and a significance threshold of 0.05. We tested both the GO (Gene Ontology) and Reactome databases. Graphics were obtained with the R package *enrichplot* (Yu, 2021).

2.6. Validation analysis

To validate the specificity of the transcriptomic findings from our blood samples, we first used two independent datasets: (a) Thair et al., (2021) included RNA-seq transcriptomes for 62 SARS-CoV-2 positive patients and 24 healthy controls; and (b) Aschenbrenner et al., (2021) included RNA-seq transcriptomes for 39 COVID-19 patients and 10 healthy controls. We additionally checked the specificity of our COVID-19 candidate genes with respect to other viral infections by correlating our results with those from microarray data comparing mixed viral infections ($n = 652$) *vs.* healthy control samples ($n = 672$) reported in Thair et al. (2021).. Independently, to further test the gene expression pattern of our best candidate genes in other viral and bacterial infections, we carried out an independent validation study by meta-analyzing other datasets ($n = 513$ viral, $n = 382$ bacterial and $n = 215$ healthy control samples) ;downloaded from the public gene expression repository Gene Expression Omnibus (GEO) under accession numbers: GSE64456 (Mahajan et al., 2016), GSE40012 (Parnell et al., 2012), GSE42026 (Herberg et al., 2013), GSE60244 (Suarez et al., 2015), also included in Thair et al., (2021) as viral cohorts; and GSE72829 (Herberg et al., 2016), GSE69529 (DeBerg et al., 2018). To merge and integrate all these data , we first normalized and pre-processed each dataset separately using *Lumi* (Du et al., 2008) for Illumina® microarrays data and *Oligo* (Carvalho and Irizarry, 2010) for Affymetrix® microarray datasets, while RNA-seq data were pre-processed as described previously (Barral-Arca et al., 2018). Then, we merged these databases retaining only the common genes. Subsequently, we used the R package *COCONUT* to combine all datasets into one and reduce batch effects in the meta-analysis (Sweeney et al., 2016). Next, we used *Limma* package (Ritchie et al., 2015) to detect DEG and calculate $\log_2FC$ values between groups. Scatterplots of $\log_2FC$ values and Pearson correlation indexes were obtained for the validation analysis.

2.7. Leukocyte proportion estimation from gene expression data

We imputed immune cell subset abundances from expression data in different patient categories, including mild and healthy controls, from which real immune cell counts from blood tests were not available. We used the CIBERSORTX (Newman et al., 2019) tool and a leukocyte gene signature matrix (Newman et al., 2015) obtained from microarray data as reference. We applied a B-mode batch correction to account for cross-platform variation and 500 permutations for the analysis. We graphically represented immune cell proportions using R software and carried out a Wilcoxon test to assess statistical significance between categories.

2.8. Weighted gene co-expression network analysis

We generated a signed weighted correlation networks from gene expression data with the *WGCNA* R package (Langfelder and Horvath, 2008) in order to find co-expression modules related to mild-severe phenotypes in the different tissues analyzed. We used normalized expression data (and corrected for differences in gender and age) from genes showing the most variant expression values between samples (top 50% with higher variance) as input, and we chose a soft-thresholding

power based on the criterion of scale-free topology after testing a set of candidate powers (powers selected were 16, 9 and 18 for blood, nasal epithelium and saliva samples, respectively). The similarity matrix between each pair of genes across all samples was calculated using correlation values and transformed into an adjacency matrix. Subsequently, the topological overlap matrix (TOM) and the corresponding dissimilarity (1–TOM) values were computed. The resulting topological overlap is a biologically meaningful measure of gene similarity based on the co-expression relationship between two genes. As co-expression module detection parameters we chose a minimum module size of 30, a medium sensitivity for cluster splitting, and a 0.2 value as dendrogram cut heigh threshold for module merging.

Correlation between module eigengenes and clinical traits were analyzed to identify modules of interest significantly associated with the mild-severe phenotype (gene significance [GS]). Module membership (MM), as a measurement of intramodular connectivity, was also calculated. We explored the correlation between GS and MM, and calculated the average absolute gene significance for all genes within a module in order to highlight the most important modules. The top hub genes within the most relevant modules were selected using both MM > 0.7 and a *P*-value for GS < 0.05 as thresholds.

We studied biological significance of the hub genes from the phenotype-related modules by performing an over-representation analysis and using the compare cluster function included in the *Clusterprofiler* R package (Yu et al., 2012).

3. Data availability

Expression data that support the findings of the present study have been deposited in Gene Expression Omnibus (GEO) database under the accession code GSE183071.

4. Results

4.1. Cohort description

Most of the clinical features were only available for hospitalized patients (Table 1). Similar median symptoms days to sampling were obtained for mild, moderate, and severe cases, as well as admission days to sampling in the case of moderate and severe patients (*P*-value = 0.236 and 0.704, respectively). During admission, oxygen support was required for 64.3% of the moderate cases and 100% of the severe cases (78.6% required invasive ventilation). All moderate cases were treated with a double therapy including antibiotic and antimalarial drugs (in 78.6% of the moderate cases, antiviral therapy was also administrated). Severe cases received a combined triple therapy of antibiotic, antiviral (lopinavir and ritonavir) and antimalarial (hydroxychloroquine) treatment, and 46.6% of them received tocilizumab.

No statistical differences were observed in viral loads at the time of sample collection depending on COVID-19 severity, nor statistical correlation between days of symptoms to sample collection and Ct values from the SARS-CoV-2 qPCR assay (Supplementary Fig. 1A) on nasopharyngeal swabs obtained at the point of sample collection. However, we obtained statistically significant differences in some analytical parameters between severe and moderate cases at the sampling timepoint: higher neutrophil counts (*P*-value = 0.012), ferritin (*P*-value = 0.002), D-Dimer (*P*-value = 0.006) and lactate (*P*-value = 0.019) values, but lower lymphocyte counts (*P*-value = 0.002) and fibrinogen (*P*-value = 0.004) values in severe cases. Maximum values for C-reactive protein (CRP) and neutrophils (*P*-values = 0.02 and 0.006, respectively) and minimum values for lymphocytes (*P*-value < 0.001) also showed statistically significant differences between these groups.

Immune cell fractions inferred from gene expression data also pointed to a significantly lower proportion of T lymphocyte subpopulations CD4 and CD8 in severe cases than in other severity categories and healthy controls (Supplementary Fig. 1B). Likewise, the proportion of B cells was significantly higher in mild when compared to moderate and severe cases. Cell numbers of CD4 and NK subpopulation also showed a significant decrease with disease severity. Finally, a higher proportion of monocytes was also observed in moderate cases, correlating well with real monocyte counts detected in the blood tests from moderate and severe cases.

4.2. Transcriptomic immune response in blood

Principal component analysis (PCA) of blood gene expression revealed a moderate differentiation by severity in the first principal component (PC1; Fig. 1A). When we compared the whole COVID-19 cohort against controls we obtained nine DEG (Supplementary Table 2, Fig. 1A), all of them over-expressed in patients except for *FCER1A* gene that was under-expressed ($\log_2$FC = −1.7; adjusted *P*-value = 2.3E-5). The most statistically significant gene was *TNFRSF17* gene ($\log_2$FC = 1.9; adjusted *P*-value = 2.4E-10), followed by *FCER1A* and *SERPING1* genes. Three out of the nine DEG were found to be still significant when using a model that accounted for differences in treatment with immunomodulators (steroids), administered only in some severe and moderate patients: *TNFRSF17* ($\log_2$FC = 1.72; adjusted *P*-value = 4.1E-6), *SERPING1* ($\log_2$FC = 2.2; adjusted *P*-value = 4.3E-4) and *MX1* ($\log_2$FC = 1.97; adjusted *P*-value = 3.0E-3) genes. To test the robustness of this COVID-19-related immune signature we carried out a validation in two recently published COVID-19 RNA-seq datasets (Aschenbrenner et al., 2021; Thair et al., 2021) (Supplementary Fig. 3A). The 9-transcript immune signature correlated very well with the data from the two validation cohorts (R = 0.96, *P*-value = 5.3E-5; R = 0.92, *P*-value = 4.0E-4; respectively). As in our study, all genes were DE in both datasets when comparing patients *vs*. healthy controls.

Next, we studied the specificity of this 9-transcript signature by examining other viral and bacterial infections. Firstly, transcriptomes of different viral infections ($n = 655$) and healthy controls ($n = 672$) from Thair et al., (2021) were contrasted with the results from our study (Supplementary Fig. 3). We found a moderate correlation (R = 0.85; *P*-value = 0.0037) between $\log_2$FC values from our study (COVID-19 *vs*. control) and their study (viral infections *vs*. healthy controls). Notably, *TNFRSF17* gene did not show appreciable differences between non-COVID-19 viral infections and controls ($\log_2$FC = 0.16; adjusted *P*-value = 0.50). To further test the differential pattern of this gene in COVID-19 patients, we used a combined dataset from different microarray and RNA-seq studies that includes transcriptomes from viral and bacterial infected patients, and healthy controls. We consistently observed a moderate correlation between our results and those from the viral infections *vs*. healthy control dataset (R = 0.81, *P*-value = 0.0078), and a weaker correlation with those from bacterial infections *vs*. healthy control dataset (R = 0.61, *P*-value = 0.081). Again, *TNFRSF17* gene showed a $\log_2$FC value very close to 0, and non-significant results in the viral combined dataset ($\log_2$FC = −0.30; adjusted *P*-value = 0.28). Furthermore, in the bacterial combined dataset, the *TNFRSF17* gene (and *MX1*) was under-expressed in bacterial infections *vs*. controls, even though this contrast was also statistically significant ($\log_2$FC = −0.81, adjusted *P*-value = 1.9E-4) as in COVID-19 datasets.

We performed a more detailed analysis of the immunological processes underlying COVID-19 by comparing different severities (Fig. 1B; Fig. 1C; Supplementary Fig. 4). We found a similar number of DEG when comparing severe cases *vs*. healthy controls and severe *vs*. mild patients (28 and 30 respectively; Supplementary Table 2; 1D), pointing to a limited systemic immune response (Fig. 1D; Supplementary Fig. 4). We detected 14 common DEG in all comparisons involving severe cases (Fig. 1D); *ARG1* gene was the most remarkable one over-expressed in severe cases (Log_2FC = 3.98–4.71; adjusted *P*-value = 1.53E-18 – 1.78E-22) (Supplementary Table 2). Other significatively over-expressed genes in severe cases were those related to pro-inflammatory pathways *IL-1* (*IL1R1* and *IL1R2*) and *IL-18* (*IL18R1*) and immunoregulatory functions (*FKBP5* and *S100A8*); Supplementary Fig. 5. To further test this 14-

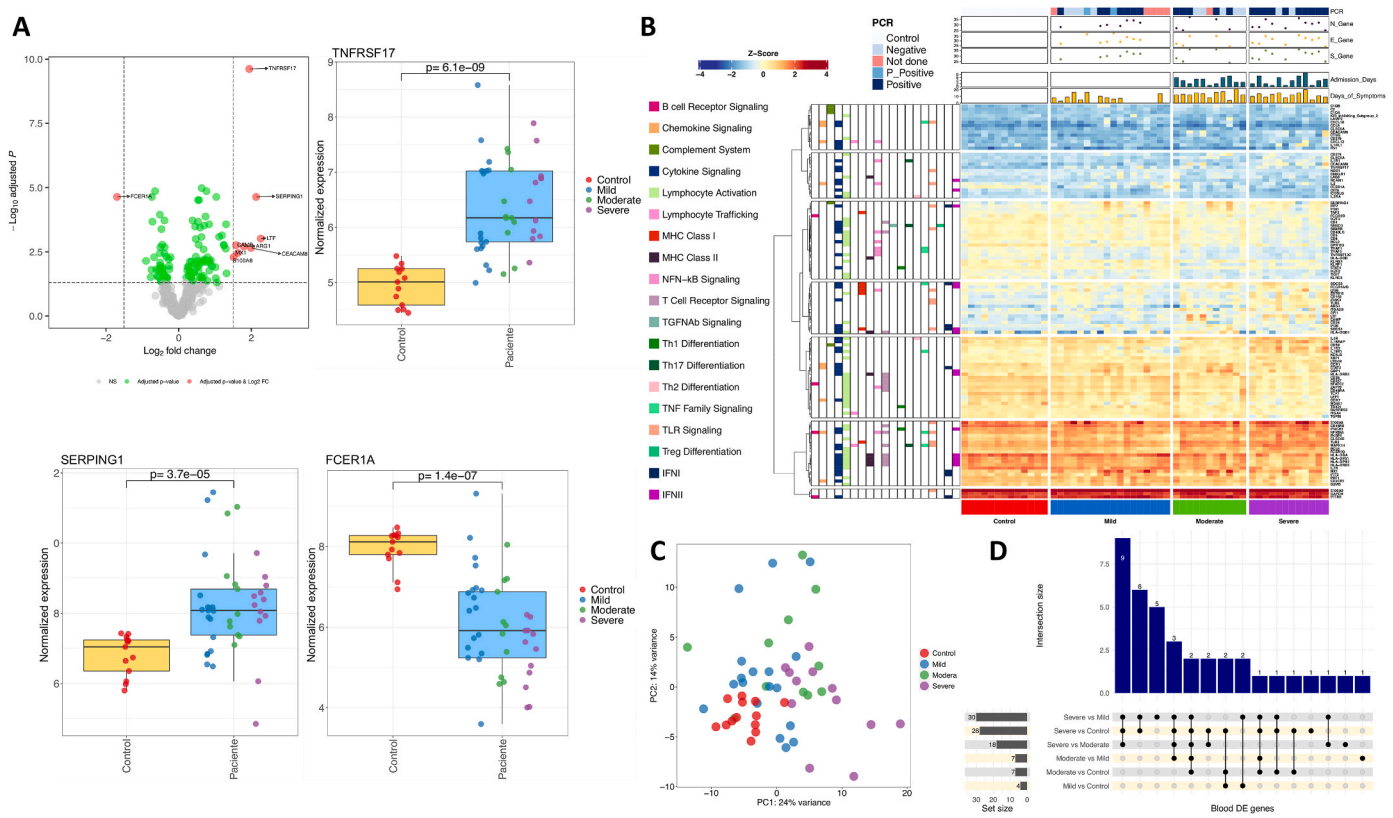


Fig. 1. A) Volcano plot showing the DEGs between cases and healthy controls in blood samples. Boxplots of the most relevant genes and the *P*-value of the statistical test are also shown. B) Heatmap and cluster analysis of the DEGs between all categories in blood samples. Only genes with a $\log_2$FC $|>1|$ were represented. Gene clusters were generated using *k*-means partitioning. C) PCA of immune transcriptomic data from blood samples of the different COVID-19 categories analyzed using all genes from the panel detected above the background. D) Upsetplot of the common DEGs between categories compared in blood samples.

transcript COVID-19 signature for severe patients, we compared our results to an independent cohort (Aschenbrenner et al., 2021); after excluding two genes (*ZBTB16* and *IL1RL1*) that were not found to be DE between their severe cases *vs*. healthy controls , we found a notable correlation between both datasets (R = 0.87; *P*-value = 0.00023); Supplementary Fig. 3.

Analysis of moderate *vs*. healthy controls and mild *vs*. healthy controls returned a few DEG coinciding with the previous 14-transcript signature, e.g. *TNFRSF17* and *SERPING1*. *TNFRSF17* represents the strongest signal of COVID-19 disease of all the genes analyzed, showing an increased gene expression associated with severity ($\log_2$FC values compared to healthy controls: 1.66 in mild, 1.95 in moderate, and 2.71 in severe cases); Supplementary Table 2. Conversely, *FCER1A* gene was under-expressed in patients and showed a gradual decreased gene expression with severity ($\log_2$FC values compared to healthy controls: not DE in mild, −1.83 in moderate, and −3.06 in severe cases). Genes *MX1* and *C1QB* showed an increased expression in mild ($\log_2$FC = 2.10; adjusted *P*-value = 1.32E-5) and moderate cases ($\log_2$FC = 1.78; adjusted *P*-value = 5.59E-5) with respect to controls (Supplementary Table 2). When comparing DEGs between hospitalized and non-hospitalized patients, the most relevant genes were *CAMP*, *LTF*, *CEACAM8* and *CEACAM6*, all involved in neutrophil degranulation processes triggered by the innate immune response (Supplementary Fig. 6).

Pathway analysis by way of a GSEA of severe cases *vs*. healthy controls and using Reactome revealed: (*i*) an up-regulation of innate immune processes (mainly driven by neutrophil activation and degranulation, the *IL-1* pathways, and the TLR cascade (Supplementary Table 3, Supplementary Fig. 7); and (*ii*) down-regulation of the adaptative immune response and related pathways such as Translocation of ZAP−70 to Immunological synapse, T cell antigen receptor (TCR) signalling, co-stimulation of CD28 family, and PD−1 signalling. Meanwhile, GSEA analysis using GO showed a comparable pattern: (i) up-regulation of the neutrophil mediated immunity and downstream processes; (*ii*) under-regulation of B cell proliferation/activation; and (*iii*) antigen receptor-mediated signalling (Supplementary Table 3, Supplementary Fig. 7). Similar findings were observed in comparisons of severe *vs*. moderate (Supplementary Table 3, Supplementary Fig. 8) and severe *vs*. mild cases (Supplementary Table 3, Supplementary Fig. 9) from Reactome, except for the IFN pathway, which was under-regulated in severe cases in comparison to moderate and mild patients. GO pathway analysis found under-expression of genes involved in antigen processing and presentation *via* MHC-II complex and viral defense. The only process detected by GSEA in mild cases were an up-regulation of the IFN signalling pathway with the Reactome database; whereas routes related to host response to viral infection were the most over-regulated in GO (Supplementary Table 3, Supplementary Fig. 10). Overall, there exists a significant up-regulation of the neutrophil activation and degranulation pathways, which increase with severity, and a lack of IFN response in severe cases. The differential neutrophil-driven immune response observed in moderate and severe patients was also detected when comparing hospitalized against non-hospitalized cases and healthy controls (Supplementary Fig. 11; Supplementary Table 3).

4.3. Transcriptomic immune response in nasal epithelium

PCA indicates a differentiation of severe cases against controls that is only notable in the PC2 dimension (12% of the variation); Fig. 2A (see also heatmaps in Fig. 2B). Conversely, mild and moderate cases showed more remarkable differences with severe and controls in the PC1 dimension (23% of the variation). Of the seven DEGs observed in the comparison patients *vs*. healthy controls, one was under-expressed and six over-expressed (Supplementary Table 4; Fig. 2C). Thus, *FCER1A* was

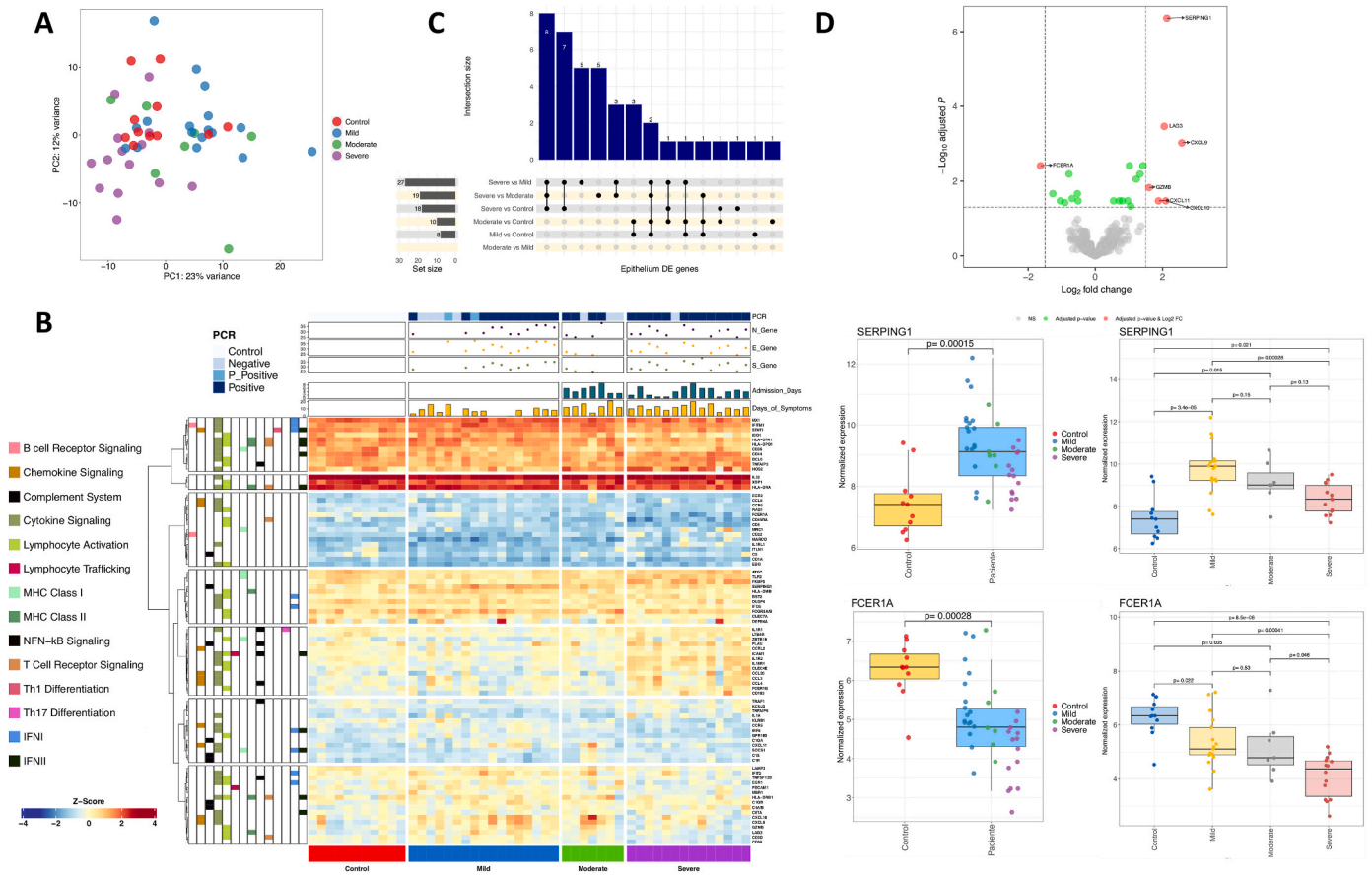


Fig. 2. A) PCA of immune transcriptomic data from nasal epithelium samples of the different disease categories analyzed using all genes from the panel detected above the background. B) Heatmap and cluster analysis of the DEGs between all categories in nasal epithelium samples. Only genes with a $\log_2$FC $|>1|$ were represented. Gene clusters were generated using *k*-means partitioning. C) Upsetplot of the common DEGs between categories compared in nasal epithelium samples. D) Volcano plot showing the DEGs between COVID-19 samples and healthy controls in nasal epithelium samples. Boxplots of the most relevant genes and the *P*-value of the statistical test are also shown.

under-expressed in patients when compared against controls ($\log_2$FC = −1.63; adjusted *P*-value = 0.004; severe *vs*. controls: $\log_2$FC = −3.28; adjusted *P*-value = 2.75E-9; severe *vs*. moderate: $\log_2$FC = −2.01; adjusted *P*-value = 0.0004; severe *vs*. mild: $\log_2$FC = −1.96; adjusted *P*-value = 0.00097), even though it was not detected as DE when comparing moderate and mild cases with healthy controls (Supplementary Table 4; note however that the absolute expression values showed only weak statistical significance in mild and moderate patients *vs*. healthy controls; Fig. 2D). Of the over-expressed genes, *SERPING1* gene leads the DEG list; the main signal comes from the high expression levels observed in moderate and mild patients against controls; this gene was however not DE in severe cases when compared with controls. Thus, *SERPING1* over-expression decreased gradually from mild to severe disease (Supplementary Table 4; Fig. 2D). A similar pattern was observed for other genes (*CXCL9*, *LAG3* and *GZMB*), all showing a gradual over-expression in moderate and mild cases against controls (Supplementary Table 4; Supplementary Fig. 12) but not in severe patients. Overall, gene expression of immune genes in nasal epithelium samples showed a very similar response to infection in mild and moderate patients, with seven overlapping DEGs (*LAG3*, *SERPING1*, *CXCL9*, *CXCL10*, *CXCL11, IFITM1* and *GZMB*; all over-expressed) between both categories *vs*. controls, but no DEG when comparing mild *vs*. moderate cases (Fig. 2C; Supplementary Fig. 13; Supplementary Fig. 14). Despite the similarities in immune response of both phenotypic categories, these genes showed a gradual increase in expression from severe to mild disease; the differences were found to be more notable when comparing mild *vs*. controls than in moderate *vs*. controls (Supplementary Table 4).

The comparison of severe *vs*. controls/moderate/mild yielded 18, 19 and 27 DEGs genes, respectively (Fig. 2C; Supplementary Table 4); eight genes were shared between these three comparisons (*FCER1A*, *IL18R1*, *IL1R2*, *CCL3 CCL4*, *CLEC4E*, *CD163* and *MARCO*), all over-expressed in severe cases (except *FCER1A*) and all mainly involved in the innate immune and inflammatory response (Supplementary Fig. 15). *IL18R1* and *IL1R2* appeared in the top three most DEGs in severe cases compared to all categories, with a $\log_2$FC ~2 in severe *vs*. moderate cases, and even higher in severe *vs*. controls/mild cohorts. Other under-expressed genes in severe compared to mild/moderate cases included *IFIT2* (Interferon Induced Protein with Tetratricopeptide Repeats 2) and *IDO1*, whose encoded protein is produced in response to inflammation and has autoimmune protective features (Supplementary Table 4; Supplementary Fig. 13).

We did not find statistically significant pathways in the nasal samples from severe cases *vs*. controls using both GSEA and ORA. Moderate *vs*. controls comparisons highlighted some up-regulated pathways in GSEA mainly connected to interaction and response to other organism processes (Supplementary Table 5), and inhibition of endopeptidase activity (Supplementary Fig. 16). Comparison of mild cases *vs*. controls using ORA suggested two main processes: (*i*) antimicrobial immune humoral response (an adaptative immune response mainly driven by B lymphocytes), and (*ii*) chemokine related pathways (Supplementary Table 5; Supplementary Fig. 17). The comparison of severe *vs*. mild patient revealed up-regulated coordinated responses in various nested pathways involving NAD metabolism (Supplementary Table 5), a regulator of infection and inflammation. Molecular functions showing down-

regulation in severe *vs*. mild cases were MHC protein binding and hydrolase activity (Supplementary Fig. 18). ORA analysis of DEGs revealed alterations in chemokine dynamics-related pathways, with over-expression of CC-type chemokines and *MARCO*, and under-expression of CXC-type chemokines and *DEFB4A* (Supplementary Fig. 18). The comparison of severe *vs*. moderate patients using GSEA analysis detected some similarities with that of severe *vs*. mild cases (Supplementary Fig. 19). Reactome and GO showed some common biological processes and/or cellular components, such as under-regulation of type I and II IFN pathway and MHC class II antigen presentation, and an over-regulation of neutrophil activation and degranulation processes. When exploring pathways associated with the comparison of moderate *vs*. mild cases, GSEA analysis found an upper-regulation of serine−type endopeptidase activity and a down-regulation of molecular transducer receptor activities (Supplementary Fig. 20; Supplementary Table 5).

4.4. Transcriptome immune response in saliva

Among the genes detected in saliva samples, there was no DEG when comparing patients and healthy controls. This fact was reflected in a PCA (and heatmap analysis), which showed a large cluster with overlapping controls, mild and moderate, and only slightly differentiation of the severe cases in the PC1 (35% of the variation) and the PC2 (20% of the variation); Fig. 3A,D. Therefore, the data seem to indicate very limited immune response to infection in the oral cavity of mild and moderate patients. A few DEGs were detected when comparing hospitalized with non-hospitalized patients (Fig. 3B; Supplementary Fig. 21) and against controls (Supplementary Table 6). As expected, most DEGs were the same in both comparisons, with 11 out of the 12 DEGs detected in hospitalized *vs*. non-hospitalized cases also appearing in hospitalized *vs*. controls, all of them over-expressed (Supplementary Fig. 21; Supplementary Table 6). The strongest differences in hospitalized *vs*. non-hospitalized patients were represented by (*i*) *PLAU*, a gene encoding a secreted serine endopeptidase that transforms plasminogen in plasmin and is implicated in the rearrangement of extracellular matrix and cell migration, and (*ii*) *LGALS3,* which plays a role in apoptosis, innate immunity (attracting macrophages and activating neutrophils), cell adhesion and T-cell regulation processes as well as in acute inflammatory responses (Fig. 3C). These two genes were the only ones detected as DE when comparing moderate *vs*. mild cases; they probably constitute the strongest and most specific immune signature in oral mucosa from hospitalized individuals (Supplementary Table 6; Supplementary Fig. 22). The other DEGs of the signature from hospitalized samples also exhibited expression levels that increased with severity (Supplementary Fig. 22).

We found 18 common over-expressed DEGs that differentiate severe cases from the other phenotypic categories (Fig. 3E; Supplementary Fig. 23). *CCL20* showed the highest $\log_2$FC value in all comparisons (Supplementary Table 6). As expected, higher differences were found in severe *vs*. mild and severe *vs*. controls (Fig. 3E). Eleven out of the 18 genes were also included in the signature that differentiated hospitalized from both non-hospitalized and healthy controls, respectively (Fig. 3E; Supplementary Fig. 24).

Pathway analysis in severe *vs*. controls using GSEA found some moderate up-regulation of processes and cell components, but only phagocytic vesicle as under-regulated cellular component

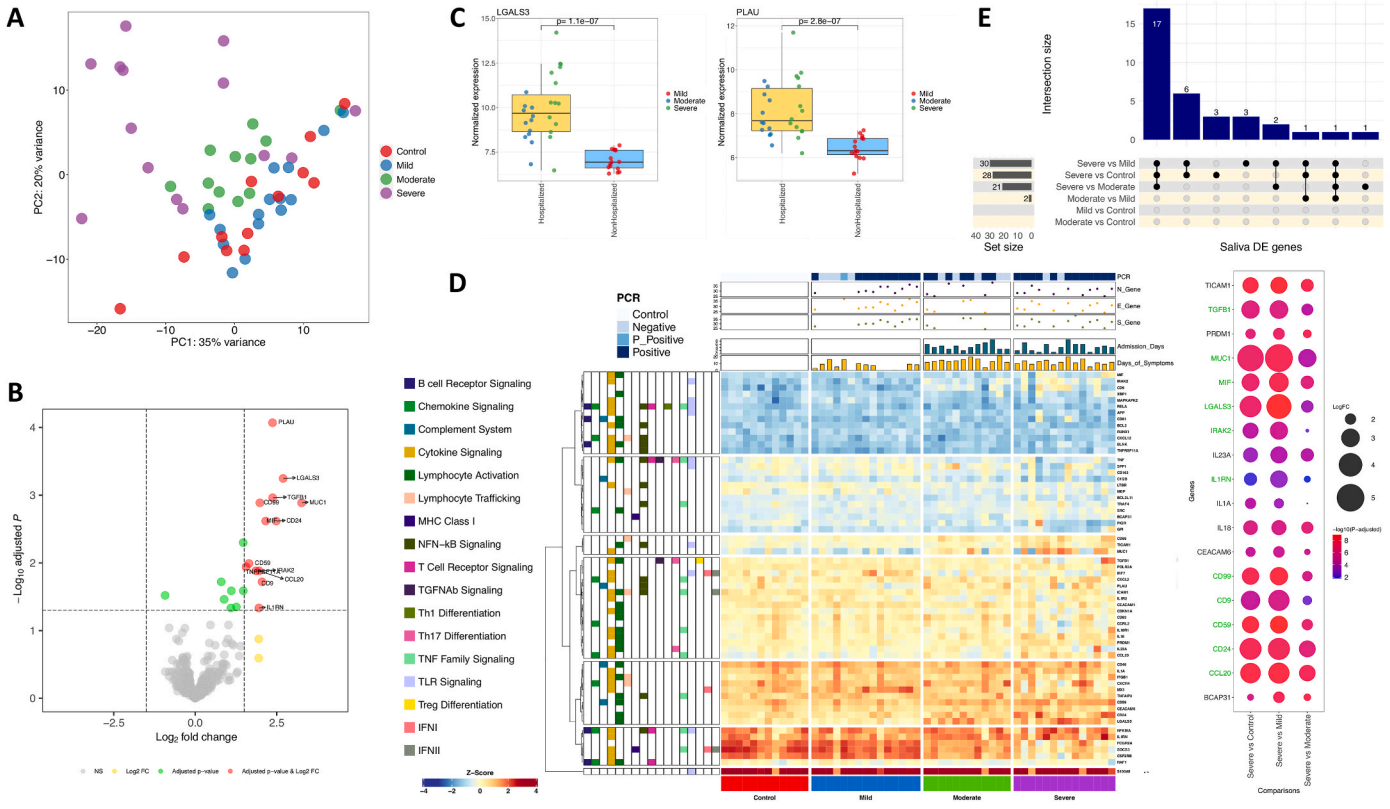


Fig. 3. A) PCA of immune transcriptomic data from saliva samples of the different disease categories analyzed using all genes from the panel detected above the background. B) Volcano plot showing the DEGs between hospitalized and non-hospitalized saliva samples. C) Boxplots of the most relevant genes from the comparison of hospitalized *vs*. non-hospitalized saliva samples and the *P*-value of the statistical test. D) Heatmap and cluster analysis of the DEGs between all categories in saliva samples. Only genes with a $\log_2$FC $|>1|$ were represented. Gene clusters were generated using *k*-means partitioning. E) Upsetplot of the common DEGs between categories compared in saliva samples. Bubble plot represents $\log_2$FC and adjusted *P*-value from the 18 DEGs between severe and all remaining categories. Green color in gene names indicates genes also included in the hospitalized signature. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

(Supplementary Fig. 25). Some of these pathways were also highlighted as up-regulated in severe *vs*. mild patients but, in this case, a positive regulation of cell migration and motility was also detected (Supplementary Table 7; Supplementary Fig. 26). GSEA analysis of severe *vs*. moderate patients provided the highest number of hits from all comparisons, all showing up-regulation in severe cases, but moderate NES values (Supplementary Table 7; Supplementary Fig. 27). Among the up-regulated pathways in moderate *vs*. mild (Supplementary Fig. 28) cases there are pathways related to cell migration, receptor activator activity and chemoattractant activity, of which the latter two were also up-regulated in moderate cases *vs*. controls (Supplementary Fig. 29). GSEA also pointed to a strong under-regulation of the IFN signalling in moderate *vs*. mild cases in all the reference databases interrogated (Supplementary Table 7; Supplementary Fig. 28). Notably, IFN pathway was not detected in severe *vs*. mild cases, however it appears as highly under-regulated in hospitalized *vs*. non-hospitalized patients, although not when comparing to controls.

Finally, comparison of mild cases and controls resulted in several under-regulated processes and cellular components mainly involved in cell junction, cell communication and secretory vesicles (Supplementary Table 7; Supplementary Fig. 30).

4.5. Transversal multi-tissue analysis

We selected a subset of severe and mild patients ($n = 21$) with samples available from all tissues collected at the same time. We explored for interactions between severity and tissue to elucidate immunological features that could illuminate the causes for the differential evolution observed in patients. Only when relaxing the thresholds for this interaction analysis (adjusted *P*-value < 0.05 and a cumulative $\log_2FC > |1.4|$) we found that the effect of severity for the *CXCL9* gene is significant for all tissue comparisons, being much lower in nasal compared to blood tissue (*P*-adjusted = 1.41E-6; cumulative $\log_2FC = -2.94$) and moderately lower for both saliva *vs*. blood (*P*-adjusted = 0.03; cumulative $\log_2FC = -1.42$) and nasal *vs*. saliva (*P*-adjusted = 0.017; cumulative $\log_2FC = -1.51$) tissues (Supplementary Table 8). Looking at the *CXCL9* normalized data, we found higher expression in severe than in mild cases for blood, with the opposite pattern in nasal and saliva epithelium samples (Supplementary Fig. 31). When we increased the significance thresholds (*P*-adjusted $<$ 10E-6; cumulative $\log_2FC > |2.5|$) a set of 17 genes were found as DE in all possible comparisons (Supplementary Table 8; Fig. 4B; Fig. 4C). Thirteen out of these 17 genes were also detected as DE between severe and mild cases in the longitudinal transcriptomic analysis (Fig. 4B). The transversal multi-tissue analysis generates four clusters for these 17 genes according to their cumulative Log_2FC values (Fig. 4C). This analysis highlights that the main specific severity signals in blood were driven by gene cluster 1 represented by *ARG1*, *CEACAM8*, *LTF* and *ITGA2B*, and to a lesser extent by cluster 2 (specially *IRAK3* and *TLR5*). This means that severity has a more marked effect in blood in comparison with nasal and saliva samples (Fig. 4C; Supplementary Table 8; Supplementary Fig. 31). Cluster 3 (Fig. 4C), represented by *CD22* and *TRAF1*, indicates that severity determines an opposite behavior in nasal (lower expression in mild and higher in severe) than in blood (higher expression in mild and lower in severe (Supplementary Fig. 31). Finally, cluster 4 (Fig. 4C) represented by seven genes indicates a differential effect of severity in saliva with respect to blood and nasal epithelium (especially remarkable for *MUC1* and *LGALS3*).

4.6. Co-expression analysis related to the mild-severe phenotypes

Through a co-expression network analysis, we aimed to detect modules of co-expressed genes that might be responsible for the extremely unbalance immune response to the SARS-CoV-2 infection

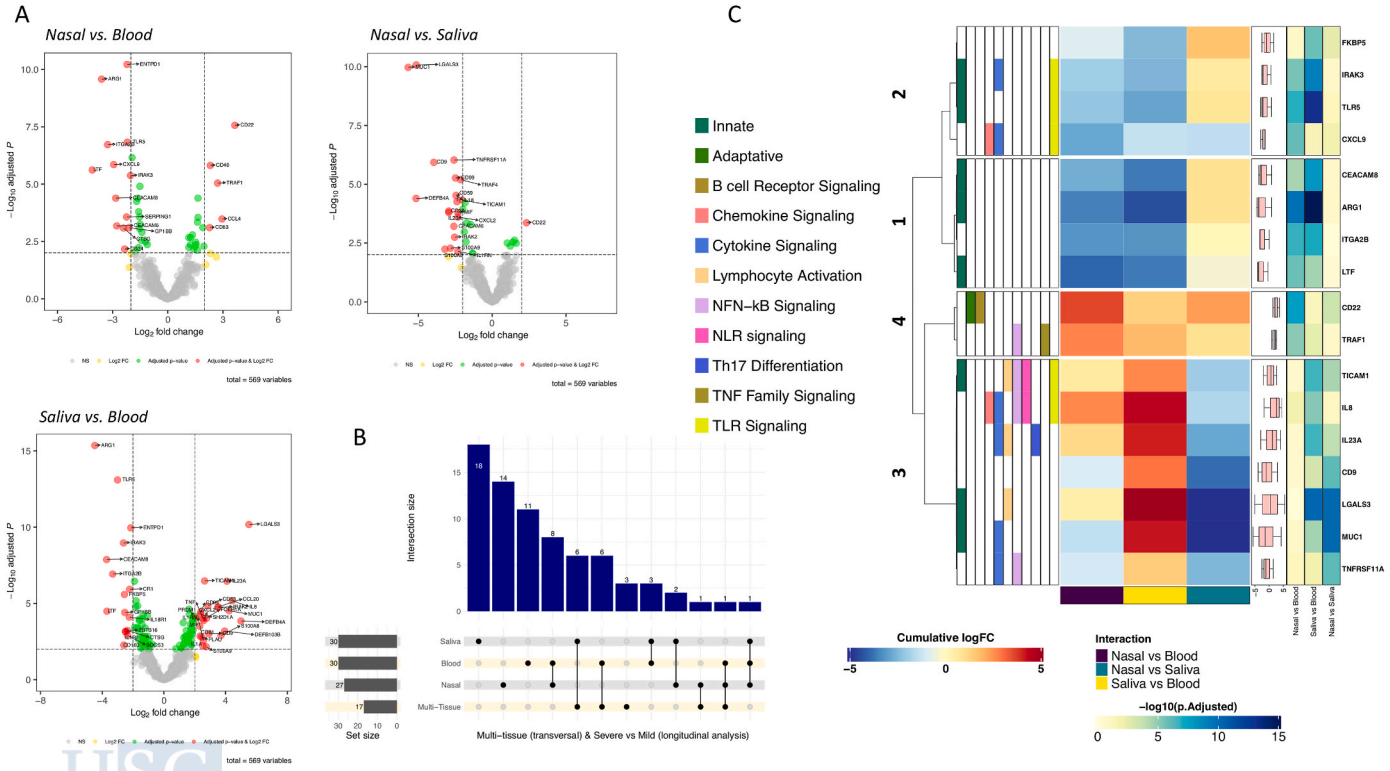


Fig. 4. A) Volcano plots showing the DEGs between severities (severe/mild) and different tissue samples (thresholds: *P*-adjusted = 0.01, cumulative $\log_2FC = |2|$). B) Upsetplot of the DEGs between severe and mild categories in all tissues as well as in the multi-tissue transversal analysis. C) Heatmap and cluster analysis of the DEGs between severities (severe/mild) and different tissue samples (thresholds: *P*-adjusted = 10E-6, cumulative $\log_2FC = |2.5|$). Gene clusters were generated using *k*-means partitioning.

between mild and severe phenotypes.

We detected four co-expression clusters of immune genes in blood samples (Supplementary Fig. 32A; Supplementary Table 9) constituting two meta-modules of highly related clusters (Supplementary Fig. 32B). Two of these modules were weakly significantly correlated with the mild-severe phenotype, one indicating over-expression ('blue' module in Supplementary Fig. 32; $n = 62$ genes; R = 0.41; *P*-value = 0.026) and the other under-expression in severe cases ('turquoise' module in Supplementary Fig. 32; $n = 105$ genes; R = −0.37; *P*-value = 0.043) (Supplementary Fig. 32C). Despite of not having large correlation with the phenotype, both modules showed highly significant correlation between gene significance (GS) and module membership (MM), meaning that genes associated to the phenotype were also hub genes within the module (higher number of connections) (Supplementary Figs. 33A and 33B; Supplementary Table 9). Hub genes of the most remarkable module (over-expressed in severe cases) are most likely involved in pathways related to cytokine production, IL-1 signalling, neutrophil mediated immunity, T-cell immunity, TLR pathways and NAD metabolism (Supplementary Fig. 34; Supplementary Table 9).

Stronger co-expression signals related to severity were detected in nasal epithelium and saliva samples. Two out of the four modules detected in nasal epithelium samples were statistically significant correlated with the phenotype, one composed by genes under-expressed in severe with respect to mild (see 'turquoise' module in Supplementary Figs. 35A, 35B, 35C; $n = 129$ genes; R = −0.46; *P*-value = 0.0099) and the other showing over-expression in severe cases (see 'brown' module in Supplementary Figs. 35A, 35B, 35C; $n = 48$ genes; R = 0.44; *P*-value = 0.014); see also Supplementary Table 10. The most important under-expressed genes in severe cases obtained in the DE analysis were also found among the genes placed in the 'turquoise' module and, by contrast, the top over-expressed genes in DE analysis from severe cases were housed in the 'brown' module, being many of them also hub genes in these modules (Supplementary Figs. 36A and 36B Supplementary Table 10). The core gene from the mild-related 'turquoise' module, defined as the significant correlated gene showing the higher MM value, was *CD3D* (MM = 0.95; *P*-value = 8.7E-17) and is part of the TCR/CD3 complex in the T-cell signal transduction with an essential role in the adaptative immune response. The core gene in the 'brown' module was *CCRL2* (MM = 0.91; *P*-value = 2.2E-12), a gene that is over-expressed during the neutrophil activation and the differentiation from monocyte to macrophage. Hub genes from these modules were also involved in different pathways (most of them overlapping to those obtained from the DE analysis) pointing to a very different response in mild and severe cases. Hub genes in 'turquoise' module were involved in CXCR chemokine receptor binding, interferon gamma signalling and adaptative T-cell immune response, such as T-cell migration, activation and chemotaxis, TCR signalling, and MHC class II receptor activity. Nevertheless, noteworthy pathways obtained from hub genes of the 'brown' module were those related to CCR chemokine receptor binding, IL-1 response, NAD metabolism, neutrophil driven immune response and TLR cascades (Supplementary Fig. 37).

From the four co-expression modules detected from saliva samples only one module showed significant correlation with the phenotype ('turquoise' module in Supplementary Figs. 38A and 38B; Supplementary Table 11) and was also the most significant phenotype-correlated module in all tissues analyzed ($n = 91$ genes; R = 0.52; *P*-value = 0.0039; Supplementary Fig. 38C). Genes from this module were over-expressed in severe cases and displayed a good significant correlation between GS and MM (Supplementary Fig. 39A). Most genes within this module (86%) were among the top DE genes and many of them were also hub genes within the module (63%). The gene with the highest intra-modular connectivity was *SRC* gene (MM = 0.95; *P*-value = 3.48E-15; Supplementary Table 11), a non-receptor tyrosine-kinase implicated in many signalling pathways controlling different processes such as immune response, cell adhesion and migration, cell cycle and apoptosis. Among the most significant pathways we found lymphocyte proliferation and activation, interleukin signalling, apoptosis, cell-cell adhesion and cell cycle arrest (Supplementary Fig. 39A, Supplementary Fig. 40).

5. Discussion

Healthcare systems worldwide collapsed during the COVID-19 pandemic due to the high number of hospital admissions, mostly with respiratory complications that, in many cases, ended up being fatal in patients with severe disease. Usually, viral colonization begins in a local tissue environment, with the upper respiratory tract being the main gateway for the virus entrance; however, the infection does not always have a significant systemic impact. To explore differential host response to infection, we analyzed the expression of immune response genes in three different tissues from COVID-19 patients affected by different disease severities.

In the analysis of blood samples, we detected specific COVID-19 DEGs, all but one (*FCER1A*) over-expressed, which were subsequently validated in two independent RNA-seq datasets. A few DEGs have comparable expression levels in other viral and bacterial infections, suggesting a common response against severe infections. Moreover, *ARG1* and *FCER1A* showed similar expression values independently of the causative pathogen origin (viral or bacterial), but with opposed $\log_2$FC values (*ARG1*: 1.5 to 3.5; *FCER1A*: −2.5 to −1.5). *ARG1* encodes for the Arginase-1 protein and is expressed in neutrophils, for which a central role in innate immune and inflammatory response has been described, while *FCER1A* is mainly expressed in basophils and mast cells and is involved in the initiation of allergic response. We also observed that *TNFRSF17* was DE in COVID-19 compared to other infections. This gene is expressed in mature B lymphocytes encoding for a member of the TNF-receptor superfamily; it is involved in MAPK8/JNK and NF-kB activation pathways and mediates the development and survival of B-cell lymphocytes to maintain humoral immunity. The number of DEGs in blood increases significantly with severity (Supplementary Fig. 4), clearly pointing to a lower immune systemic response in mild and moderate than in severe patients. Moreover, mild and moderate patients over-expressed the interferon type I (IFN–I) pathway in response to viral infection; a timely IFN-I response is critical as it represents a successful front-line protection against viral infections. IFN-I is produced by cells after pathogen-associated molecular patterns (PAMPs) recognition, inducing downstream expression of IFN stimulated genes (ISGs). Furthermore, it has an important role in immune control through the regulation of proinflammatory cytokines, activation and recruitment of different immune cells types, and shaping a successful adaptative immune response (Crouse et al., 2015). SARS-CoV-2 seems to be more susceptible to IFN-I and produces various proteins that antagonize IFN production through different mechanisms, suppressing IFN-I pathway in a more efficient manner than other coronaviruses (Lokugamage et al., 2020; Mantlo et al., 2020). Interestingly, low IFN-I alpha level has been recently related to a more severe disease outcome, and proposed as a potential therapeutic treatment (Hadjadj et al., 2020; Contoli et al., 2021). The IL-1 pathway related genes were detected as over-activated in severe cases. Conversely, moderate and severe patients showed an up-regulation of innate immune response (not observed in mild patients and healthy controls), driven by a different degree of neutrophil response pathways activation. Severe disease was associated with high neutrophil counts and increased levels of pro-inflammatory cytokines, producing microthrombus formation and organ damage. The release of neutrophil extracellular traps (NETs) could explain (at least partially) COVID-19 vascular complications through forming aggregates that produce vessel occlusion and endothelial damage (Leppkes et al., 2020; Veras et al., 2020); unbalanced NET production seems to have an important role in viral immunopathology and respiratory disease complications (Twaddell et al., 2019). Consistently, we observed a DE of *S100A8/S100A9* and *CAMP* genes (all encoding proteins usually identified in NETs) in severe cases *vs*. all other categories. Likewise, the

multi-tissue transversal analysis consistently indicated a higher difference in severity between systemic and local immune response in neutrophil related genes (*ARG1*, *LTF*, *CEACAM8*), and other blood coagulation genes that participate in the process of platelet aggregation (*ITGA2B*), which could explain the formation of microthrombus in severe phenotypes. Patterns of exacerbated innate response in hospitalized individuals were accompanied by a global failure of the adaptative immune response in severe cases, with under-regulation of several routes involved in T-cell signalling and maintenance of immune homeostasis, like ZAP-70 signalling. ZAP-70 kinase regulates TCR downstream signaling of antigen presentation, and an adequate regulation of its activity is key for a proper T cell proliferation; furthermore, ZAP protein can restrict SARS-CoV-2 replication (Nchioua et al., 2020). It has been described the importance of a solid T cell response in COVID-19 recovery (Zheng et al., 2020a) and, therefore, the suppression of the adaptative immune system could explain in part the characteristic lymphopenia observed in severe patients. In agreement with an over-activation of the innate and under-activation of adaptative immune response, an unbalanced activation of HLA class I and II antigen presentation was also found in severe cases. Overall, our results are in the line with previous gene expression studies carried out in blood samples (Zheng et al., 2020a; Yan et al., 2021), indicating that severe cases exhibit signals of a systemic impairment of IFN response characterized by a non-robust IFN production, a suppression of T and B responses and lymphopenia, neutrophil recruitment (with over-expression of ARG1 gene), abnormal increase of low density neutrophils (with high capacity to release NETs) and also enrichment in blood coagulation–related genes.

The most prominently over-expressed genes in the nasal epithelium of severe cases were anti-inflammatory IL-1 family receptors IL-1R2 and IL-18R1 (mainly expressed by monocytes-macrophages). Both receptors are involved in the IL-1 and IL-18 cascades, which play a role in the initiation of the inflammation process after PAMPs recognition. Nevertheless, we did not detect DE of the pro-inflammatory cytokines *IL1A*, *IL1B*, *IL1RAP* or *IL18* in severe samples, just a moderate up-regulation of *IL1R1* gene. This imbalance between agonists-antagonists/soluble receptors in severe disease with respect to mild cases was previously noted in broncho-alveolar lavage fluids (BALF) from COVID-19 patients (Xu et al., 2020), and it could lead to exacerbated inflammatory responses (Italiani et al., 2020). In addition, pathways analysis detected an alteration of NAD metabolism in severe cases with respect to those with mild disease, probably pointing to a NAD (and ATP) depletion in host cells from severe patients. It is known that some intracellular pathogens can modulate host cell NAD, impairing several innate immune routes (Mesquita et al., 2016). Nasal samples from severe cases also showed similar expression patterns of CXC-chemokines, *GZMB* and antiviral *IFITM1/IFIT2* genes compared to controls. Unlike in mild cases, in which main signals came from cell-mediated immunity (Th1, NK and T cytotoxic cells responses) as well as from a robust IFN-II induction, severe cases showed an over-expression of genes associated to monocyte-macrophage recruitment; these cells could play a key role in the cytokine storm observed in severe patients (Merad and Martin, 2020). For instance, CC pro-inflammatory chemokines genes *CCL3* and *CCL4* act as chemoattractant cytokines in response to tissue injury or inflammation. Some of them have already been highlighted as severity-related biomarkers of COVID-19 in airway epithelial cells (Chua et al., 2020), and they were found highly expressed in BALF of severe compared to mild cases (Xu et al., 2020). They are ligands of chemokine receptors CCR1 (CCL3) and CCR5 (CCL3 and CCL4) and have a role in monocyte recruitment. *CD163*, *CLEC4E* and *MARCO* are expressed in macrophages, and the latter two can act as pattern recognition receptors (PRR) for PAMPs. The protein encoded by *CLEC4E* can interact with *FCER1G* which, unlike *FCRE1A* gene, is over-expressed in severe cases when compared to controls. Nasal epithelium response in moderate and mild patients was mainly characterized by an over-expression of CXCR3 receptor ligands chemokines *CXCL9*, *CXCL10* and *CXCL11*. These chemokines are highly inducible by IFN gamma, and once joined to the receptor IFN is activated and triggers calcium release to the cytosol, acting as a second messenger to activate downstream signalling pathways (also highlighted in the pathway analysis of nasal samples from mild patients). These chemokines were previously described as biomarkers of respiratory viral infection in nasopharyngeal samples (Landry and Foxman, 2018). Similarly, other antiviral genes were also over-expressed in mild cases: *IFITM1* has antiviral activity against several enveloped viruses, hampering the viral membrane fusion, and has demonstrated a strong inhibition to SARS-CoV-2 infection by blocking the virus entrance into cells (Buchrieser et al., 2020); and *IFIT2* is involved in the inhibition of viral replication.

Co-expression network analysis yielded comparable results to those observed in DE analysis, reporting two significant modules with opposite behavior in nasal epithelium of severe *vs*. mild cases. One of the modules consists of genes showing over-expression in the mild phenotype, while the other includes genes over-expressed in severe patients. Hub genes from the mild-related module are mostly implicated in CXCR chemokine receptor binding processes, interferon gamma signalling and adaptative T-cell immune response. By contrast, hub genes within the severe-related module were linked to IL-1 response, CCR chemokine receptor binding, NAD metabolism and an innate response involving neutrophils and TLR cascades.

Contrary to the expression profile observed in nasal epithelium samples, there is an absence of appreciable DE in the saliva of mild cases, suggesting a lack of oral immune response against viral infection. Nevertheless, in severe, and to a lesser extent in moderate samples, some up-regulated severity related hits were detected, many of them linked to the viral dynamics. The virus has mechanisms to hijack cellular machinery to their own benefit, altering the levels of several metabolites and cellular processes; one of these molecules are galectins, which have been proposed as biomarkers to monitor viral infections as well as therapeutic targets or antagonists (Wang et al., 2020). One of the most remarkable over-expressed hits in moderate and severe saliva samples was *LGALS3* (galectin-3), a contributor to pro-inflammatory response against different viral infections (Wang et al., 2020), enhancing macrophage survival, and positively regulating macrophage recruitment (Liu and Hsu, 2007), even promoting a more severe course of the infection (Chen et al., 2018). Elevated levels of galectin-3 were observed in macrophages, monocytes and dendritic cells in severe *vs.* mild COVID-19 patients (Caniglia et al., 2020); the regulatory role of galectin-3 in inflammatory response, fibrosis and disease progression, makes it a promising therapeutic target for severe cases (Garcia-Revilla et al., 2020). We did not observe systemic or nasal epithelium over-expression of this gene. Another important over-expressed gene in saliva was *PLAU* (urokinase-type plasminogen activator), which, together with *CD59* (inhibitor of complement-mediated lysis) and *CD9* (tetraspanin), can facilitate virus-host cell interaction, host immune system viral evasion and, thus, dissemination and viral pathogenesis through buccal epithelium. Urokinase transforms plasminogen to the active form serine protease plasmin; this airway protease (like several others e.g. trypsin, cathepsins, elastase, and *TMPRSS* family members) are required for coronavirus entry enhancement (Millet and Whittaker, 2015). After viral interaction with *ACE2* receptor, furin site in S protein of SARS-CoV-2 is proteolytically cleaved by *TMPRSS2* to enable viral internalization (Devaux et al., 2020). Plasmin has demonstrated to cleave several viral proteins of many respiratory pathogens (Dubovi et al., 1983; Goto et al., 2001; Berri et al., 2013) and also S protein in SARS-CoV *in vitro* (Kam et al., 2009), increasing the ability of the virus to enter host cells. Even though the plasmin protein S cleavage in SARS-CoC-2 has not been demonstrated, there is growing evidence pointing towards a probable role of the plasmin-plasminogen system and fibrinolytic pathway in the COVID-19 severity (Gralinski et al., 2013; D'Alonzo et al., 2020; Ji et al., 2020; Kwaan and Lindholm, 2021). Co-expression and upregulation of *PLAU*, *TMPRSS2*, and *ACE2* was observed in airway epithelial cells from severe/moderate patients and

SARS-CoV-2 infected cell lines. This, together with a plasmin-mediated cleavage, could make cells more susceptible to infection (Hou et al., 2021). Interestingly, we also observed an enhanced expression of *PLAU* in nasal epithelium from severe cases when compared to mild and controls, highlighting the crosslink between both tissues. In the same vein, tetraspanins are molecules associated to specific viruses and seem to play a role in different stages of viral infections (Martin et al., 2005). Tetraspanin CD9 was involved in early MERS-CoV lung cell entry (Earnest et al., 2017). Recently, it has also been proposed that extracellular vesicles may contribute to viral dissemination by transferring viral receptors, such as CD9 and ACE2, to other host cells making them more susceptible to the infection after exosomes endocytosis (Hassanpour et al., 2020). Our functional analysis revealed down-regulation of some pathways related to cell communication, adhesion and signal transduction in mild cases, probably as a mechanism to minimize viral spread. Conversely, severe cases showed up-regulation of genes involved in cell junction, adhesion and extracellular exosomes, all processes that could facilitate the viral dissemination. CD59 is a non-IFN inducible protein that acts as inhibitor of complement-mediated lysis by disruption of the membrane attack complex. Its implication in viral complement evasion of many enveloped viruses (Li and Parks, 2018), including coronaviruses (Wei et al., 2017), has been demonstrated. This escape mechanism can be mediated by incorporating CD59 to the viral envelope, controlling cellular CD59 expression or producing a counterpart protein that mimics host CD59, or even use it as cell entry receptor (Maloney et al., 2020).

We have found an important co-expression module defining the severe phenotype in saliva samples containing all these aforementioned genes. In addition, most of them were also hub genes in the module, indicating that genes highly significantly associated with the severe phenotype were also key elements within this module. These hub genes were related to lymphocyte proliferation and activation, but also to apoptosis, cell-cell adhesion and cell cycle arrest. The gene that was found to be more connected within the module was the *SRC* gene, a member of the Src family kinases (SFKs) that demonstrated to have an important role in the life cycle of a many viruses and have emerged as an important factor in the interplay between cell functions and viral demands (Pagano et al., 2013). It has been shown that inhibitors of SFKs can significantly reduce the replication of MERS-CoV (Shin et al., 2018) and some candidates are being proposed as treatment for COVID-19 because of its potential to effectively block the viral dissemination (Weisberg et al., 2020; Pillaiyar and Laufer, 2021).

Overall, immune response in buccal cavity, due to its anatomical location and interplay between upper and lower respiratory tract (LRT), could play a key role in viral spread and transmission not only at inter-individual level (Huang et al., 2021) but also in intra-individual viral dissemination to LRT, leading to a more severe phenotype.

Transversal analysis of different tissue samples from the same patients showed a tissue-specific behavior depending on illness severity. Most of the tissue-specific severity signals observed were represented by genes involved in innate immune response and cytokine signaling. Genes that appeared differentially expressed with severities in nasal and blood samples are related to leukocyte chemotaxis, cytokine-cytokine signaling, cell adhesion and extracellular exosomes, with higher expression of severe cases (with respect to those with mild disease) in blood and, in contrast, lower expression in nasal samples; B-cell antigen receptor signaling genes showed an opposite pattern. Specific expression signatures of severity in saliva samples correspond to genes involved in inflammatory and NK activation routes, cell adhesion, and vesicle trafficking.

Our results point to a set of genes and pathways that are tissue-specific and correlated to severity (Fig. 5); many of these are involved in the innate immune system and cytokine/chemokine signaling pathways. Disease severity at systemic level might be strongly dependent on local immune response. Thus, the nasal epithelium shows a robust activation of the antiviral immune machinery in mild cases, characterized by IFN pathway involvement, thus triggering the IFN induced genes, chemokine release and transduction signaling in the cells, leading

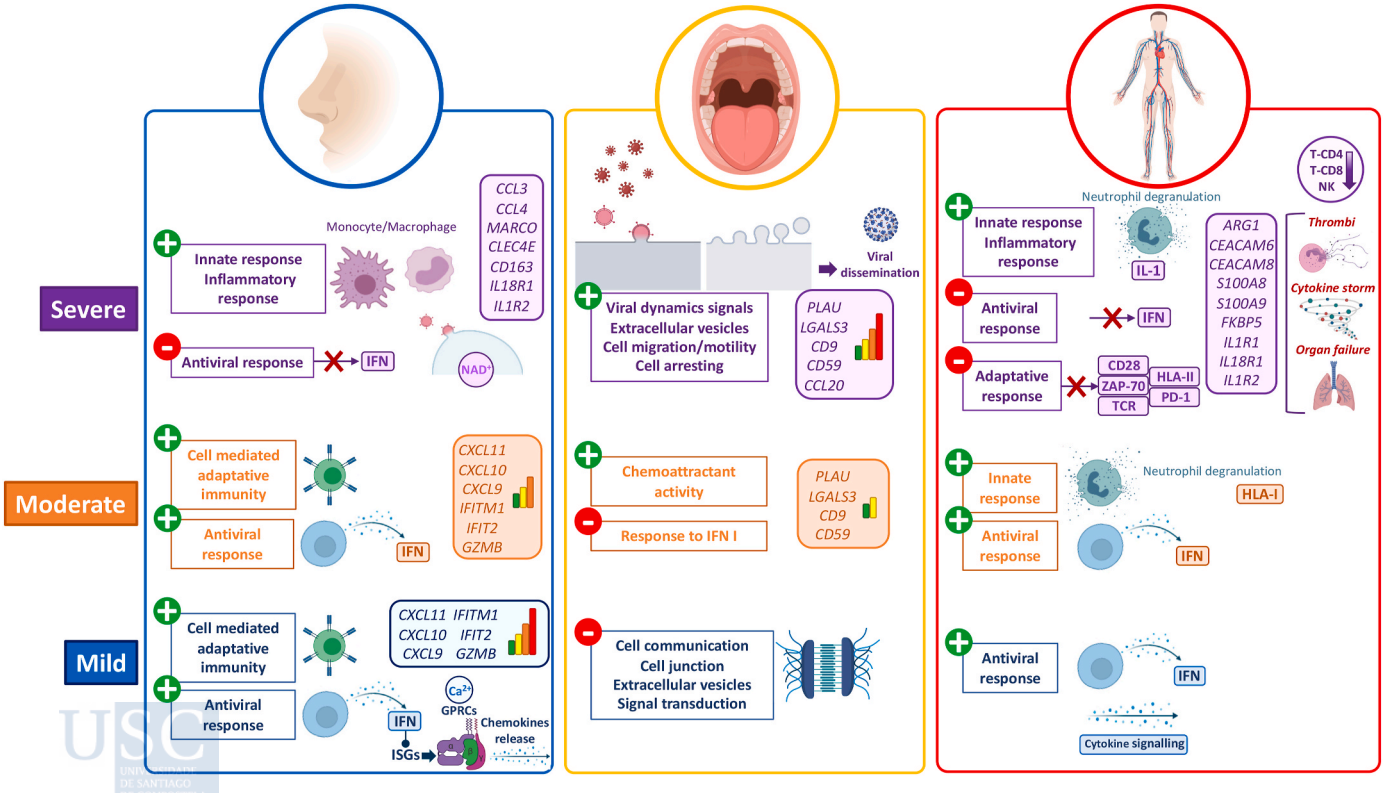


Fig. 5. Schematic representation of the main findings in gene expression and pathway analysis of COVID-19 severity in nasal, oral and blood tissues. The figure was built with Biorender (https://biorender.com/) resources.

to the activation of downstream cascades related to Th1, NK and T-cytotoxic adaptative immune response. Instead, severe cases showed these local antiviral reactions switched off; they develop an exacerbated innate response represented by an over-expression of genes related to monocyte-macrophage recruitment in nasal epithelium (Fig. 5). We did not find signals of immune viral response in buccal mucosa in mild patients, which probably indicates a successful local control of the infection and, therefore, prevention of viral dissemination and systemic colonization. On the contrary, buccal mucosa in moderate and, to a greater extent, severe cases, provided evident signals of viral activity, cell arresting and viral dissemination to the LRT (Fig. 5); ultimately generating an exacerbated innate and impaired adaptative systemic immune responses. In addition, the buccal cavity might play a key role in body infection dissemination: a severe phenotype seems to be associated to limited immunological resistance to the virus in the mouth, therefore facilitating the dissemination of the virus towards the LRT. This could make saliva not only a good prognostic proxy of severity, but also a source of therapeutic targets to contain viral dissemination in early stages of the disease.

The present study has a few limitations. The number of genes that could be explored is limited by the technology employed (*n-Counter* from NanoString); in counterpart, this is the only one available that allows the analysis of RNA expression patterns from different tissues (usually involving limited and/or degraded RNA). Although we used a sample size substantially higher than usual in transcriptomic studies, results from the present study should be further validated using a similar design and other independent cohorts and, ideally, performing functional studies aimed at validating the severity related pathways detected. The growing number of high-throughput gene expression data available from COVID-19 patients could also help to validate tissue specific findings in a near future.

Overall, our study highlights the key role of the nasal epithelium in COVID-19 severity, and include transcript biomarkers and pathways that signal severity at different tissue layers. The findings might illuminate new diagnostic, prognostic, or therapeutic targets in COVID-19.

Author contribution

AS and FMT conceived, designed, and provided financial support to the study. IRC, JGR, CRV, NRN, GBC, HPF, MLC, ADU, CRT, SP, MJCT, SVL and FMT were involved in sample recruitment and analysis of the clinical data. AGC carried out the experimental lab work. AS, AGC, JPS, XB, RBA analyzed the data. AGC and AS wrote the initial draft of the article. All the authors revised and contributed to the final version of the manuscript.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This study received support from Instituto de Salud Carlos III (ISCIII): GePEM (PI16/01478/Cofinanciado FEDER; A.S.), DIAVIR (DTS19/00049/Cofinanciado FEDER, A.S.), Resvi-Omics (PI19/01039/Cofinanciado FEDER, A.S.), ReSVinext (PI16/01569/Cofinanciado FEDER, F.M.T.), Enterogen (PI19/01090/Cofinanciado FEDER, F.M.T.); Agencia Gallega para la Gestión del Conocimiento en Salud (ACIS): BI-BACVIR (PRIS-3, A.S.), and CovidPhy (SA 304 C, A.S.); Agencia Gallega de Innovación (GAIN): Grupos con Potencial de Crecimiento (IN607B 2020/08, A.S.), GEN-COVID (IN845D 2020/23, F.M.T.); Framework Partnership Agreement between the Consellería de Sanidad de la XUNTA de Galicia and GENVIP-IDIS - 2021–2024 (SERGAS-IDIS march 2021); and consorcio Centro de Investigación Biomédica en Red de Enfermedades Respiratorias (CB21/06/00103; F.M.T.). We also thank Aida Freire Valls from Nanostring for her support.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.envres.2022.112890.

References

Alfi, O., Yakirevitch, A., Wald, O., Wandel, O., Izhar, U., Oiknine-Djian, E., Nevo, Y., Elgavish, S., Dagan, E., Madgar, O., et al., 2021. Human nasal and lung tissues infected *ex vivo* with SARS-CoV-2 provide insights into differential tissue-specific and virus-specific innate immune responses in the upper and lower respiratory tract. J. Virol.

Althouse, B.M., Wenger, E.A., Miller, J.C., Scarpino, S.V., Allard, A., Hebert-Dufresne, L., Hu, H., 2020. Superspreading events in the transmission dynamics of SARS-CoV-2: opportunities for interventions and control. PLoS Biol. 18, e3000897.

Aschenbrenner, A.C., Mouktaroudi, M., Krämer, B., Oestreich, M., Antonakos, N., Nuesch-Germano, M., Gkizeli, K., Bonaguro, L., Reusch, N., Baßler, K., et al., 2021. Disease severity-specific neutrophil signatures in blood transcriptomes stratify COVID-19 patients. Genome Med. 13, 7.

Barral-Arca, R., Pardo-Seco, J., Bello, X., Martinón-Torres, F., Salas, A., 2019. Ancestry patterns inferred from massive RNA-seq data. RNA 25, 857–868.

Barral-Arca, R., Pardo-Seco, J., Martinón-Torres, F., Salas, A., 2018. A 2-transcript host cell signature distinguishes viral from bacterial diarrhea and it is influenced by the severity of symptoms. Sci. Rep. 8, 8043.

Berri, F., Rimmelzwaan, G.F., Hanss, M., Albina, E., Foucault-Grunenwald, M.L., Lê, V.B., Vogelzang-van Trierum, S.E., Gil, P., Camerer, E., Martinez, D., et al., 2013. Plasminogen controls inflammation and pathogenesis of influenza virus infections via fibrinolysis. PLoS Pathog. 9, e1003229.

Bhattacharya, A., Hamilton, A.M., Furberg, H., Pietzak, E., Purdue, M.P., Troester, M.A., Hoadley, K.A., Love, M.I., 2020. An approach for normalization and quality control for NanoString RNA expression data. Briefings Bioinf.

Blighe, K., Rana, S., Lewis, M., 2020. EnhancedVolcano: publication-ready volcano plots with enhanced colouring and labeling. *version 180, R package*.

Buchrieser, J., Dufloo, J., Hubert, M., Monel, B., Planas, D., Rajah, M.M., Planchais, C., Porrot, F., Guivel-Benhassine, F., Van der Werf, S., et al., 2020. Syncytia formation by SARS-CoV-2-infected cells. EMBO J. 39, e106267.

Caniglia, J.L., Asuthkar, S., Tsung, A.J., Guda, M.R., Velpula, K.K., 2020. Immunopathology of galectin-3: an increasingly promising target in COVID-19. F1000Res 9, 1078.

Carvalho, B.S., Irizarry, R.A., 2010. A framework for oligonucleotide microarray preprocessing. Bioinformatics 26, 2363–2367.

Chen, Y.J., Wang, S.F., Weng, I.C., Hong, M.H., Lo, T.H., Jan, J.T., Hsu, L.C., Chen, H.Y., Liu, F.T., 2018. Galectin-3 enhances avian H5N1 influenza A virus-induced pulmonary inflammation by promoting NLRP3 inflammasome activation. Am. J. Pathol. 188, 1031–1042.

Chua, R.L., Lukassen, S., Trump, S., Hennig, B.P., Wendisch, D., Pott, F., Debnath, O., Thürmann, L., Kurth, F., Völker, M.T., et al., 2020. COVID-19 severity correlates with airway epithelium-immune cell interactions identified by single-cell analysis. Nat. Biotechnol. 38, 970–979.

Contoli, M., Papi, A., Tomassetti, L., Rizzo, P., Vieceli Dalla Sega, F., Fortini, F., Torsani, F., Morandi, L., Ronzoni, L., Zucchetti, O., et al., 2021. Blood interferon-α levels and severity, outcomes, and inflammatory profiles in hospitalized COVID-19 patients. Front. Immunol. 12, 648004.

Crouse, J., Kalinke, U., Oxenius, A., 2015. Regulation of antiviral T cell responses by type I interferons. Nat. Rev. Immunol. 15, 231–242.

D'Alonzo, D., De Fenza, M., Pavone, V., 2020. COVID-19 and pneumonia: a role for the uPA/uPAR system. Drug Discov. Today 25, 1528–1534.

Davies, N.G., Abbott, S., Barnard, R.C., Jarvis, C.I., Kucharski, A.J., Munday, J.D., Pearson, C.A.B., Russell, T.W., Tully, D.C., Washburne, A.D., et al., 2021. Estimated transmissibility and impact of SARS-CoV-2 lineage B.1.1.7 in England. Science 372.

DeBerg, H.A., Zaidi, M.B., Altman, M.C., Khaenam, P., Gersuk, V.H., Campos, F.D., Perez-Martinez, I., Meza-Segura, M., Chaussabel, D., Banchereau, J., et al., 2018. Shared and organism-specific host responses to childhood diarrheal diseases revealed by whole blood transcript profiling. PLoS One 13, e0192082.

Devaux, C.A., Rolain, J.M., Raoult, D., 2020. ACE2 receptor polymorphism: susceptibility to SARS-CoV-2, hypertension, multi-organ failure, and COVID-19 disease outcome. J. Microbiol. Immunol. Infect. 53, 425–435.

Du, P., Kibbe, W.A., Lin, S.M., 2008. lumi: a pipeline for processing Illumina microarray. Bioinformatics 24, 1547–1548.

Dubovi, E.J., Geratz, J.D., Tidwell, R.R., 1983. Enhancement of respiratory syncytial virus-induced cytopathology by trypsin, thrombin, and plasmin. Infect. Immun. 40, 351–358.

Earnest, J.T., Hantak, M.P., Li, K., McCray, P.B., Perlman, S., Gallagher, T., 2017. The tetraspanin CD9 facilitates MERS-coronavirus entry by scaffolding host cell receptors and proteases. PLoS Pathog. 13, e1006546.

Gamage, A.M., Tan, K.S., Chan, W.O.Y., Liu, J., Tan, C.W., Ong, Y.K., Thong, M., Andiappan, A.K., Anderson, D.E., Wang, Y., et al., 2020. Infection of human Nasal Epithelial Cells with SARS-CoV-2 and a 382-nt deletion isolate lacking ORF8 reveals similar viral kinetics and host transcriptional profiles. PLoS Pathog. 16, e1009130.

Garcia-Revilla, J., Deierborg, T., Venero, J.L., Boza-Serrano, A., 2020. Hyperinflammation and fibrosis in severe COVID-19 patients: galectin-3, a target molecule to consider. Front. Immunol. 11, 2069.
Gomez-Rial, J., Rivero-Calle, I., Salas, A., Martinon-Torres, F., 2020. Role of monocytes/macrophages in covid-19 pathogenesis: implications for therapy. Infect. Drug Resist. 13, 2485–2493.
Goto, H., Wells, K., Takada, A., Kawaoka, Y., 2001. Plasminogen-binding activity of neuraminidase determines the pathogenicity of influenza A virus. J. Virol. 75, 9297–9301.
Gralinski, L.E., Bankhead, A., Jeng, S., Menachery, V.D., Proll, S., Belisle, S.E., Matzke, M., Webb-Robertson, B.J., Luna, M.L., Shukla, A.K., et al., 2013. Mechanisms of severe acute respiratory syndrome coronavirus-induced acute lung injury. mBio 4.
Gu, Z., Eils, R., Schlesner, M., 2016. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. Bioinformatics 32, 2847–2849.
Gómez-Carballa, A., Bello, X., Pardo-Seco, J., Del Molino, M.L.P., Martinón-Torres, F., Salas, A., 2020a. Phylogeography of SARS-CoV-2 pandemic in Spain: a story of multiple introductions, micro-geographic stratification, founder effects, and super-spreaders. Zool. Res. 41, 605.
Gómez-Carballa, A., Bello, X., Pardo-Seco, J., Martinón-Torres, F., Salas, A., 2020b. Mapping genome variation of SARS-CoV-2 worldwide highlights the impact of COVID-19 super-spreaders. Genome Res. 30, 1434–1448.
Gómez-Carballa, A., Pardo-Seco, J., Bello, X., Martinón-Torres, F., Salas, A., 2021. Superspreading in the emergence of COVID-19 variants. Trends Genet. 37, 1069–1080.
Gómez-Rial, J., Currás-Tuala, M.J., Rivero-Calle, I., Gómez-Carballa, A., Cebey-López, M., Rodríguez-Tenreiro, C., Dacosta-Urbieta, A., Rivero-Velasco, C., Rodríguez-Núñez, N., Trastoy-Pena, R., 2020. Increased serum levels of sCD14 and sCD163 indicate a preponderant role for monocytes in COVID-19 immunopathology. Front. Immunol. 11, 2436.
Hadjadj, J., Yatim, N., Barnabei, L., Corneau, A., Boussier, J., Smith, N., Péré, H., Charbit, B., Bondet, V., Chenevier-Gobeaux, C., et al., 2020. Impaired type I interferon activity and inflammatory responses in severe COVID-19 patients. Science 369, 718–724.
Hassanpour, M., Rezaie, J., Nouri, M., Panahi, Y., 2020. The role of extracellular vesicles in COVID-19 virus infection. Infect. Genet. Evol. 85, 104422.
Herberg, J.A., Kaforou, M., Gormley, S., Sumner, E.R., Patel, S., Jones, K.D., Paulus, S., Fink, C., Martinon-Torres, F., Montana, G., et al., 2013. Transcriptomic profiling in childhood H1N1/09 influenza reveals reduced expression of protein synthesis genes. J. Infect. Dis. 208, 1664–1668.
Herberg, J.A., Kaforou, M., Wright, V.J., Shailes, H., Eleftherohorinou, H., Hoggart, C.J., Cebey-López, M., Carter, M.J., Janes, V.A., Gormley, S., et al., 2016. Diagnostic test accuracy of a 2-transcript host RNA signature for discriminating bacterial vs viral infection in febrile children. JAMA 316, 835–845.
Hou, Y., Ding, Y., Nie, H., Ji, H.L., 2021. Fibrinolysis Influences SARS-CoV-2 Infection in Ciliated Cells. bioRxiv.
Huang, C., Wang, Y., Li, X., Ren, L., Zhao, J., Hu, Y., Zhang, L., Fan, G., Xu, J., Gu, X., et al., 2020. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. Lancet 395, 497–506.
Huang, N., Pérez, P., Kato, T., Mikami, Y., Okuda, K., Gilmore, R.C., Conde, C.D., Gasmi, B., Stein, S., Beach, M., et al., 2021. SARS-CoV-2 infection of the oral cavity and saliva. Nat. Med.
Italiani, P., Mosca, E., Della Camera, G., Melillo, D., Migliorini, P., Milanesi, L., Boraschi, D., 2020. Profiling the course of resolving vs. Persistent inflammation in human monocytes: the role of IL-1 family molecules. Front. Immunol. 11, 1426.
Jain, R., Ramaswamy, S., Harilal, D., Uddin, M., Loney, T., Nowotny, N., Alsuwaidi, H., Varghese, R., Deesi, Z., Alkhajeh, A., et al., 2021. Host transcriptomic profiling of COVID-19 patients with mild, moderate, and severe clinical outcomes. Comput. Struct. Biotechnol. J. 19, 153–160.
Ji, H.L., Zhao, R., Matalon, S., Matthay, M.A., 2020. Elevated plasmin(ogen) as a common risk factor for COVID-19 susceptibility. Physiol. Rev. 100, 1065–1075.
Jin, X., Lian, J.S., Hu, J.H., Gao, J., Zheng, L., Zhang, Y.M., Hao, S.R., Jia, H.Y., Cai, H., Zhang, X.L., et al., 2020. Epidemiological, clinical and virological characteristics of 74 cases of coronavirus-infected disease 2019 (COVID-19) with gastrointestinal symptoms. Gut 69, 1002–1009.
Kam, Y.W., Okumura, Y., Kido, H., Ng, L.F., Bruzzone, R., Altmeyer, R., 2009. Cleavage of the SARS coronavirus spike glycoprotein by airway proteases enhances virus entry into human bronchial epithelial cells in vitro. PLoS One 4, e7870.
Kwaan, H.C., Lindholm, P.F., 2021. The central role of fibrinolytic response in COVID-19-A hematologist's perspective. Int. J. Mol. Sci. 22.
Lamers, M.M., Beumer, J., van der Vaart, J., Knoops, K., Puschhof, J., Breugem, T.I., Ravelli, R.B.G., Paul van Schayck, J., Mykytyn, A.Z., Duimel, H.Q., et al., 2020. SARS-CoV-2 productively infects human gut enterocytes. Science 369, 50–54.
Landry, M.L., Foxman, E.F., 2018. Antiviral response in the nasopharynx identifies patients with respiratory virus infection. J. Infect. Dis. 217, 897–905.
Langfelder, P., Horvath, S., 2008. WGCNA: an R package for weighted correlation network analysis. BMC Bioinf. 9, 559.
Lemieux, J.E., Siddle, K.J., Shaw, B.M., Loreth, C., Schaffner, S.F., Gladden-Young, A., Adams, G., Fink, T., Tomkins-Tinch, C.H., Krasilnikova, L.A., et al., 2020. Phylogenetic analysis of SARS-CoV-2 in Boston highlights the impact of superspreading events. Science 371, eabe3261.
Leppkes, M., Knopf, J., Naschberger, E., Lindemann, A., Singh, J., Herrmann, I., Stürzl, M., Staats, L., Mahajan, A., Schauer, C., et al., 2020. Vascular occlusion by neutrophil extracellular traps in COVID-19. EBioMedicine 58, 102925.
Li, Y., Parks, G.D., 2018. Relative contribution of cellular complement inhibitors CD59, CD46, and CD55 to parainfluenza virus 5 inhibition of complement-mediated neutralization. Viruses 10.
Lieberman, N.A.P., Peddu, V., Xie, H., Shrestha, L., Huang, M.L., Mears, M.C., Cajimat, M.N., Bente, D.A., Shi, P.Y., Bovier, F., et al., 2020. In vivo antiviral host transcriptional response to SARS-CoV-2 by viral load, sex, and age. PLoS Biol. 18, e3000849.
Liu, F.T., Hsu, D.K., 2007. The role of galectin-3 in promotion of the inflammatory response. Drug News Perspect. 20, 455–460.
Lokugamage, K.G., Hage, A., de Vries, M., Valero-Jimenez, A.M., Schindewolf, C., Dittmann, M., Rajsbaum, R., Menachery, V.D., 2020. Type I interferon susceptibility distinguishes SARS-CoV-2 from SARS-CoV. J. Virol. 94.
Love, M.I., Huber, W., Anders, S., 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 15, 550.
Mahajan, P., Kuppermann, N., Mejias, A., Suarez, N., Chaussabel, D., Casper, T.C., Smith, B., Alpern, E.R., Anders, J., Atabaki, S.M., et al., 2016. Association of RNA biosignatures with bacterial infections in febrile infants aged 60 Days or younger. JAMA 316, 846–857.
Maloney, B.E., Perera, K.D., Saunders, D.R.D., Shadipeni, N., Fleming, S.D., 2020. Interactions of viruses and the humoral innate immune response. Clin. Immunol. 212, 108351.
Mantlo, E., Bukreyeva, N., Maruyama, J., Paessler, S., Huang, C., 2020. Antiviral activities of type I interferons to SARS-CoV-2 infection. Antivir. Res. 179, 104811.
Martin, F., Roth, D.M., Jans, D.A., Pouton, C.W., Partridge, L.J., Monk, P.N., Moseley, G. W., 2005. Tetraspanins in viral infections: a fundamental role in viral biology? J. Virol. 79, 10839–10851.
McGonagle, D., Sharif, K., O'Regan, A., Bridgewood, C., 2020. The role of cytokines including interleukin-6 in COVID-19 induced pneumonia and macrophage activation syndrome-like disease. Autoimmun. Rev. 19, 102537.
Merad, M., Martin, J.C., 2020. Pathological inflammation in patients with COVID-19: a key role for monocytes and macrophages. Nat. Rev. Immunol. 20, 355–362.
Mesquita, I., Varela, P., Belinha, A., Gaifem, J., Laforge, M., Vergnes, B., Estaquier, J., Silvestre, R., 2016. Exploring NAD+ metabolism in host-pathogen interactions. Cell. Mol. Life Sci. 73, 1225–1236.
Millet, J.K., Whittaker, G.R., 2015. Host cell proteases: critical determinants of coronavirus tropism and pathogenesis. Virus Res. 202, 120–134.
Nchioua, R., Kmiec, D., Müller, J.A., Conzelmann, C., Groß, R., Swanson, C.M., Neil, S.J. D., Stenger, S., Sauter, D., Münch, J., et al., 2020. SARS-CoV-2 is restricted by zinc finger antiviral protein despite preadaptation to the low-CpG environment in humans. mBio 11.
Newman, A.M., Liu, C.L., Green, M.R., Gentles, A.J., Feng, W., Xu, Y., Hoang, C.D., Diehn, M., Alizadeh, A.A., 2015. Robust enumeration of cell subsets from tissue expression profiles. Nat. Methods 12, 453–457.
Newman, A.M., Steen, C.B., Liu, C.L., Gentles, A.J., Chaudhuri, A.A., Scherer, F., Khodadoust, M.S., Esfahani, M.S., Luca, B.A., Steiner, D., et al., 2019. Determining cell type abundance and expression from bulk tissues with digital cytometry. Nat. Biotechnol. 37, 773–782.
Pagano, M.A., Tibaldi, E., Palù, G., Brunati, A.M., 2013. Viral proteins and Src family kinases: mechanisms of pathogenicity from a "liaison dangereuse. World J. Virol. 2, 71–78.
Parnell, G.P., McLean, A.S., Booth, D.R., Armstrong, N.J., Nalos, M., Huang, S.J., Manak, J., Tang, W., Tam, O.Y., Chan, S., et al., 2012. A distinct influenza infection signature in the blood transcriptome of patients with severe community-acquired pneumonia. Crit. Care 16, R157.
Pasomsub, E., Watcharananan, S.P., Boonyawat, K., Janchompoo, P., Wongtabtim, G., Suksuwan, W., Sungkanuparph, S., Phuphuakrat, A., 2020. Saliva sample as a non-invasive specimen for the diagnosis of coronavirus disease 2019: a cross-sectional study. Clin. Microbiol. Infect.
Pierce, C.A., Sy, S., Galen, B., Goldstein, D.Y., Orner, E., Keller, M.J., Herold, K.C., Herold, B.C., 2021. Natural mucosal barriers and COVID-19 in children. JCI Insight 6.
Pillaiyar, T., Laufer, S., 2021. Kinases as potential therapeutic targets for anti-coronaviral therapy. J. Med. Chem.
Pipes, L., Wang, H., Huelsenbeck, J.P., Nielsen, R., 2020. Assessing uncertainty in the rooting of the SARS-CoV-2 phylogeny. Mol. Biol. Evol.
Qin, C., Zhou, L., Hu, Z., Zhang, S., Yang, S., Tao, Y., Xie, C., Ma, K., Shang, K., Wang, W., et al., 2020. Dysregulation of immune response in patients with coronavirus 2019 (COVID-19) in wuhan, China. Clin. Infect. Dis. 71, 762–768.
Risso, D., Ngai, J., Speed, T.P., Dudoit, S., 2014. Normalization of RNA-seq data using factor analysis of control genes or samples. Nat. Biotechnol. 32, 896–902.
Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W., Smyth, G.K., 2015. Limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 43, e47.
Sajuthi, S.P., DeFord, P., Li, Y., Jackson, N.D., Montgomery, M.T., Everman, J.L., Rios, C. L., Pruesse, E., Nolin, J.D., Plender, E.G., et al., 2020. Type 2 and interferon inflammation regulate SARS-CoV-2 entry factor expression in the airway epithelium. Nat. Commun. 11, 5139.
Salas, A., Pardo-Seco, J., Barral-Arca, R., Cebey-López, M., Gómez-Carballa, A., Rivero-Calle, I., Pischedda, S., Currás-Tuala, M.-J., Amigo, J., Gómez-Rial, J., 2018. Whole exome sequencing identifies new host genomic susceptibility factors in empyema caused by streptococcus pneumoniae in children: a pilot study. Genes 9, 240.
Salas, A., Pardo-Seco, J., Cebey-López, M., Gómez-Carballa, A., Obando-Pacheco, P., Rivero-Calle, I., Currás-Tuala, M.-J., Amigo, J., Gómez-Rial, J., Martinón-Torres, F., 2017. Whole exome sequencing reveals new candidate genes in host genomic susceptibility to respiratory syncytial virus disease. Sci. Rep. 7, 1–13.

USC DE COMPOSTELA

Shin, J.S., Jung, E., Kim, M., Baric, R.S., Go, Y.Y., 2018. Saracatinib inhibits Middle East respiratory syndrome-coronavirus replication in vitro. Viruses 10.
Suarez, N.M., Bunsow, E., Falsey, A.R., Walsh, E.E., Mejias, A., Ramilo, O., 2015. Superiority of transcriptional profiling over procalcitonin for distinguishing bacterial from viral lower respiratory tract infections in hospitalized adults. J. Infect. Dis. 212, 213–222.
Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., et al., 2005. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci. U. S. A. 102, 15545–15550.
Sungnak, W., Huang, N., Bécavin, C., Berg, M., Queen, R., Litvinukova, M., Talavera-López, C., Maatz, H., Reichart, D., Sampaziotis, F., et al., 2020. SARS-CoV-2 entry factors are highly expressed in nasal epithelial cells together with innate immune genes. Nat. Med. 26, 681–687.
Sweeney, T.E., Wong, H.R., Khatri, P., 2016. Robust classification of bacterial and viral infections via integrated host gene expression diagnostics. Sci. Transl. Med. 8, 346ra391.
Thair, S.A., He, Y.D., Hasin-Brumshtein, Y., Sakaram, S., Pandya, R., Toh, J., Rawling, D., Remmel, M., Coyle, S., Dalekos, G.N., et al., 2021. Transcriptomic similarities and differences in host response between SARS-CoV-2 and other viral infections. iScience 24, 101947.
To, K.K., Tsang, O.T., Yip, C.C., Chan, K.H., Wu, T.C., Chan, J.M., Leung, W.S., Chik, T.S., Choi, C.Y., Kandamby, D.H., et al., 2020. Consistent detection of 2019 novel coronavirus in saliva. Clin. Infect. Dis. 71, 841–843.
Twaddell, S.H., Baines, K.J., Grainge, C., Gibson, P.G., 2019. The emerging role of neutrophil extracellular traps in respiratory disease. Chest 156, 774–782.
Vandesompele, J., De Preter, K., Pattyn, F., Poppe, B., Van Roy, N., De Paepe, A., Speleman, F., 2002. Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes. Genome Biol. 3. RESEARCH0034.
Veras, F.P., Pontelli, M.C., Silva, C.M., Toller-Kawahisa, J.E., de Lima, M., Nascimento, D.C., Schneider, A.H., Caetité, D., Tavares, L.A., Paiva, I.M., et al., 2020. SARS-CoV-2-triggered neutrophil extracellular traps mediate COVID-19 pathology. J. Exp. Med. 217.
Wang, W.H., Lin, C.Y., Chang, M.R., Urbina, A.N., Assavalapsakul, W., Thitithanyanont, A., Chen, Y.H., Liu, F.T., Wang, S.F., 2020. The role of galectins in virus infection - a systemic literature review. J. Microbiol. Immunol. Infect. 53, 925–935.
Wei, Y., Ji, Y., Guo, H., Zhi, X., Han, S., Zhang, Y., Gao, Y., Chang, Y., Yan, D., Li, K., et al., 2017. CD59 association with infectious bronchitis virus particles protects against antibody-dependent complement-mediated lysis. J. Gen. Virol. 98, 2725–2730.
Weisberg, E., Parent, A., Yang, P.L., Sattler, M., Liu, Q., Wang, J., Meng, C., Buhrlage, S. J., Gray, N., Griffin, J.D., 2020. Repurposing of kinase inhibitors for treatment of COVID-19. Pharm. Res. (N. Y.) 37, 167.
Wickham, H., 2016. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag, New York.
Wyllie, A.L., Fournier, J., Casanovas-Massana, A., Campbell, M., Tokuyama, M., Vijayakumar, P., Warren, J.L., Geng, B., Muenker, M.C., Moore, A.J., et al., 2020. Saliva or nasopharyngeal swab specimens for detection of SARS-CoV-2. N. Engl. J. Med. 383, 1283–1286.
Xu, G., Qi, F., Li, H., Yang, Q., Wang, H., Wang, X., Liu, X., Zhao, J., Liao, X., Liu, Y., et al., 2020. The differential immune responses to COVID-19 in peripheral and lung revealed by single-cell RNA sequencing. Cell Discov. 6, 73.
Yan, Q., Li, P., Ye, X., Huang, X., Feng, B., Ji, T., Chen, Z., Li, F., Zhang, Y., Luo, K., et al., 2021. Longitudinal peripheral blood transcriptional analysis reveals molecular signatures of disease progression in COVID-19 patients. J. Immunol. 206, 2146–2159.
Yang, X., Yu, Y., Xu, J., Shu, H., Xia, J., Liu, H., Wu, Y., Zhang, L., Yu, Z., Fang, M., et al., 2020. Clinical course and outcomes of critically ill patients with SARS-CoV-2 pneumonia in Wuhan, China: a single-centered, retrospective, observational study. Lancet Respir. Med. 8, 475–481.
Yu, G., 2021. Enrichplot: visualization of functional enrichment result. Vol R package version 1.10.2.
Yu, G., Wang, L.G., Han, Y., He, Q.Y., 2012. clusterProfiler: an R package for comparing biological themes among gene clusters. OMICS A J. Integr. Biol. 16, 284–287.
Zhang, C., Wu, Z., Li, J.W., Zhao, H., Wang, G.Q., 2020. Cytokine release syndrome in severe COVID-19: interleukin-6 receptor antagonist tocilizumab may be the key to reduce mortality. Int. J. Antimicrob. Agents 55, 105954.
Zheng, H.Y., Xu, M., Yang, C.X., Tian, R.R., Zhang, M., Li, J.J., Wang, X.C., Ding, Z.L., Li, G.M., Li, X.L., et al., 2020a. Longitudinal transcriptome analyses show robust T cell immunity during recovery from COVID-19. Signal Transduct. Target Ther. 5, 294.
Zheng, H.Y., Zhang, M., Yang, C.X., Zhang, N., Wang, X.C., Yang, X.P., Dong, X.Q., Zheng, Y.T., 2020b. Elevated exhaustion levels and reduced functional diversity of T cells in peripheral blood may predict severe progression in COVID-19 patients. Cell. Mol. Immunol. 17, 541–543.
Zhong, S., 2019. ctrlGene: Assess the Stability of Candidate Housekeeping Genes. *Version 101, R Package*.
Zhou, Y., Zhang, J., Wang, D., Guan, W., Qin, J., Xu, X., Fang, J., Fu, B., Zheng, X., Zhao, H., et al., 2021. Profiling of the immune repertoire in COVID-19 patients with mild, severe, convalescent, or retesting-positive status. J. Autoimmun. 118, 102596.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

5.2. ARTICLE 2: SEX-BIASED EXPRESSION OF THE TLR7 GENE IN SEVERE COVID-PATIENTS: INSIGHTS FROM TRANSCRIPTOMICS AND EPIGENOMICS

The results from this chapter have already been published as A. Gómez-Carballa [a,b,c], J. Pardo-Seco [a,b,c], S. Pischedda [a,b,c], I. Rivero-Calle [a,b,d], G. Butler-Laporte [e,f], J.B. Richards [e,f], S. Viz-Lasheras [a,b,c], F. Martinón-Torres [a,b,d], A. Salas [a,b,c,*], GENDRES network. DOI: 10.1016/j.envres.2022.114288

Affiliations

a- Genetics, Vaccines and Infections Research Group (GENVIP), Instituto de Investigación Sanitaria (IDIS), Universidade de Santiago de Compostela, Santiago de Compostela, Spain

b- Centro de Investigación Biomédica en Red de Enfermedades Respiratorias (CIBER-ES), Madrid, Spain.

c- Unidade de Xenética, Instituto de Ciencias Forenses, Facultade de Medicina, Universidade de Santiago de Compostela, and GenPoB Research Group, Instituto de Investigación Sanitaria (IDIS), Hospital Clínico Universitario de Santiago (SERGAS), Galicia, Spain.

d- Translational Pediatrics and Infectious Diseases, Department of Pediatrics, Hospital Clínico Universitario de Santiago de Compostela, Santiago de Compostela, Spain.

e- Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montréal, Québec, Canada.

f- Lady Davis Institute, Jewish General Hospital, McGill University, Montréal, Québec, Canada.

Link: https://www.sciencedirect.com/science/article/pii/S0013935122016152?via%3Dihub

Abstract

There is abundant epidemiological data indicating that the incidence of severe cases of coronavirus disease (COVID-19) is significantly higher in males than females worldwide. Moreover, genetic variation at the X-chromosome linked *TLR7* gene has been associated with COVID-19 severity. It has been suggested that the sex-biased incidence of COVID-19 might be related to the fact that *TLR7* escapes X-chromosome inactivation during early embryogenesis

in females, thus encoding a doble dose of its gene product compared to males. We analysed *TLR7* expression in two acute phase cohorts of COVID-19 patients that used two different technological platforms, one of them in a multi-tissue context including saliva, nasal, and blood samples, and a third cohort that included different post-infection timepoints of long-COVID-19 patients. We additionally explored methylation patterns of *TLR7* using epigenomic data from an independent cohort of COVID-19 patients stratified by severity and sex. In line with genome-wide association studies, we provide supportive evidence indicating that *TLR7* has altered CpG methylation patterns, and it is consistently downregulated in males compared to females in the most severe cases of COVID-19.

DOI: 10.1016/j.envres.2022.114288

JOURNAL: Environmental Research (ISSN: 1096-0953)

IMPACT FACTOR: **8.3** (2022)

CATHEGORY: PUBLIC, ENVIRONMENTAL & OCCUPATIONAL HEALTH; ENVIRONMENTAL SCIENCES

RELATIVE POSSITION IN THE CATHEGORY: **16/207** in PUBLIC, ENVIRONMENTAL & OCCUPATIONAL HEALTH (**Q1 and D1**) and **32/275 in** ENVIRONMENTAL SCIENCES (**Q1**).

OTHER QUALITY EVIDENCE: Environmental Research (ISSN: 1096-0953) is an international peer-reviewed open access journal focused on publishing high-quality research in environmental science, biomedicine, and public health. Rank by Journal Citation Indicator situate this journal in the **15/334 (D1)** position for ENVIROMENTAL SCIENCES and **21/400 (D1)** in PUBLIC, ENVIRONMENTAL & OCCUPATIONAL HEALTH. The journal had **16,997** citations in 2022.

AUTHOR CONTRIBUTION: In the present study, the doctoral student took part in the study design, data analysis and the writing process.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Environmental Research 215 (2022) 114288

Contents lists available at ScienceDirect

Environmental Research

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

Sex-biased expression of the *TLR7* gene in severe COVID-19 patients: Insights from transcriptomics and epigenomics

A. Gómez-Carballa [a,b,c], J. Pardo-Seco [a,b,c], S. Pischedda [a,b,c], I. Rivero-Calle [a,b,d], G. Butler-Laporte [e,f], J.B. Richards [e,f], S. Viz-Lasheras [a,b,c], F. Martinón-Torres [a,b,d], A. Salas [a,b,c,*], GENDRES network

[a] *Genetics, Vaccines and Infections Research Group (GENVIP), Instituto de Investigación Sanitaria (IDIS), Universidade de Santiago de Compostela, Santiago de Compostela, Spain*
[b] *Centro de Investigación Biomédica en Red de Enfermedades Respiratorias (CIBER-ES), Madrid, Spain*
[c] *Unidade de Xenética, Instituto de Ciencias Forenses, Facultade de Medicina, Universidade de Santiago de Compostela, and GenPoB Research Group, Instituto de Investigación Sanitaria (IDIS), Hospital Clínico Universitario de Santiago (SERGAS), Galicia, Spain*
[d] *Translational Pediatrics and Infectious Diseases, Department of Pediatrics, Hospital Clínico Universitario de Santiago de Compostela, Santiago de Compostela, Spain*
[e] *Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montréal, Québec, Canada*
[f] *Lady Davis Institute, Jewish General Hospital, McGill University, Montréal, Québec, Canada*

ARTICLE INFO

Keywords:
COVID-19
Long-COVID-19
Host
TLR7
RNAseq
Transcriptomics
Epigenomics
n-Counter
DNA methylation

ABSTRACT

There is abundant epidemiological data indicating that the incidence of severe cases of coronavirus disease (COVID-19) is significantly higher in males than females worldwide. Moreover, genetic variation at the X-chromosome linked *TLR7* gene has been associated with COVID-19 severity. It has been suggested that the sex-biased incidence of COVID-19 might be related to the fact that *TLR7* escapes X-chromosome inactivation during early embryogenesis in females, thus encoding a doble dose of its gene product compared to males. We analyzed *TLR7* expression in two acute phase cohorts of COVID-19 patients that used two different technological platforms, one of them in a multi-tissue context including saliva, nasal, and blood samples, and a third cohort that included different post-infection timepoints of long-COVID-19 patients. We additionally explored methylation patterns of *TLR7* using epigenomic data from an independent cohort of COVID-19 patients stratified by severity and sex. In line with genome-wide association studies, we provide supportive evidence indicating that *TLR7* has altered CpG methylation patterns and it is consistently downregulated in males compared to females in the most severe cases of COVID-19.

1. Introduction

It has been reported that women have a stronger immune system than men, which generally leads to a more efficient response to infections and vaccines (Zuk, 2009; Cordoba-Aguilar and Munguia-Steyer, 2013; Fink and Klein, 2018); conversely, women also suffer significantly much more from autoimmune diseases, such as rheumatoid arthritis, multiple sclerosis, etc (Youness et al., 2021). A complex interaction of chromosome X-linked genetic factors coupled with sex-specific hormones and environmental factors, may be involved in this phenomenon. For instance, it has been claimed that altered (X-linked) Toll-Like Receptor (TLR) 7 gene (*TLR7*) expression is involved in several autoimmune disorders (Brown et al., 2022). In addition, deficiencies in *TLR7*, together with other inborn errors of type 1 interferons (IFNs), have been found in several studies on infectious diseases (Asano et al. 2021).

TLR7 has also received much attention in the context of the coronavirus disease 2019 (COVID-19). This gene is essential for host defense against the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) infection, as its protein plays a fundamental role in pathogen recognition and activation of innate immunity (Petes et al., 2017). *TLR7* encodes for an endosomal innate immune sensor capable to detect single-stranded RNA (ssRNA) of the virus, inducing the production of type 1 IFN and the release of other proinflammatory cytokines. It has

* Corresponding author. Unidade de Xenética, Instituto de Ciencias Forenses (INCIFOR), Facultade de Medicina, Universidade de Santiago de Compostela (USC), and GenPoB Research Group, Instituto de Investigación Sanitaria (IDIS), Hospital Clínico Universitario de Santiago (SERGAS), Galicia, Spain.
E-mail address: antonio.salas@usc.es (A. Salas).

https://doi.org/10.1016/j.envres.2022.114288
Received 20 July 2022; Received in revised form 30 August 2022; Accepted 4 September 2022
Available online 21 September 2022
0013-9351/© 2022 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

been suggested that insufficient type 1 IFN immunity in the respiratory tract during the initial days of the infection may be critical for the virus spread, potentially constituting a decisive factor leading to pulmonary and systemic inflammation (Zhang et al., 2022).

At the same time, it is also known that severe COVID-19 has a marked male sex bias, as also found in other viral infections (Guerra-Silveira and Abad-Franch, 2013; Wang et al., 2022). Thus, mortality and hospitalization rates are higher in male COVID-19 patients (Spiering and de Vries, 2021). It has been suggested that the *TLR7* gene might play a role in this differential sex incidence of COVID-19. Most of the evidence comes from genomic studies. The first studies pointing to this link were small-scale both family- and population-based genetic association studies carried out in selected severe COVID-19 patients (van der Made et al., 2020; Asano et al. 2021; Fallerini et al., 2021). The role of *TLR7* in COVID-19 pathogenesis has found further support in more recent large-scale Whole-Exome and Whole-Genome Sequencing studies (WES and WGS, respectively). Thus, for instance, the WES association study of >580K COVID-19 patients of the COVID-19 Host Genetics Initiative revealed several genes of marginal genome-wide significance. One of these genes was *TLR7*, which emerged as a good association candidate when contrasting COVID-19 severe patients *versus* COVID-19 negative or unknown donors (P-value $= 4.28 \times 10^{-8}$ in a burden association test) (Kosmicki et al., 2021). The most definitive evidence of this association has been reported in a recent study that combined WES and WGS from 624K COVID-19 patients including a much larger number of severe cases (Butler-Laporte et al., 2022); these authors showed that carrying a rare predicted loss of function or missense variant in the *TLR7* gene was associated with a 5.3-fold increase in odds of severe disease (P-value $= 5.41 \times 10^{-7}$); this study indicated that deleterious variation at this gene could potentially affect both, males and females. This finding found strong evidence with exome-wide significance in the discovery phase, and solid replication in a large independent cohort (Butler-Laporte et al., 2022).

It has been proposed that the rationale behind the role of *TLR7* in the differential sex incidence of COVID-19 may derive from the phenomenon of escape from X-chromosome inactivation, already described in the early 1960s (Lyon, 1962). To avoid double dosage of X-linked chromosomal genes in females compare to males, one of the female X-chromosomes is randomly silenced during early steps of embryonic development, thereby generating a mosaicism of activated X-chromosomes in females (X-chromosome inactivation or XCI). However, it seems that some X-linked genes, especially those in the pseudoautosomal regions (PARs) of pairing between the X and Y chromosomes, escape from this inactivation (escape from XCI). This phenomenon is not necessarily complete, it has inter-individual variability, and it also can differ between tissues (Spiering and de Vries, 2021).

However, little effort has been so far devoted to understanding the gene expression patterns of *TLR7* and their connection with different COVID-19 severe outcomes, as well as its presumed relationship with the reported sex-bias in this disease.

To shed light on this issue, we analyzed the sex differential transcription pattern of the *TLR7* gene in three recently published COVID-19 cohorts. Samples from two of these cohorts were selected to study *TLR7* expression during acute phase of infection: one dataset originated from a RNAseq study (Jackson et al., 2022), and the other one used the *n*-Counter (Nanostring) platform (Gómez-Carballa et al., 2022). The first study was carried out on blood samples extracted from COVID-19 patients (total $n = 69$), while the second one was carried out on blood, saliva, and nasopharyngeal samples (total $n = 52$); the two datasets overlap partially ($n = 23$ blood samples from the same timepoint), with this overlap serving as a technical validation. The third independent gene expression cohort, also stratified by sex and severity, was used to investigate the *TLR7* expression during COVID-19 convalescence and comes from a follow-up study also including long-COVID-19 patients (Ryan et al., 2022). In addition, we investigated the contribution of *TLR7* whole blood methylation patterns to the severity sex-bias observed in COVID-19 patients by exploring a recent epigenomic study ($n = 473$) (Barturen et al., 2022).

2. Methods

First, we used gene expression data from two previous studies carried out on acute phase COVID-19 patients stratified by sex and severity. We have selected these two cohorts due to the accessibility to full raw expression data as well as to availability of complete clinical data; therefore allowing a homogeneous and acurate classification of patients by severity (ICU non-ICU and WHO score). The first one used RNAseq data (Jackson et al., 2022) while the second one used *n*-Counter data (Gómez-Carballa et al., 2022); the two cohorts overlap in 23 patients. Samples were collected during acute phase of infection. Processing and normalization of raw data generated were carried out as in the original articles (Gómez-Carballa et al., 2022; Jackson et al., 2022). More details about both cohorts, samples sizes, and clinical data associated are provided in Table S1.

The classification of patients into ICU and non-ICU categories or the computation of the WHO score of severity refers to the time of sample collection. We used a Wilcoxon test to assess statistical significance between severity groups, and a Pearson test for the correlation indices (r) and P-values (after checking the data for normality). Log_2FC was calculated by subtracting the arithmetic mean of Log_2 counts of gene A in the patient sample, and the arithmetic mean of Log_2 counts of gene A in the reference sample.

Next, raw counts data and metadata from the COVID-19 post infection RNAseq study (12-, 16-, and 24-weeks post infection) were downloaded from the link provided by the authors in the original publication (Ryan et al., 2022); only convalescence samples from critical patients (severity classification during the acute phase) were used in the present study. This cohort was the only one available including follow-up blood whole transcriptome data from COVID-19 patients stratified by severity in the acute phase of the infection. Data normalization was carried out using Deseq2 R package (Love et al., 2014) with default parameters and after removing low expressed genes (genes with sum of counts lower than five in all the samples).

In addition, DNA methylation data were obtained from a recent epigenomic study (Barturen et al., 2022) including a total of 473 COVID-19 positive patients. Whole blood samples were recruited during the acute phase of the infection, and according to the clinical outcome, they were subdivided into two groups: severe ($n = 113$) and mild patients ($n = 360$). Severity classification was made by their authors considering a WHO scale (mild: 1–4 scale, severe: 5–8 scale). Their WHO score division was comparable to the one used in the mentioned gene expression studies (WHO 1–3 *vs.* 4–8); a full unification of the scores was not possible due to the unavailability of clinical data in the later study (in contrast to the data in the expression studies that were generated by our group). Raw data were downloaded, filtered, and normalized using *minfi* (Aryee et al., 2014) package. To identify differentially methylated positions (DMPs) between the phenotype groups, the *limma* package (Ritchie et al., 2015) was used assuming a linear model adjusted for age.

Sexual chromosomes representation and locations of immune-related genes was generated using *KaryoploteR* R package (Gel and Serra, 2017). We used the R package *Enrichmentbrowser* (Geistlinger et al., 2016) to collect genes involved in immune response processes from the Gene Ontology (GO; http://geneontology.org) database. Graphing and statistical analyses were carried out using R software v4.1.1 (www.r-project.org).

3. Results

3.1. Systemic TLR7 gene expression and methylation patterns

We analyzed the transcription patterns of *TLR7* in COVID-19 patients stratified by sex and severity. The RNAseq data from Jackson et al.

(2022) indicate that *TLR7* expresses almost equally in male and females in the non-ICU group of patients, but it is downexpressed in males with respect to females in the ICU group (although the difference is not statistically significant; Fig. 1A). This pattern is also suggested when stratified patients by WHO severity score (Fig. 1A). Notably, there is a very weak positive correlation (not statistically significant) between *TLR7* expression and the WHO score in samples within the score range 1–3 (r = 0.11; *P*-value = 0.59) but this correlation is strongly negative and highly statistically significant for the more severe patients with WHO score values ≥ 4 (r = −0.47; *P*-value = 0.0071); Fig. 1B.

The data generated in the *n*-Counter study by Gómez-Carballa et al. (2022) showed that, for the overlapping samples with the RNAseq dataset, there is a significant correlation of gene expression values for the *TLR7* gene (r = 0.82; *P*-value = 1.9×10^{-6}; Fig. 1C). In agreement with the RNAseq data above, the *n*-Counter data also indicate that *TLR7* expresses less in males than females in the most severe patients. Although this difference is more subtle and non-statistically significant when considering the WHO score, it is clearly statistically significant when using the ICU *vs*. non-ICU criterion (*P*-value = 0.004); Fig. 1A.

Finally, in a comparison of ICU *vs*. non-ICU males, *TLR7* is downexpressed in severe male patients in both the RNAseq (*P*-value = 0.0054) and the *n*-Counter (*P*-value = 0.0045) datasets (Fig. S1A).

DNA methylation analysis of *TLR7* on blood samples revealed the presence of nine CpGs within the gene, three of them located within the gene body, while the remaining six were annotated to its promoter region (TSS200, TSS1550, 5′UTR). The nine positions did not show the same trend of methylation for the whole gene. When we interrogated the methylation status of these nine positions in females and males, separately, we observed not significant DMPs in females when comparing the severe group (WHO ≥ 5) against the mild group (WHO < 5); however, when the same analysis was performed in males, a statistically significant difference was observed in four out of the nine positions (Fig. S2). Three of these DMPs (cg08029608: *P*-value = 0.005; cg09542796: *P*-value = 0.0001; cg22859180: *P*-value = 0.003) showed a hypermethylation pattern in severe males, while one DMP (cg24735671: *P*-value = 0.045) presented an opposite methylation status, with higher methylation levels in mild patients.

3.2. TLR7 gene expression in nasal and saliva samples

TLR7 expression has also been analyzed in saliva and nasal samples from COVID-19 patients (Gómez-Carballa et al., 2022) stratified by severity and sex.

It is again remarkable that *TLR7* is always downregulated in males with respect to females in both nasal and saliva samples and independently of the severity criteria used (Fig. 1D).

Thus, the *TLR7* expression in the comparison males *vs*. females is statistically significant in saliva samples in both non-ICU (*P*-value = 6.7×10^{-6}) and ICU patients (*P*-value = 0.022). This difference is also notable and in the same direction when using the WHO score, although it is only statistically significant in non-ICU patients (*P*-value = 0.036, Fig. 1D).

Females and males do not differ significantly in *TLR7* expression from nasal samples, but the data always show a decreasing expression pattern in males when compared to females, and this difference is more marked when considering the WHO score (although this comparison is particularly affected by a low sample size).

3.3. TLR7 expression, admission days, days of symptoms and convalescence

We also analyzed the expression of *TLR7* in the context of

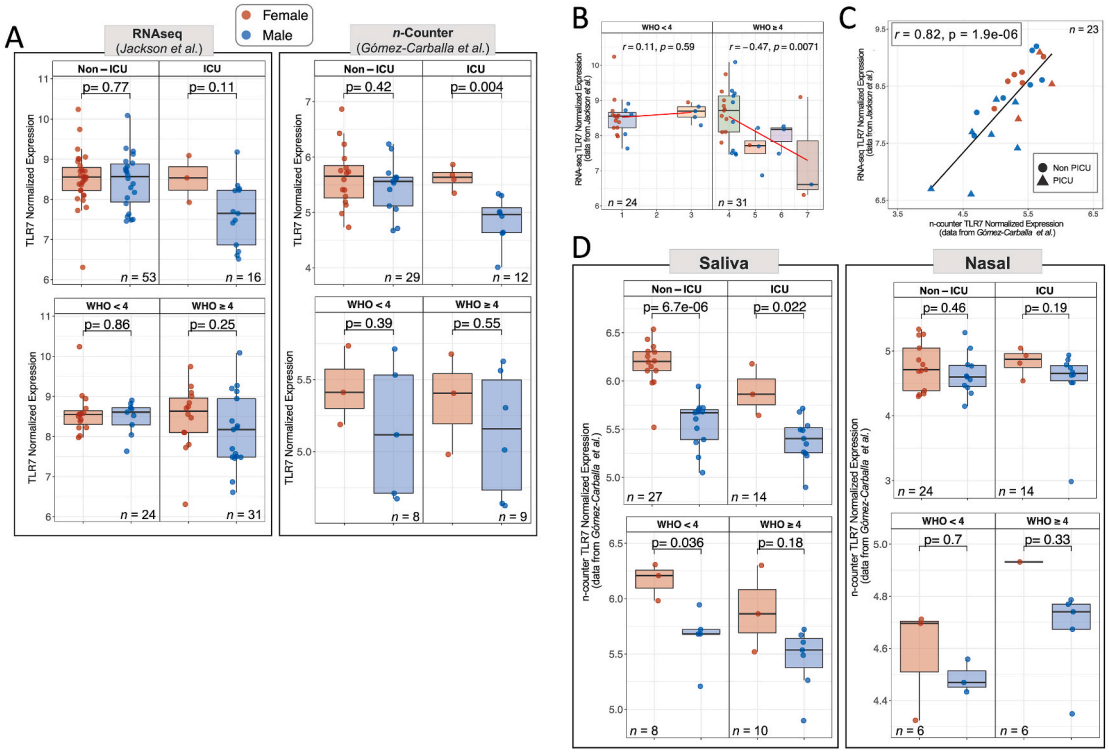


Fig. 1. (A) *TLR7* gene expression in blood samples from COVID-19 patients using two datasets, RNAseq from (Jackson et al., 2022) and *n*-Counter from (Gómez-Carballa et al., 2022), and stratifying by sex and severity: (*i*) ICU *vs*. non-ICU ($n = 69$ from (Jackson et al., 2022), and $n = 41$ from (Gómez-Carballa et al., 2022)) and (*ii*) WHO score ($n = 55$ from (Jackson et al., 2022), and $n = 17$ (Gómez-Carballa et al., 2022)). (B) Correlation between *TLR7* expression ($n = 55$) and the WHO score using RNAseq data (Jackson et al., 2022); *P*-values are also provided. (C) Correlation between *TLR7* expression in the RNAseq (Jackson et al., 2022) and the *n*-Counter (Gómez-Carballa et al., 2022) datasets ($n = 23$). (D) *TLR7* gene expression in nasal and saliva samples in the *n*-Counter dataset (Gómez-Carballa et al., 2022) stratified by sex and severity: ICU *vs*. non-ICU ($n = 41$ in saliva and $n = 38$ in nasal samples) and WHO score ($n = 18$ in saliva and $n = 12$ in nasal samples).

hospitalization days and days of symptoms to sample collection. It is most remarkable that expression of *TLR7* is negatively correlated with days from the onset of symptoms (whether hospitalized or not) to sample collection but only in ICU patients (r = −0.43; Fig. 2A). This negative correlation is even more clear when considering only those patients for which there is information on the WHO score (r = −0.66; Fig. S1B). A similar correlation is observed when using the WHO score as severity parameter, although yielding a more discrete value (r = −0.14; Fig. 2B).

TLR7 expression is also negatively correlated with admission days until sample collection (Fig. 2C); despite being a moderate correlation value, it is in the limit of the nominal statistical significance (r = −0.32; *P*-value = 0.059).

Additionally, we investigated how *TLR7* expression changes at different convalescence times after infection in patients classified as critical in the acute phase of the disease. Even though results are not statistically significant (probably owing to the small sample size), we observed that *TLR7* is again downexpressed in the acute phase of infection of the most critical males, and this *TLR7* reduction is maintained over time in convalescence samples, even 24 weeks post infection (Fig. 2 D).

4. Discussion

It has been reported that several genes involved in the regulation of the innate and adaptive immune response are strategically placed on the X-chromosome (Fig. 3A), including pathogen-related receptors (PRRs) such as the *TLR7*, *ACE2* and *TMPRSS2*. These and other genes have been found to be associated in COVID-19 in family-based, genome-wide and NGS population-based association studies. We examined expression patterns of *TLR7* because of its singularity of being located at a XCI escape region and, therefore, having (theoretically) a predominantly 'biallelic' behavior in females but 'monoallelic' in males (Fig. 3A). Although genetic variation at this gene has been noted several times in genomic COVID-19 association studies, its gene expression patterns in whole blood, saliva and nasal samples has not been analyzed so far.

Our data indicate that *TLR7* is downregulated in blood from males with respect to females in the most severe cases of COVID-19 when stratifying the cohort in those attending ICU *vs.* non-ICU, and when using WHO scores for severity. Consistently, a statistically significant downregulation of *TLR7* occurs in males with a more severe disease (ICU) when compared to non-ICU males (who show similar *TLR7* expression levels as females). Moreover, it seems that *TLR7* expresses in a sex-biased manner in saliva and nasal samples. Note that, while in blood from ICU patients, *TLR7* expresses less in males than in females, in saliva *TLR7* follows this pattern independently of the severity classification used. A trend towards increased *TLR7* expression in females than in males is also observed when examaning Log_2FC values and median *TLR7* expression differences in females *vs.* males, in both, ICU and WHO ≥ 4 and in all tissues analyzed (Fig. 3B).

Therefore, the *TLR7* expression in severe females was always higher than in severe males, independently of the tissue and severity classification used, with blood being the tissue showing the largest expression differences. Even though not very large expression changes, our results indicate that *TLR7* may play a role in the sex-bias observed in the more severe outcomes of COVID-19.

Two of the three gene expression datasets used in the present study overlap in a few blood samples, and this fact offered an opportunity to assess the power of both RNAseq and *n*-Counter transcriptomic techniques to capture the real gene expression of the *TLR7* gene. The results of these two analyses correlate significantly, therefore providing a technical validation of the expression data.

In addition, we examined methylation patterns of *TLR7* in COVID-19 patients. We found four (statistically significant) DMPs in *TLR7* when comparing severe males (WHO ≥ 5) *vs.* males with mild disease (WHO < 5) but not in females. Interestingly, while three of them were

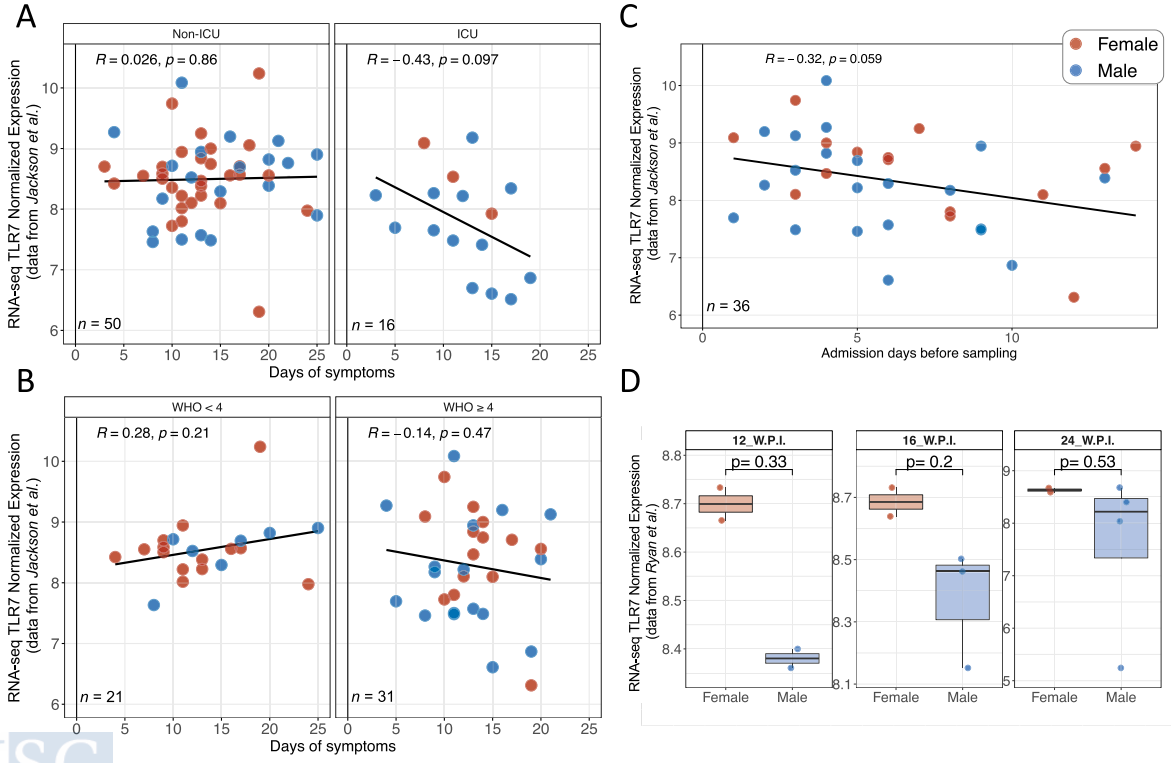


Fig. 2. Correlation between *TLR7* gene expression using RNAseq data from (Jackson et al., 2022) and days of symptoms (asymptomatic patients were not included) since disease onset to sampling using (A) ICU (n = 66) and (B) WHO classification (n = 52); and (C) admission days since hospitalization to sampling (only hospitalized samples were included; n = 36). (D) *TLR7* expression in blood convalescence samples from critical COVID patients collected 12 w.p.i (n = 4), 16 w.p.i (n = 5) and 24 w.p.i (n = 6) (w.p.i = weeks post infection) from the study by (Ryan et al., 2022).

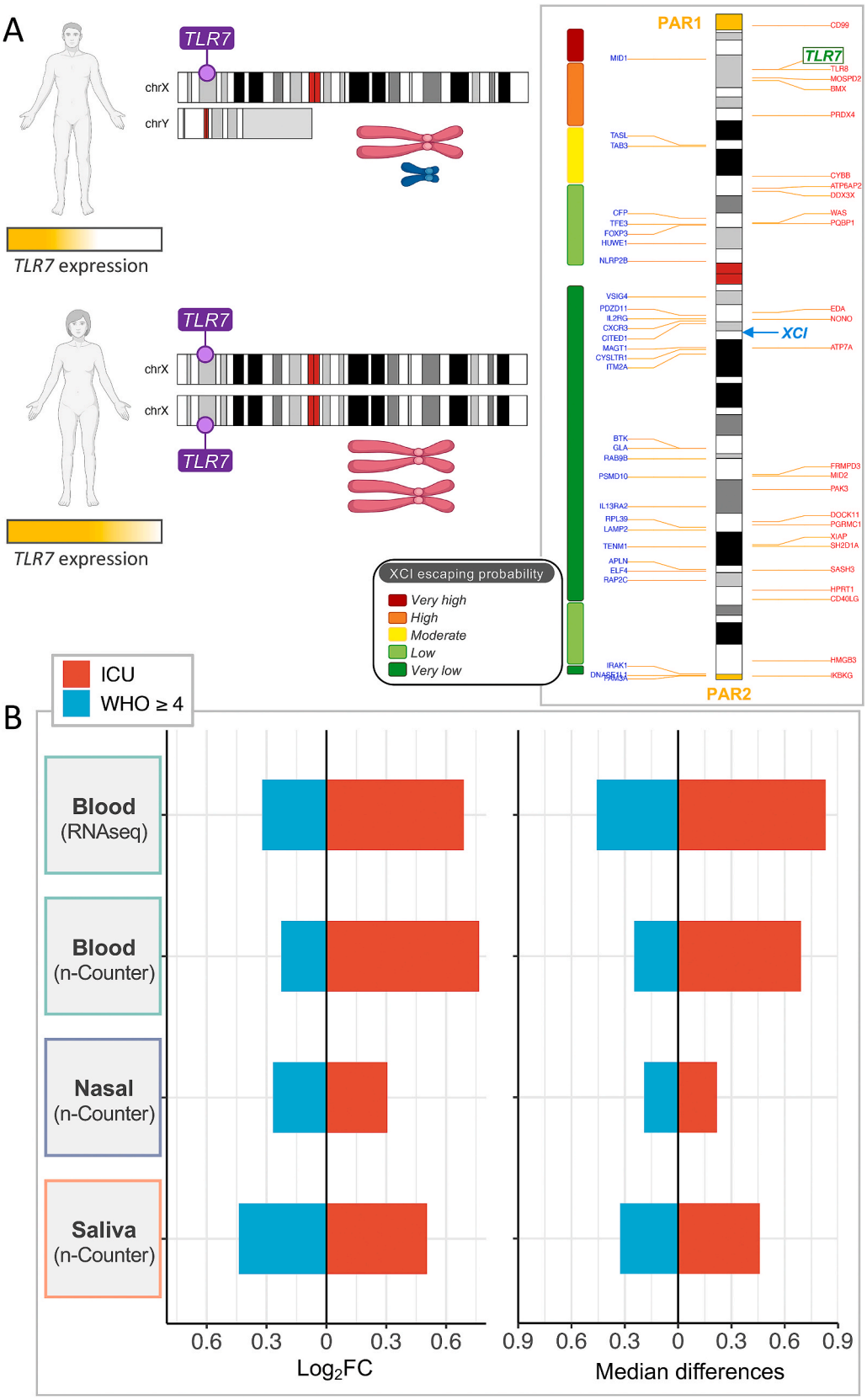


Fig. 3. (A) Schematic representation of sex chromosomes (left); X-chromosome location of *TLR7* and main immune-related genes is also displayed (gene names in red indicate location at positive strand, and in blue at negative strand); XCI escaping probability regions were obtained from (Schurz et al., 2019) (right). Male and female figures as well as illustrations of X- Y-chromosomes were downloaded from Biorender (https://biorender.com/). (B) *TLR7* expression changes in severe females with respect to severe males (for both criteria, ICU and WHO ≥ 4) in all tissues analyzed from RNA-seq (Jackson et al., 2022) and *n*-Counter (Gómez-Carballa et al., 2022) data expressed as Log_2FC and median differences. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

hypermethylated in severe males, the other one showed higher methylation in male patients with a mild phenotype. This might suggest that the interaction between methylation and demethylation of *TLR7* gene could be contributing to COVID-19 severity condition in males. The coexistence of both hypo- and hypermethylated sites in the same gene region has not been fully understood (He et al., 2019). The 'dogma' defining the classical correlation between DNA methylation and gene expression asserts that when methylation occurs at the promoter region, the expression of the gene is inhibited or downregulated (Baylin, 2005). Actually, whole-genome studies have revealed that the effect of DNA methylation is not restricted to the promoter regions, but epigenetic changes occurring in low CpG density regions (such as gene bodies), could also impact on the regulation of gene expression (Jones, 2012). With the data available, it is not possible to assess to what extent, the different patterns of methylation observed in the *TLR7* gene (in the promoter region [three DMPs] and in the gene body [one DMP]), could contribute to a global downregulation of the gene in blood (as showed by the transcriptomic data).

The data examined in the present study are consistent in the different cohorts analyzed, using different technical procedures for gene expression measurement (RNAseq and *n*-Counter from Nanostring) and different '-omic' approaches (transcriptomics and epigenomics). The patterns observed are in line to expectations according to genomic association studies. However, many questions remain unanswered regarding the role of *TLR7* in COVID-19. For instance, as inferred from large-scale genome studies and the gene expression results of the present study, *TLR7* alone cannot fully explain severity in COVID-19 patients. For instance, this gene seems to be upregulated in a minor proportion of females who also have a severe outcome. This observation would be in good agreement with the fact that variation at the deleterious variants in *TLR7* is related to severity in both, males and females (Butler-Laporte et al., 2022).

There are three main limitations in the present study. First, the relatively small sample size of the studies, which might explain the lack of statistical significance of a few pairwise comparisons. Second, *TLR7* gene expression was measured in whole blood, nasal and saliva samples showing certain tissue variability with respect to sex differences and severity; therefore, its expression in other tissues that are directly involved in the immune response to SARS-CoV-2 infection (i.e. lower respiratory tract tissues) remains to be analyzed. Third, our study does not allow to investigate the level of *TLR7* expression in different cell populations in e.g blood. Finally, other external or internal factors might be modulating *TLR7* expression to some extent, including the different treatments given to patients with a more severe course of the disease (especially those from ICU) or sexual hormone levels, which were recently associated to an increased type 1 IFN response in females (Webb et al., 2018). Instead, the main strength of the study comes from the consistent trend seen in these comparisons, which basically indicates a more limited gene expression of *TLR7* in severely affected males (even in the convalescence phase) and a negative correlation with admission days and days of symptoms.

To the best of our knowledge, this is the first study investigating expression levels of *TLR7* in different COVID-19 scenarios and tissues. The findings suggest that its gene expression could (at least in part) explain the sex-biased incidence in severe patients. Further studies are necessary to investigate more thoroughly the relationship between *TLR7* expression, sex, and COVID-19 severity.

Funding sources

This study received support from Instituto de Salud Carlos III (ISCIII): GePEM (PI16/01478/Cofinanciado FEDER; A.S.), DIAVIR (DTS19/00049/Cofinanciado FEDER, A.S.), Resvi-Omics (PI19/01039/Cofinanciado FEDER, A.S.), Agencia Gallega de Innovación (GAIN): Grupos con Potential de Crecimiento (IN607B 2020/08, A.S.); Agencia Gallega para la Gestión del Conocimiento en Salud (ACIS): BI-BACVIR (PRIS-3, A.S.), and CovidPhy (SA 304C, A.S.); ReSVinext (PI16/01569/Cofinanciado FEDER, F.M.T.), Enterogen (PI19/01090/Cofinanciado FEDER, F.M.T.) and consorcio Centro de Investigación Biomédica en Red de Enfermedades Respiratorias (CB21/06/00103; F.M.T.); GEN-COVID (IN845D 2020/23, F.M-T.) and Grupos de Referencia Competitiva (IIN607A2021/05, F.M-T). The funders were not involved in the study design, collection, analysis, interpretation of data, the writing of this article or the decision to submit it for publication.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data from four other studies were used. We did not generated new data.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.envres.2022.114288.

References

Aryee, M.J., Jaffe, A.E., Corrada-Bravo, H., Ladd-Acosta, C., Feinberg, A.P., Hansen, K.D., Irizarry, R.A., 2014. Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. Bioinformatics 30, 1363–1369.

Asano, T., Boisson, B., Onodi, F., Matuozzo, D., Moncada-Velez, M., Maglorius, Renkilaraj, MRL, Zhang, Meertens, P., Bolze, L., Materna, A., et al., 2021. X-linked recessive TLR7 deficiency in ~1% of men under 60 years old with life-threatening COVID-19. Sci Immunol 6.

Barturen, G., Carnero-Montoro, E., Martinez-Bueno, M., Rojo-Rello, S., Sobrino, B., Porras-Perales, O., Alcantara-Dominguez, C., Bernardo, D., Alarcon-Riquelme, M.E., 2022. Whole blood DNA methylation analysis reveals respiratory environmental traits involved in COVID-19 severity following SARS-CoV-2 infection. Nat. Commun. 13, 4597.

Baylin, S.B., 2005. DNA methylation and gene silencing in cancer. Nat. Clin. Pract. Oncol. 2 (Suppl. 1), S4–S11.

Brown, G.J., Canete, P.F., Wang, H., Medhavy, A., Bones, J., Roco, J.A., He, Y., Qin, Y., Cappello, J., Ellyard, J.I., et al., 2022. TLR7 gain-of-function genetic variation causes human lupus. Nature 605, 349–356.

Butler-Laporte, G., Povysil, G., Kosmicki, J.A., Cirulli, E.T., Drivas, T., Furini, S., Saad, C., Schmidt, A., et al., 2022. Exome-wide Association Study to Identify Rare Variants Influencing COVID-19 Outcomes: Results from the Host Genetics Initiative. *PLoS Genet* **In press**.

Cordoba-Aguilar, A., Munguia-Steyer, R., 2013. The sicker sex: understanding male biases in parasitic infection, resource allocation and fitness. PLoS One 8, e76246.

Fallerini, C., Daga, S., Mantovani, S., Benetti, E., Picchiotti, N., Francisci, D., Paciosi, F., Schiaroli, E., Baldassarri, M., Fava, F., et al., 2021. Association of Toll-like receptor 7 variants with life-threatening COVID-19 disease in males: findings from a nested case-control study. Elife 10, e67569.

Fink, A.L., Klein, S.L., 2018. The evolution of greater humoral immunity in females than males: implications for vaccine efficacy. Curr Opin Physiol 6, 16–20.

Geistlinger, L., Csaba, G., Zimmer, R., 2016. Bioconductor's EnrichmentBrowser: seamless navigation through combined results of set- & network-based enrichment analysis. BMC Bioinf. 17, 45.

Gel, B., Serra, E., 2017. karyoploteR: an R/Bioconductor package to plot customizable genomes displaying arbitrary data. Bioinformatics 33, 3088–3090.

Gómez-Carballa, A., Rivero-Calle, I., Pardo-Seco, J., Gómez-Rial, J., Rivero-Velasco, C., Rodriguez-Nunez, N., Barbeito-Castineiras, G., Pérez-Freixo, H., Cebey-López, M., Barral-Arca, R., et al., 2022. A multi-tissue study of immune gene expression profiling highlights the key role of the nasal epithelium in COVID-19 severity. Environ. Res. 210, 112890.

Guerra-Silveira, F., Abad-Franch, F., 2013. Sex bias in infectious disease epidemiology: patterns and processes. PLoS One 8, e62390.

He, L., Khanal, P., Morse, C.I., Williams, A., Thomis, M., 2019. Differentially methylated gene patterns between age-matched sarcopenic and non-sarcopenic women. J Cachexia Sarcopenia Muscle 10, 1295–1306.

Jackson, H., Rivero-Calle, I., Broderick, C., Habgood-Coote, D., D'Souza, G., Nichols, S., Vito, O., Gómez-Rial, J., Rivero-Velasco, C., Rodríguez-Núñez, N., et al., 2022. Characterisation of the blood RNA host response underpinning severity in COVID-19 patients. Sci. Rep. 12, 12216.

Jones, P.A., 2012. Functions of DNA methylation: islands, start sites, gene bodies and beyond. Nat. Rev. Genet. 13, 484–492.

Kosmicki, J.A., Horowitz, J.E., Banerjee, N., Lanche, R., Marcketta, A., Maxwell, E., Bai, X., Sun, D., Backman, J.D., Sharma, D., et al., 2021. Pan-ancestry exome-wide association analyses of COVID-19 outcomes in 586,157 individuals. Am. J. Hum. Genet. 108, 1350–1355.

Love, M.I., Huber, W., Anders, S., 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 15, 550.

Lyon, M.F., 1962. Sex chromatin and gene action in the mammalian X-chromosome. Am. J. Hum. Genet. 14, 135–148.

Petes, C., Odoardi, N., Gee, K., 2017. The toll for trafficking: Toll-like receptor 7 delivery to the endosome. Front. Immunol. 8, 1075.

Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W., Smyth, G.K., 2015. *Limma* powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 43, e47.

Ryan, F.J., Hope, C.M., Masavuli, M.G., Lynn, M.A., Mekonnen, Z.A., Yeow, A.E.L., Garcia-Valtanen, P., Al-Delfi, Z., Gummow, J., Ferguson, C., et al., 2022. Long-term perturbation of the peripheral immune system months after SARS-CoV-2 infection. BMC Med. 20, 26.

Schurz, H., Salie, M., Tromp, G., Hoal, E.G., Kinnear, C.J., Moller, M., 2019. The X chromosome and sex-specific effects in infectious disease susceptibility. Hum. Genom. 13, 2.

Spiering, A.E., de Vries, T.J., 2021. Why females do better: the X chromosomal *TLR7* gene-dose effect in COVID-19. Front. Immunol. 12, 756262.

van der Made, C.I., Simons, A., Schuurs-Hoeijmakers, J., van den Heuvel, G., Mantere, T., Kersten, S., van Deuren, R.C., Steehouwer, M., van Reijmersdal, S.V., Jaeger, M., et al., 2020. Presence of genetic variants among young men with severe COVID-19. JAMA 324, 663–673.

Wang, C., Lashua, L.P., Carter, C.E., Johnson, S.K., Wang, M., Ross, T.M., Ghedin, E., Zhang, B., Forst, C.V., 2022. Sex disparities in influenza: a multiscale network analysis. iScience 25, 104192.

Webb, K., Peckham, H., Radziszewska, A., Menon, M., Oliveri, P., Simpson, F., Deakin, C. T., Lee, S., Ciurtin, C., Butler, G., et al., 2018. Sex and pubertal differences in the Type 1 Interferon pathway associate with both X chromosome number and serum sex hormone concentration. Front. Immunol. 9, 3167.

Youness, A., Miquel, C.H., Guery, J.C., 2021. Escape from X chromosome inactivation and the female predominance in autoimmune diseases. Int. J. Mol. Sci. 22.

Zhang, Q., Bastard, P., Effort, C.H.G., Cobat, A., Casanova, J.L., 2022. Human genetic and immunological determinants of critical COVID-19 pneumonia. Nature 603, 587–598.

Zuk, M., 2009. The sicker sex. PLoS Pathog. 5, e1000267.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

5.3. ARTICLE 3: WHOLE EXOME SEQUENCING IDENTIFIES NEW SUSCEPTIBILITY CANDIDATES UNDERLYING COMMUNITY-ACQUIRED PNEUMONIA

The results from this chapter have already been published as Jacobo Pardo-Seco [a,b,c], Sandra Viz-Lasheras [a,b,c], Xabier Bello [a,b,c], Alberto Gómez-Carballa [a,b,c],Alba Camino-Mera [a,b,c], Sara Pischedda [a,b,c], María José Currás-Tuala [a,b,c], Irene Rivero-Calle [b,c,d], Ana Dacosta-Urbieta [b,c,d], Fernando Caamaño-Viña [b,c,d], Carmen Rodríguez-Tenreiro Sánchez [b,c,d], Isabel Cifuentes [e], Cristina Méndez [e], Chiea Chuen Khor [f], Federico Martinón-Torres [b,c,d], Antonio Salas [a,b,c,*]on behalf of the CAPPRIC study group[1]. DOI: https://doi.org/10.1016/j.gendis.2023.101170

Affiliations

a- Unidade de Xenética, Instituto de Ciencias Forenses,Facultade de Medicina, Universidade de Santiago de Compostela, and GenPoB Research Group, Instituto de Investigación Sanitaria (IDIS), Hospital Clínico Universitario de Santiago de Compostela (SERGAS), Galicia 15706, Spain.

b -Genetics, Vaccines and Infections Research Group (GENVIP), Instituto de Investigación Sanitaria de Santiago, Universidade de Santiago de Compostela, Santiago de Compostela, Galicia 15706, Spain.

c- Centro de Investigación Biomédica en Red de Enfermedades Respiratorias (CIBERES), Madrid 28029, Spain.

d- Translational Pediatrics and Infectious Diseases, Department of Pediatrics, Hospital Clínico Universitario de Santiago de Compostela (SERGAS), Galicia 15706, Spain.

Link: https://www.sciencedirect.com/science/article/pii/S2352304223004531?via%3Dihub

Abstract

Community-acquired pneumonia (CAP) is one of the main causes of hospitalization and mortality involving infectious diseases. With the aim of investigating the genetic susceptibility causes underlying CAP, we carried out a multi-centric and prospective study involving the recruitment of 300 adults with pneumonia and 438 healthy controls. Whole exome-wide sequencing data allowed to pinpoint a few SNP candidates (with rs1056322 being the most

statistically significant in a recessive genetic model; *P*-value = 4.71×10^{-6}, OR = 2.64) falling in a linkage disequilibrium region covering the *FAM118A* gene. A SKAT-O gene-based association test highlighted five additional gene candidates, namely *COL13A1*, *C20orf96*, *HEXIM1*, and *NPAS4*, and *STRC*. By meta-analyzing gene expression data from public repositories, we found several disease scenarios where these top genes are significantly dysregulated in non-COVID-19 pneumonia and in severe COVID-19 pneumonia patients. Remarkably, in COVID-19 patients, while the expression of *FAM1188A* and *COL13A1* decay with the time from symptoms onset, the expression of *HEXIM1* shows the opposite trend. Predictive performances were assessed using the area under the receiver operating characteristic curve (AUC-ROC), showing the predictive power of different combinations of the best candidate genes to predict the disease outcome. We found strong supportive genetic evidence underlying pneumonia outcome beyond its etiology.

DOI: 10.1016/j.gendis.2023.101170

JOURNAL: Genes and Diseases (ISSN: 1096-0953)

IMPACT FACTOR: **6.9** (2023)

CATHEGORY: BIOCHEMISTRY & MOLECULAR BIOLOGY; GENETICS & HEREDITY

RELATIVE POSSITION IN THE CATHEGORY: **36/313** in BIOCHEMISTRY & MOLECULAR BIOLOGY (**Q1**) and **16/191** in GENETICS & HEREDITY (**Q1 and D1**).

OTHER QUALITY EVIDENCE: Genes and Diseases (ISSN: 1096-0953) is an international peer-reviewed open access journal focused on publishing cutting-edge research in molecular and translational medicine. The journal primarily focuses on publishing investigations on the molecular bases and experimental therapeutics of human diseases.

AUTHOR CONTRIBUTION: In the present study, the doctoral student took part in the study design, data analysis and the writing process.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Genes & Diseases (2024) 11, 101170

Available online at www.sciencedirect.com

ScienceDirect

journal homepage: www.keaipublishing.com/en/journals/genes-diseases



RAPID COMMUNICATION

Whole exome sequencing identifies new susceptibility candidates underlying community-acquired pneumonia

Pneumonia is an inflammatory condition of the lung with symptoms that include productive dry cough, fever, chest pain, and difficulty breathing, and it is usually caused by viruses and bacteria, but also other microorganisms (such as fungi and parasites). Community-acquired pneumonia (CAP) is a major cause of infectious diseases, hospitalization, and mortality, especially in the elderly population.[1] We have undertaken a case—control study on CAP patients by way of sequencing the complete exome of 300 patients and 438 healthy controls (Table S1 for demographic and clinical characteristics of patients). This study is by far the largest exome sequencing project to date aimed at exploring the potential genetic causes behind CAP.

We have sequenced the whole-exome of 300 saliva samples from CAP patients older than 18 years recruited under the umbrella of the CAPPRIC study group.[2] Sequence analysis and statistical procedures were carried out as indicated in Supplementary Text (sequencing details are described in the previous study[3]). For the best candidate genes, we explored public resources to further investigate gene expression in COVID-19 and non-COVID-19 pneumonia patients (details in Supplementary Text).

A total of 517,502 biallelic single nucleotide polymorphisms (SNPs) were obtained after filtering indels, no-biallelic positions, monomorphic SNPs, and variants with genotyping rate >99.9%. We initially conducted a population stratification analysis by merging our dataset with 1000 Genome Project (1000 GP) population data, resulting in an overlapping set of 301,412 SNPs. Analysis of the ancestral background of samples showed the CAP exomes falling within the Western European variation, using both a multidimensional scaling analysis and a genome-ancestry analysis (Fig. 1A; Supplementary Text).

To carry out the single-point association test, 447,239 and 265 SNPs were removed by minor allele frequency (MAF) and Hardy-Weinberg criteria, respectively. Totally 69,998 markers survived these initial filters. Association test results revealed the most statistically significant SNPs when comparing CAP patients with controls: rs2239813, rs2753492, rs10039601, and rs700082 (Fig. S1). For the genotyping association test (additive, dominant, and recessive models), we evaluated 47,635 SNPs; being the most statistically significant one rs1056322 (Fig. S2) when considering a recessive model, which falls within the *FAM118A* gene. By further exploring the genome region closely related to this SNP, we also found two additional SNPs showing highly significant values under the same recessive model (rs6007010 and rs6007594); these three SNPs are in a region showing high linkage disequilibrium (rs1056322 *vs.* rs6007010: $r^2 = 0.89$; rs1056322 *vs.* rs6007594: $r^2 = 0.76$) (Fig. S3, 4). We ran the SKAT-O collapsing association tests using whole gene variability and considering only the rare variants (MAF < 0.05). Five genes passed the false discovery rate test for multiple corrections: *C20orf96*, *COL13A1*, *HEXIM1*, *NPAS4*, and *STRC* (Table S2 and Fig. 1B).

We additionally examined the expression of genes related to the most significant SNPs and genes. We used whole blood gene expression datasets available in public repositories — three of these analyzing pneumonia in adults, and another one analyzing pneumonia in children. *FAM118A* gene is significantly down-regulated in adult pneumonia patients, when compared with healthy controls (Fig. 1C and Table S3), and also in the pediatric cohort when compared with controls (Fig. 1D). Next, we examined the gene expression patterns for the five most significant genes emerging from the collapsing association tests. Except for the *HEXIM1* gene, we were able to capture the other four genes differentially expressed in some of the patient cohorts when compared with their respective controls (Fig. 1C

Peer review under responsibility of Chongqing Medical University.

https://doi.org/10.1016/j.gendis.2023.101170

2352-3042/© 2023 The Authors. Publishing services by Elsevier B.V. on behalf of KeAi Communications Co., Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

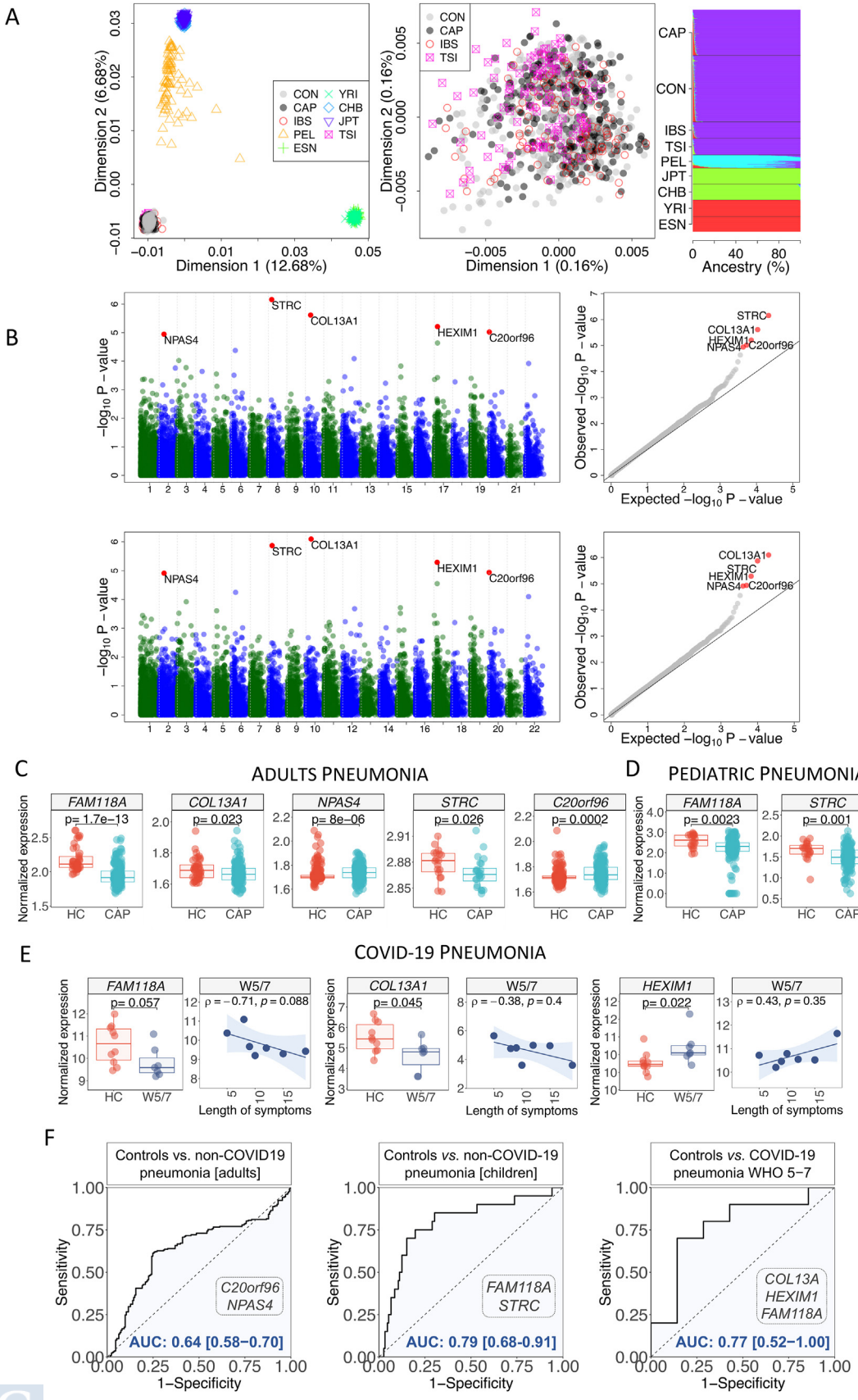


Figure 1 Ancestral genetic analysis of the CAPPRIC cohort, allele/gene-based association tests, and differential expression of candidate genes with best ROC-AUC predictions. **(A)** Analysis of the ancestral background of the studied cohort using reference continental population dataset from the 1000 GP. Left: Multidimensional scaling analysis for the patient's cohort and continental

and Table S3). The most significant findings are: i) the *NPAS4* is highly significant in the GSE65682 cohort and even more significant when meta-analyzed with the other two adult cohorts; ii) *C20orf96* is significantly up-regulated when meta-analyzed in the three adult cohorts; iii) *COL13A1* appears up-regulated only in the largest cohort GSE65682; and iv) the *STRC* gene appears significantly up-regulated in the adult cohort GSE40012 and in the pediatric cohort (Fig. 1C, D and Table S3). Analysis of the AUC (Area Under the Curve) -ROC (Receiver Operating Charasteristic) curves indicates that the combination of these candidate genes that best predicts non-COVID-19 pneumonia in adults is *C20orf96* and *NPAS4*, with an AUC-ROC value of 0.64 (0.58—0.70), while in the pediatric cohort, the combination is *FAM118* and *STRC*, with an AUC-ROC value of 0.79 (0.68—0.91) (Fig. 1F).

We also analyzed whole blood transcriptome data from pneumonia in severe COVID-19 patients (with WHO score values between 5 and 7 (W5/7)). Two candidate genes were found to be expressed differentially in cases and controls (Fig. 1E and Table S3). The *FAM118A* gene was found to be down-regulated in severe COVID-19 cases when compared with healthy controls. Most interestingly, the expression of this gene decays with the time from symptom onset (Spearman's $\rho = -0.71$). The *HEXIM1* gene was up-regulated in the most severe patients when compared with healthy controls. Most remarkable is that the expression of *HEXIM1* appears to increase with the time from symptom onset (Spearman's $\rho = 0.43$; Fig. 1E). The *COL13A1* gene is however down-regulated in the most severe patients; the expression of this gene decays over time from symptom onset (Spearman's $\rho = -0.38$; Fig. 1E). Analysis of the AUC-ROC curves indicates that the combination of genes (among the best candidates) that best predicts severe COVID-19 pneumonia is *COL13A1*, *FAM118A*, and *HEXIM1*, with an AUC-ROC value of 0.77 (0.52—1.00) (Fig. 1F).

Statistical association tests on allele and genotype frequencies, and the analysis of collapsing variants per gene in cases and controls, have allowed to shed light on new candidate susceptibility variation to pneumonia. Single-allele association tests highlight the variant rs1056322, located in the 3'UTR at "the family with sequence similarity 118 member A" (*FAM118A*) gene as the most significant one (under a recessive model), with two other closely related SNPs that are in high linkage disequilibrium and show suggestive statistical significance, namely rs6007010 (synonymous SNP falling in the neighboring gene *SMC1B*) and rs6007594 (non-synonymous variant located in *FAM118A*). The association of *FAM118A* with pneumonia is unclear (Supplementary Text) and needs to be validated in an independent cohort. The most remarkable findings of the present study come from exploring the pathogenicity accumulated in single genes; the most significant ones being *COL13A1*, *C20orf96*, *HEXIM1*, *NPAS4*, and *STRC*. The main interest of these candidates is that none of them has been previously signaled as possible candidates for pneumonia; this may be explained, at least in part, by the fact that our study is based on whole exome sequencing, in contrast to previous studies that focused on candidate genes or GWAS approaches. At least, *COL13A1*, *HEXIM1*, and *NPAS4*, have been previously reported to be involved in respiratory-related disorders (Supplementary Text). By exploring gene expression of the top genes in available datasets of non-COVID-19 and COVID-19 pneumonia patients, we found supporting evidence for these new candidates. AUC-ROC curves of gene expression values were evaluated for the most significant genes. We found that for both severe COVID-19 and non-COVID-19 pneumonia conditions, combinations of these genes might jointly predict the disease outcome (Fig. 1F). This adds indirect evidence supporting the involvement of these genes in the pneumonia phenotype. The cohorts and datasets analyzed do not allow us to determine whether the statistical associations observed are specific to the European population. In the case of exome data, the primary findings were derived from a gene-based test that assesses variation by aggregating SNP variations. When it comes to gene expression data, it is not possible to infer genetic background from microarray data.[4] Although with limitations (Supplementary Text), we have gathered reasonable genomic evidence at the gene association and gene expression levels that provides a suggestive association between these genes and their contribution to pneumonia susceptibility.

Our exploratory analysis of genome-wide susceptibility to CAP has revealed a few new gene candidates. The causal link between these candidates and the disease is unclear, but the suggestive statistical association evidence, coupled with consistent gene expression data, opens new research scenarios that need further investigation.

1000 GP datasets; Middle: Multidimensional scaling analysis for the patient's cohort together with other European 1000 GP datasets; Right: Estimated ancestral individual values in the sampled cohort and reference 1000 GP datasets using ADMIXTURE. CAP, community-acquired pneumonia; CON, healthy controls; CHB, Han Chinese in Beijing (China); ESN, Esan (Nigeria); IBS, Iberian population (Spain); JPT, Japanese in Tokyo (Japan); PEL, Peruvians in Lima (Peru); TSI, Tuscans (Italy); YRI, Yoruba in Ibadan (Nigeria). **(B)** Gene-base association analysis using the SKAT-O collapsing association test. Manhattan plots (left) and QQ-plots (right) of *P*-values for the SKAT-O collapsing association test for all the variants in the dataset (top), and only the rare variants (minor allele frequency (MAF) < 0.05) (bottom). **(C—E)** Differential gene expression between cases and healthy controls for the best gene candidates. Only the most significant findings for each group of analysis (adult [C] and pediatric [D] non-COVID-19 pneumonia, and COVID-19 pneumonia [E]) are shown (see Table S3). In the COVID-19 dataset, we only show the severity category (WHO score > 5; W5/7). Correlation between gene expression of candidate genes and the time from symptom onset to sample collection is also displayed for the COVID-19 dataset. **(F)** ROC-AUC for the best candidate genes and the disease scenarios analyzed in the present study. Left: Non-COVID-19 pneumonia in adults; Middle: Non-COVID-19 pneumonia in children; Right: Severe COVID-19 pneumonia. We display the curves using the combination of the genes that show the best AUC-ROC values.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Ethics declaration

The study was conducted according to the guidelines of the Declaration of Helsinki, and approved by the Institutional Regional Ethics Committee of Madrid (CEIC-R) with the reference PFI-PRE-2015-01 on 12 January 2016. Informed consent was obtained from all subjects included in the study.

Author contributions

F.M.-T., I.C., and C.M. contributed to the conception and design of the work; The CAPPRIC study group contributed the samples and performed the acquisition of the clinical data and the follow-up; I.R.-C., A.D., F.C.-V., and C.R.-T. contributed to the interpretation of the clinical data; J.P.-S., S.V.-L., X.B., A.G.-C., A.C.-M., S.P., M.J.C.-T., and A.S. analyzed the sequence genome data; J.P.-S. and A.S. drafted the manuscript; all authors reviewed the manuscript critically for important intellectual content and provided final approval of the version to be published. All authors read and agreed to the published version of the manuscript.

Conflict of interests

F.M.-T. reports grants to their institutions from Pfizer S.L.U., Madrid, Spain for this study, and he has received honoraria from GSK, Pfizer S.L.U., Sanofi Pasteur, MSD, Seqirus, and Janssen for taking part in advisory boards and expert meetings, and for acting as speaker in congresses outside the scope of the submitted work. F.M.-T. has also acted as principal investigator in RCTs of the above-mentioned companies, and Ablynx, Regeneron, Roche, Abbott, Novavax, and MedImmune, with honoraria paid to his institution. I.R.-C. has participated in advisory boards organized by MSD, GSK, Sanofi, and Pfizer. I.R.-C. has been involved in clinical trials funded by Ablynx, Abbot, Seqirus, Sanofi Pasteur MSD, Cubist, Wyeth, Merck, Pfizer, Roche, Regeneron, Jansen, Medimmune, Novavax, Novartis, and GSK, with the funds were awarded to the institution; A.D. received funds (awarded to the institution) by participating as researcher of GSK and Pfizer S.L.U. pneumococcal vaccines clinical trials. I.C. and C.M. are employees of Pfizer S.L.U., Madrid, Spain.

Funding

This study was sponsored by Pfizer. It also received support from Instituto de Salud Carlos III ([ISCIII] TRINEO: No. PI22/00162; DIAVIR: No. DTS19/00049; Resvi-Omics: No. PI19/01039 (to A.S.); ReSVinext: No. PI16/01569; Enterogen: No. PI19/01090 (to F.M.-T).; cofinanciados FEDER, GAIN: Grupos con Potential de Crecimiento No. IN607B 2020/08 and Grupos de Referencia Competitiva No. IN607A 2023/02 (to A.S.); ACIS: BI-BACVIR No. PRIS-3 (to A.S.), and CovidPhy No. SA 304C (to A.S.); consorcio Centro de Investigación Biomédica en Red de Enfermedades Respiratorias No. CB21/06/00103 (to F.M.-T.); GEN-COVID No. IN845D2020/23 (to F.M.-T.) and Grupos de Referencia Competitiva No. IIN607A2021/05 (to F.M.-T).

Acknowledgements

The authors sincerely thank all the members of the CAPPRIC study group for their contribution to the recruitment of the study participants and their dedication and support to the CAPPRIC study.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.gendis.2023.101170.

References

1. Vila-Corcoles A, Ochoa-Gondar O, Rodriguez-Blanco T, Raga-Luria X, Gomez-Bertomeu F, EPIVAC Study Group. Epidemiology of community-acquired pneumonia in older adults: a population-based study. *Respir Med*. 2009;103(2):309—316.
2. Molina J, González-Gamarra A, Ginel L, et al. CAPPRIC study-characterization of community-acquired pneumonia in Spanish adults managed in primary care settings. *Microorganisms*. 2021;9(3):508.
3. Partnership GOES, Li Z, Wang Z, et al. Association of rare CYP39A1 variants with exfoliation syndrome involving the anterior chamber of the eye. *JAMA*. 2021;325(8):753—764.
4. Barral-Arca R, Pardo-Seco J, Bello X, Martinón-Torres F, Salas A. Ancestry patterns inferred from massive RNA-seq data. *RNA*. 2019;25(7):857—868.

Jacobo Pardo-Seco [a,b,c], Sandra Viz-Lasheras [a,b,c], Xabier Bello [a,b,c], Alberto Gómez-Carballa [a,b,c], Alba Camino-Mera [a,b,c], Sara Pischedda [a,b,c], María José Currás-Tuala [a,b,c], Irene Rivero-Calle [b,c,d], Ana Dacosta-Urbieta [b,c,d], Fernando Caamaño-Viña [b,c,d], Carmen Rodríguez-Tenreiro Sánchez [b,c,d], Isabel Cifuentes [e], Cristina Méndez [e], Chiea Chuen Khor [f], Federico Martinón-Torres [b,c,d], Antonio Salas [a,b,c,*] on behalf of the CAPPRIC study group[1]

[a] *Unidade de Xenética, Instituto de Ciencias Forenses, Facultade de Medicina, Universidade de Santiago de Compostela, and GenPoB Research Group, Instituto de Investigación Sanitaria (IDIS), Hospital Clínico Universitario de Santiago de Compostela (SERGAS), Galicia 15706, Spain*

[b] *Genetics, Vaccines and Infections Research Group (GENVIP), Instituto de Investigación Sanitaria de Santiago, Universidade de Santiago de Compostela, Santiago de Compostela, Galicia 15706, Spain*

[c] *Centro de Investigación Biomédica en Red de Enfermedades Respiratorias (CIBERES), Madrid 28029, Spain*

[d] *Translational Pediatrics and Infectious Diseases, Department of Pediatrics, Hospital Clínico Universitario de Santiago de Compostela (SERGAS), Galicia 15706, Spain*

[1] The CAPPRIC study group includes the following institutes and contributors.

[e] *Medical Department, Pfizer S.L.U., Alcobendas, Madrid 28108, Spain*
[f] *Genome Institute of Singapore, Agency for Science, Technology and Research, Singapore 138632, Singapore*

*Corresponding author. Unidade de Xenética, Instituto de Ciencias Forenses, Facultade de Medicina, Universidade de Santiago de Compostela, and Genética de Poblaciones en Biomedicina (GenPoB) Research Group, Instituto de Investigación Sanitaria (IDIS), Hospital Clínico Universitario de Santiago de Compsotela (SERGAS), Galicia 15706, Spain.
E-mail address: antonio.salas@usc.es (A. Salas)
30 June 2023
Available online 19 November 2023

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

5.4. ARTICLE 4: A 5-TRANSCRIPT SIGNATURE FOR DISCRIMINATING VIRAL AND BACTERIAL ETIOLOGY IN PEDIATRIC PNEUMONIA

Sandra Viz-Lasheras[1,2,3,4,#], Alberto Gómez-Carballa[1,2,3,4,#], Xabier Bello[1,2,3,4], Irene Rivero-Calle[1,4,5], Ana Isabel Dacosta[1,4,5], Myrsini Kaforou[6], Dominic Habgood-Coote[6], Aubrey J. Cunnington[6], Marieke Emonts[7,8,9], Jethro A. Herberg[6], Victoria J. Wright[6], Enitan D. Carrol[10,11], Stephane C. Paulus[12], Werner Zenz[13], Daniela S. Kohlfürst[13], Nina Schweintzger[13], Michiel Van der Flier[14,15], Ronald de Groot[15], Luregn J. Schlapbach[16], Philipp Agyeman[17], Andrew J. Pollard[12], Colin Fink[18], Taco T. Kuijpers[19], Suzanne Anderson[20], Ulrich Von Both[21], Marko Pokorn[22], Dace Zavadska[23], María Tsolia[24], Henriëtte A. Moll[25], Clementien Vermont[25], Michael Levin[6], Federico Martinón-Torres[1,4,5], Antonio Salas[1,2,3,4,&]; on behalf of EUCLIDS, PERFORM, and DIAMONDS consortia. DOI: https://doi.org/10.1016/j.isci.2025.111747

Affiliations

1 - Genetics, Vaccines and Infections Research Group (GenViP), Instituto de Investigación Sanitaria de Santiago (IDIS), 15706 Hospital Clínico Universitario de Santiago (SERGAS), Galicia, Spain.

2 - Unidade de Xenética, Instituto de Ciencias Forenses, Facultade de Medicina, Universidade de Santiago de Compostela, Spain.

3 - Genética de Poblaciones en Biomedicina (GenPoB) Research Group, Instituto de Investigación Sanitaria (IDIS), 15706 Hospital Clínico Universitario de Santiago (SERGAS), Galicia, Spain.

4 - Centro de Investigación Biomédica en Red de Enfermedades Respiratorias (CIBER-ES), Madrid, Spain.

5 - Translational Pediatrics and Infectious Diseases, Department of Pediatrics, 15706 Hospital Clínico Universitario de Santiago de Compostela, Santiago de Compostela, Galicia, Spain.

6 - Department of Infectious Disease, Imperial College London, London, W2 1PG, United Kingdom.

7 - Paediatric Immunology, Infectious Diseases & Allergy, Great North Children's Hospital, Newcastle upon Tyne Hospitals NHS Foundation Trust, Newcastle upon Tyne, NE1 4LP, United Kingdom.

USC DE SANTIAGO DE COMPOSTELA

8 - Translational and Clinical Research Institute, Newcastle University, Newcastle upon Tyne, NE2 4HH, United Kingdom.

9 - NIHR Newcastle Biomedical Research Centre based at Newcastle upon Tyne Hospitals NHS Trust and Newcastle University, Newcastle upon Tyne, NE4 5PL, United Kingdom.

10 - Department of Infectious Diseases, Alder Hey Children's NHS Foundation Trust, Liverpool, L12 2AP, United Kingdom.

11 - Department of Clinical Infection, Microbiology and Immunology, Institute of Infection, Veterinary and Ecological Sciences, University of Liverpool, Liverpool, L69 7BE, United Kingdom.

12 - Department of Paediatrics, University of Oxford and the NIHR Oxford Biomedical Research Centre, Oxford, OX3 9DU, United Kingdom.

13 - Department of General Paediatrics, Medical University of Graz, Graz, Auenbruggerplatz 34/2, 8036, Graz, Austria.

14 - Pediatric Infectious Diseases and Immunology, Wilhelmina Children's Hospital, University Medical Center Utrecht, Utrecht, 3508 AB, The Netherlands.

15 - Pediatric Infectious Diseases and Immunology, Amalia Children's Hospital, and Section Pediatric Infectious Diseases, Laboratory of Medical Immunology, Department of Laboratory Medicine, Radboud Institute for Molecular Life Sciences, Radboud University Medical Center, Nijmegen, 6500 HB, The Netherlands.

16 - Department of Intensive Care and Neonatology, and Children's Research Center, University Children's Hospital Zürich, University of Zürich, Zürich, Switzerland.

17 - Department of Pediatrics, Inselspital, Bern University Hospital, University of Bern, Bern, Switzerland.

18 - Micropathology Ltd, University of Warwick, Warwick, CV4 7EZ, United Kingdom.

19 - Division of Pediatric Immunology, Rheumatology and Infectious Diseases, Emma Children's Hospital, Amsterdam University Medical Center (Amsterdam UMC), Amsterdam, 1105 AZ, The Netherlands.

20 - Medical Research Council Unit at the London School of Hygiene & Tropical Medicine, Banjul, The Gambia.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

21 - Infectious Diseases, Department of Pediatrics, Dr von Hauner Children's Hospital, University Hospital, LMU Munich, Munich, Germany.

22 - Division of Paediatrics, University Medical Centre Ljubljana and Medical Faculty, University of Ljubljana, Ljubljana, Slovenia.

23 - Children's Clinical University Hospital, Rīga Stradins University, Rīga, Latvia.

24 - Second Department of Paediatrics, National and Kapodistrian University of Athens (NKUA), School of Medicine, Panagiotis & Aglaia, Kyriakou Children's Hospital, Athens, Greece.

25 - Department of Paediatric Infectious Diseases and Immunology, Erasmus MC Sophia Children's Hospital, Rotterdam, Netherlands.

Link: https://www.sciencedirect.com/science/article/pii/S2589004225000069

Abstract

Pneumonia stands as the primary cause of death among children under five, yet current diagnosis methods often result in inadequate or unnecessary treatments. Our research seeks to address this gap by identifying host transcriptomic biomarkers in the blood of children with definitive viral and bacterial pneumonia. We performed RNA sequencing on 192 prospectively collected whole blood samples, including 38 controls and 154 pneumonia cases, uncovering a 5-transcript signature (genes *FAM20A, BAG3, TDRD9, MXRA7* and *KLF14*) that effectively distinguishes bacterial from viral pneumonia (AUC: 0.95 [0.88–1.00]) Initial validation using combined definitive and probable cases yielded an AUC of 0.87 [0.77–0.97], while full validation in a new prospective cohort of 32 patients achieved an AUC of 0.92 [0.83–1]. This robust signature holds significant potential to enhance diagnostics accuracy for pediatric pneumonia, reducing diagnostic delays and unnecessary treatments, and potentially transforming clinical practice.

DOI: 10.1016/j.isci.2025.111747

JOURNAL: iScience (ISSN: 2589-0042)

IMPACT FACTOR: **4.6** (2023)

CATHEGORY: MULTIDISCIPLINARY SCIENCES

RELATIVE POSSITION IN THE CATHEGORY: **19/134 (Q1)**

OTHER QUALITY EVIDENCE: iScience (ISSN: 2589-0042) is an international peer-reviewed open access journal focused on publishing high-quality multidisciplinary research that advances knowledge across the life sciences, physical sciences, and interdisciplinary areas, promoting scientific innovation and discovery. The journal had **2.114** citations in 2023.

PATENT APPLICATTION: There is a European Patent application derived from this work under application number No. 24 382 240.0.

AUTHOR CONTRIBUTION: In the present study, the doctoral student took part in the study conceptualization and design, computational model and machine learning approaches development, data analysis and writing process.

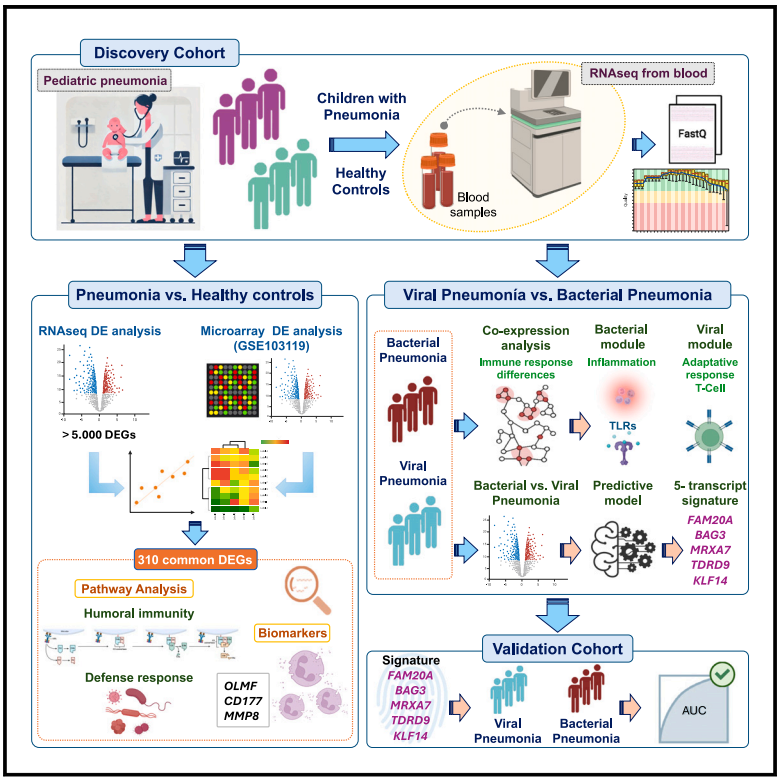
USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

iScience

Article

A 5-transcript signature for discriminating viral and bacterial etiology in pediatric pneumonia

Graphical abstract



Highlights

- Over 5,000 genes are differentially expressed in pneumonia vs. healthy controls
- Pneumonia DEPs: neutrophil degranulation, humoral response, and inflammation
- A 5-transcript host signature differentiates bacterial from viral pneumonia
- Enhances diagnosis, reducing unnecessary tests and treatments

Authors

Sandra Viz-Lasheras, Alberto Gómez-Carballa, Jacobo Pardo-Seco, ..., DIAMONDS, GENDRES and, PERFORM consortia

Correspondence

antonio.salas@usc.es

In brief

Pediatrics; Diagnostics; Body substance sample; Clinical microbiology; Transcriptomics

Viz-Lasheras et al., 2025, iScience *28*, 111747
February 21, 2025 © 2025 The Author(s). Published by Elsevier Inc.
https://doi.org/10.1016/j.isci.2025.111747

CellPress

iScience

CellPress
OPEN ACCESS

Article

A 5-transcript signature for discriminating viral and bacterial etiology in pediatric pneumonia

Sandra Viz-Lasheras,[1,2,3,37] Alberto Gómez-Carballa,[1,2,3,37] Jacobo Pardo-Seco,[1,2,3] Xabier Bello,[1,2,3] Irene Rivero-Calle,[2,3,4] Ana Isabel Dacosta,[2,3,4] Myrsini Kaforou,[5] Dominic Habgood-Coote,[5] Aubrey J. Cunnington,[5] Marieke Emonts,[6,7,8] Jethro A. Herberg,[5] Victoria J. Wright,[5] Enitan D. Carrol,[9,10] Stephane C. Paulus,[11] Werner Zenz,[12] Daniela S. Kohlfürst,[12] Michiel Van der Flier,[13,14] Ronald de Groot,[14] Luregn J. Schlapbach,[15] Philipp Agyeman,[16] Andrew J. Pollard,[11] Colin Fink,[17] Taco T. Kuijpers,[18] Suzanne Anderson,[19] Cristina Calvo,[20,21,22,23,24]

(Author list continued on next page)

SUMMARY

Pneumonia stands as the primary cause of death among children under five, yet current diagnosis methods often result in inadequate or unnecessary treatments. Our research seeks to address this gap by identifying host transcriptomic biomarkers in the blood of children with definitive viral and bacterial pneumonia. We performed RNA sequencing on 192 prospectively collected whole blood samples, including 38 controls and 154 pneumonia cases, uncovering a 5-transcript signature (genes *FAM20A*, *BAG3*, *TDRD9*, *MXRA7*, and *KLF14*) that effectively distinguishes bacterial from viral pneumonia (area under the curve (AUC): 0.95 [0.88–1.00]). Initial validation using combined definitive and probable cases yielded an AUC of 0.87 [0.77–0.97], while full validation in a new prospective cohort of 32 patients achieved an AUC of 0.92 [0.83–1.00]. This robust signature holds significant potential to enhance diagnostics accuracy for pediatric pneumonia, reducing diagnostic delays and unnecessary treatments and potentially transforming clinical practice.

INTRODUCTION

Pneumonia is an acute respiratory infection and one of the leading causes of morbidity and mortality worldwide. In children younger than 5 (excluding the neonatal period), it remains the primary cause of death.[1,2] Common symptoms in children include fever, tachypnea, cough, chest pain, and difficulty in breathing.[3] Pneumonia can be caused by a variety of microorganisms, including bacteria, viruses, and fungi, depending on age groups and the specific epidemiological situation. However, *Streptococcus pneumoniae* is the most common bacterial cause of pneumonia across all age group.[2,4] The causative microorganisms also vary significantly between the two main types of pneumonia: community-acquired pneumonia (CAP) and hospital-acquired pneumonia (HAP). *Pseudomonas aeruginosa*, *Staphylococcus aureus*, and *Enterobacter* are the most common causes of HAP[5] whereas *Streptococcus pneumoniae*; *Staphylococcus aureus*; *Streptococcus pyogenes*; *Haemophilus influenzae*; *Mycoplasma pneumoniae*; *Legionella pneumophila*; and respiratory viruses, such as respiratory syncytial virus, rhinovirus, or influenza virus A/B, are the most common microorganisms responsible for CAP.[3,6] The introduction of improved conjugate vaccines against *Haemophilus influenzae* type b and *Streptococcus pneumoniae* has led to a reduction in the incidence and severity of childhood pneumonia, resulting in significant changes in the proportions of CAP cases etiologies.[7,8]

The clinical presentation of pneumonia is diverse and may overlap with other respiratory conditions such as asthma, bronchiolitis, pertussis, lung abscess, bronchiectasis, or malaria, among others.[6,9–14] Accurate assessment of disease severity[15] is crucial for effective management, influencing decisions on antibiotic prescriptions and hospitalization.[16,17] Despite efforts

[1]Unidade de Xenética, Instituto de Ciencias Forenses, Facultade de Medicina, Universidade de Santiago de Compostela, and Genética de Poblaciones en Biomedicina (GenPoB) Research Group, Instituto de Investigación Sanitaria (IDIS), 15706 Hospital Clínico Universitario de Santiago (SERGAS), Galicia, Spain
[2]Genetics, Vaccines and Infections Research Group (GenViP), Instituto de Investigación Sanitaria de Santiago, 15706 Universidade de Santiago de Compostela, Santiago de Compostela, Galicia, Spain
[3]Centro de Investigación Biomédica en Red de Enfermedades Respiratorias (CIBER-ES), Madrid, Spain
[4]Translational Pediatrics and Infectious Diseases, Department of Pediatrics, 15706 Hospital Clínico Universitario de Santiago de Compostela, Santiago de Compostela, Galicia, Spain
[5]Department of Infectious Disease, Imperial College London, London W2 1PG, UK
[6]Great North Children's Hospital, Paediatric Immunology, Infectious Diseases & Allergy, Newcastle upon Tyne Hospitals NHS Foundation Trust, Newcastle upon Tyne NE1 4LP, UK

(Affiliations continued on next page)

This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

María del Carmen Martínez-Padilla,[25] Ana Pérez-Aragón,[26] Esteban Gómez-Sánchez,[27] Juan Valencia-Ramos,[27] Francisco Giménez-Sánchez,[28] Paula Alonso-Quintela,[29] Laura Moreno-Galarraga,[30,31] Ulrich von Both,[32] Marko Pokorn,[33] Dace Zavadska,[34] María Tsolia,[35] Clementien L. Vermont,[36] Henriëtte A. Moll,[36] Michael Levin,[5] Federico Martinón-Torres,[2,3,4] and Antonio Salas[1,2,3,38,*] on behalf of EUCLIDS, DIAMONDS, GENDRES and, PERFORM consortia

[7]Translational and Clinical Research Institute, Newcastle University, Newcastle upon Tyne NE2 4HH, UK
[8]NIHR Newcastle Biomedical Research Centre based at Newcastle upon Tyne Hospitals NHS Trust and Newcastle University, Newcastle upon Tyne NE4 5PL, UK
[9]Department of Infectious Diseases, Alder Hey Children's NHS Foundation Trust, Liverpool L12 2AP, UK
[10]Department of Clinical Infection, Microbiology and Immunology, Institute of Infection, Veterinary and Ecological Sciences, University of Liverpool, Liverpool L69 7BE, UK
[11]Department of Paediatrics, University of Oxford and the NIHR Oxford Biomedical Research Centre, Oxford OX3 9DU, UK
[12]Department of General Paediatrics, Medical University of Graz, Graz, Auenbruggerplatz 34/2 8036, Graz, Austria
[13]Pediatric Infectious Diseases and Immunology, Wilhelmina Children's Hospital, University Medical Center Utrecht, Utrecht 3508 AB, the Netherlands
[14]Pediatric Infectious Diseases and Immunology, Amalia Children's Hospital, and Section Pediatric Infectious Diseases, Laboratory of Medical Immunology, Department of Laboratory Medicine, Radboud Institute for Molecular Life Sciences, Radboud University Medical Center, Nijmegen 6500 HB, the Netherlands
[15]Department of Intensive Care and Neonatology, and Children's Research Center, University Children's Hospital Zürich, University of Zürich, Zürich, Switzerland
[16]Department of Pediatrics, Inselspital, Bern University Hospital, University of Bern, Bern, Switzerland
[17]Micropathology Ltd, University of Warwick, Warwick CV4 7EZ, UK
[18]Division of Pediatric Immunology, Rheumatology and Infectious diseases, Emma Children's Hospital, Amsterdam Univiersyt Medical Center (Amsterdam UMC), Amsterdam 1105 AZ, the Netherlands
[19]Medical Research Council Unit at the London School of Hygiene & Tropical Medicine, Banjul, The Gambia
[20]General Pediatrics, Infectious and Tropical Diseases Department, Hospital La Paz, 28046 Madrid, Spain
[21]La Paz Research Institute (IdiPAZ), 28029 Madrid, Spain
[22]Faculty of Medicine, Universidad Autónoma de Madrid (UAM), 28049 Madrid, Spain
[23]Centro de Investigación Biomédica en Red en Enfermedades Infecciosas (CIBERINFEC), Madrid, Spain
[24]Red de Investigación Traslacional en Infectología Pediátrica (RITIP), Madrid, Spain
[25]Unidad de Cuidados Intensivos Pediátricos, Complejo Hospitalario de Jaen, Jaen, Spain
[26]Hospital Universitario Virgen de las Nieves, Servicio de Pediatría, Granada, Spain
[27]Department of Pediatric Intensive Care Unit, Hospital Universitario de Burgos, Burgos, Spain
[28]Instituto Hispalense de Pediatría, Instituto Balmis de Vacunas, Almería, Spain
[29]Neonatal Intensive Care Unit, Complejo Asistencial Universitario de León, León, Spain
[30]Department of Pediatrics, Complejo Hospitalario de Navarra, Servicio Navarro de Salud, Pamplona, Spain
[31]IdiSNA (Instituto de Investigación Sanitaria de Navarra), Navarra Institute for Health Research, Pamplona, Spain
[32]Infectious Diseases, Department of Pediatrics, Dr von Hauner Children's Hospital, University Hospital, LMU Munich, Munich, Germany
[33]Division of Paediatrics, University Medical Centre Ljubljana and Medical Faculty, University of Ljubljana, Ljubljana, Slovenia
[34]Children's Clinical University Hospital, Rīga Stradins University, Rīga, Latvia
[35]Second Department of Paediatrics, National and Kapodistrian University of Athens (NKUA), School of Medicine, Panagiotis & Aglaia, Kyriakou Children's Hospital, Athens, Greece
[36]Department of Pediatrics, Erasmus MC, Rotterdam, the Netherlands
[37]These authors contributed equally
[38]Lead contact
*Correspondence: antonio.salas@usc.es
https://doi.org/10.1016/j.isci.2025.111747

to enhance international guidelines,[9,10,18] the variability in causative microorganisms and symptoms poses a significant challenge for clinicians in healthcare centers.[14,19–22] Current diagnostic criteria rely on nonspecific symptoms, chest imaging, and inconclusive laboratory analysis,[14,19,23,24] with limitations in differentiating the etiology of the disease (i.e., viral and bacterial). Obtaining lower respiratory tract samples in children is often difficult.[19,20] Consequently, misdiagnosis of bacterial pneumonia is common, leading to increased costs, unnecessary medical tests, hospital admissions, and incorrect antibiotic prescriptions.[6,25]

Given these challenges, numerous studies have suggested the potential use of biomarkers as supplementary tools in the managing and diagnosis of pneumonia.[9,16,25–27] Serum or plasma C-reactive protein (CRP) and procalcitonin (PCT) have been identified as potential biomarkers to assist in the diagnosis and prognosis of pneumonia patients.[9,16,28,29] The study of host transcriptomic in the field of infectious disease has become increasingly important in recent years, leading to advancements in understanding host-pathogen interactions and the development of useful tools for diagnosing and prognosing diseases, including the development of point-of-care devices.[30–35] There are a growing number of studies investigating host gene expression biomarkers related to different infectious diseases, such as tuberculosis,[36–38] H1N1,[39] Respiratory Syncitial Virus (RSV),[40] and rotavirus,[41] and also aiming at differentiating bacterial from viral infection.[30,35,42,43] However, the study of host transcriptomics in pneumonia patients has not received as much attention, and only a few studies

describing blood transcriptomic biomarkers related to CAP in adults are available in the literature.[27,44] Some studies have recently focused on pneumonia outcome prediction[45–47] or pneumonia etiology discrimination.[48,49]

Regarding pediatric pneumonia, most studies have focused on the potential use of already described biomarkers, such as white blood cell (WBC) count, neutrophil percentage (NP), serum CRP, and PCT. However, results indicate that these biomarkers are not good predictors for pediatric pneumonia due to their inability to differentiate pneumonia from bacterial and viral origins.[50,51] To our knowledge, there is only one study investigating blood transcriptomics in the context of pediatric pneumonia etiology, but in a very specific population from a malaria-endemic area and using a signature with a high number of genes.[52] Therefore, further studies are necessary to approach this complex disease in children.

In light of these challenges in pneumonia diagnosis and management in children, we analyze the whole blood transcriptome of the largest pediatric pneumonia cohort recruited to date with the aim of (1) shedding light on the molecular mechanisms underlying pediatric pneumonia, providing insights on new therapeutic targets; (2) investigating host's differential response to pneumonia of bacterial and viral origin; and (3) proposing a minimal transcriptomic signature that allows for the differentiation between viral and bacterial pneumonia in children.

RESULTS

Discovery cohort description

In this study, blood samples were collected from 192 children (see Figure 1 and STAR Methods for details on the experimental design). Among them, 154 were hospitalized with pneumonia (median age: 3.4 years; 51.3% male) and 38 were healthy controls (median age: 3.3 years; 68.4% male). Clinical and demographic details of the entire cohort and sub-categories used in different analyses are provided in Table 1. Among the pneumonia patients, 25% were admitted to Pediatric Intensive Care Unit (PICU), 3% died, and 39% required oxygen support (16.2% with invasive ventilation). No statistically significant differences were found between the bacterial and viral pneumonia sub-groups for these parameters. The majority of bacterial pneumonias were caused by *Streptococcus pneumoniae* (47.5%) followed by *Staphylococcus aureus* (10.0%) and *Streptococcus pyogenes* (7.5%) among others (35.0%). Viral pneumonias were caused by influenza virus (22.2%), bocavirus (22.2%), RSV (22.2%), rhinovirus (11.1%), adenovirus (11.1%), and parainfluenza (11.1%).

As expected, CRP values were significantly higher in definitive bacterial (DB) + probable bacterial (PB) compared to definitive viral (DV) + probable viral (PV) (p value = 0.001). CRP values were higher in the DB group compared to DV; this difference approached the significance threshold (p value = 0.062), likely due to limited statistical power, especially given the low number of samples in the DV group, coupled with missing data. For the remaining blood tests, there were no significant statistical differences between bacterial and viral sub-groups.

Differentially expressed genes in pneumonia

Through a comparative analysis of transcriptomes in children with pneumonia vs. healthy controls, we identified 5,486 differentially expressed genes (DEGs), using a significance threshold of Benjamini-Hochberg false discovery rate (FDR) 5% (see Table S1). Within this set of DEGs, 2,716 were found to be upregulated, while 2,770 were downregulated.

To gain insight into the overall variation in gene expression, we performed a principal component analysis (PCA) on a subset of 100 highly variable genes out of a total of 192 samples. The first principal component (PC1), explaining approximately 30% of the variation, effectively segregated the samples into two main clusters, corresponding to pneumonia samples and healthy control samples (Figure 2A).

To prioritize significant changes in gene expression, we applied a selective criterion, including an adjusted p value < 0.05 and a $\log_2$FC > |2|. This selection process yielded 272 DEGs (Figure 2B), with 208 exhibiting upregulation and 64 genes showing downregulation, indicating a prevailing pattern of overexpression among these genes.

Next, we compared the DEGs identified in our RNA sequencing (RNA-seq) cohort, with those inferred from the analysis of the Wallihan et al. study[53] (accession number GSE103119). The dataset by Wallihan et al. included transcriptomic data from blood samples of pediatric pneumonia cases (n = 152) and healthy controls (n = 20). Our comparison revealed 1,729 common DEGs (Table S2). Focusing on the DEGs with significant expression changes compared to healthy controls (adjusted p value < 0.05 and $\log_2$FC > |1|; n = 310 genes), we found a strong positive correlation (ρ = 0.81; p value < $2.2e^{-16}$) between $\log_2$FC values obtained from both datasets (Table S2; Figure 2C). Furthermore, the expression pattern of these 310 shared genes clearly segregated pneumonia cases from healthy controls, forming two distinct clusters in both our RNA-seq discovery cohort and Wallihan's microarray dataset (Figure 2D).

Functional enrichment analysis of DEGs

To further characterize DEGs from a functional point of view, we conducted an over-representation analysis (ORA) and a gene set enrichment analysis (GSEA) based on Gene Ontology (GO) and Reactome using only validated DEGs with a positive correlation between the discovery and the validation datasets. We set a threshold of an adjusted p value < 0.05 for significance and a $|\log_2FC|$ > 1.5. The analysis revealed that the most significant pathways (adjusted p value < 0.01) are associated with essential immune processes, primarily related to the innate response to severe CAP (Table S3; Figure 2E).

The Reactome-ORA analysis indicated a robust association between the selected DEGs, including *matrix metalloproteinase-8 (MMP8)*, *OLFM4*, or *CD177*, and an innate response characterized by "neutrophil degranulation." This pathway displayed the highest significance (adjusted p value = $6.5e^{-12}$; Table S3). Consistently, GO-ORA results identified differences in neutrophil-related immune pathways, including "leukocyte mediated immunity" (adjusted p value = $7.4e^{-5}$), "myeloid leukocyte mediated immunity" (adjusted p value = 0.001), "neutrophil mediated immunity" (adjusted p value = $5.1e^{-5}$), "myeloid leukocyte activation" (adjusted p value = $1.4e^{-5}$), "granulocyte activation" (adjusted p value = 0.002), and "neutrophil activation" (adjusted p value = 0.001) (Table S3; Figure 2E).

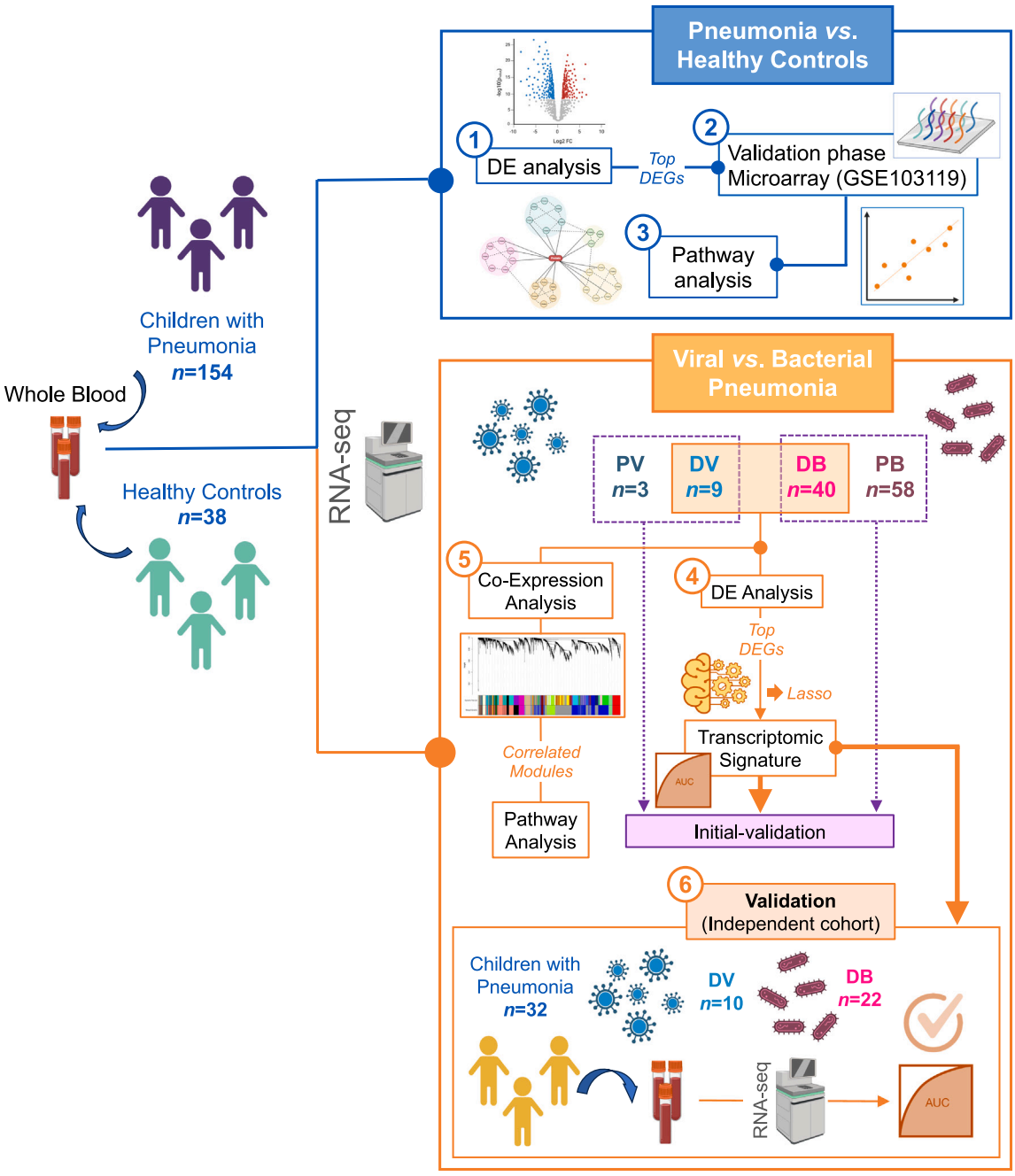


Figure 1. Scheme of the study design

The figure was built using BioRender resources; created in BioRender. BioRender.com/b32x666.

Furthermore, there is a significant overlap in enrichment results from GO and Reactome, particularly regarding humoral immunity processes. Notably, GO-ORA identified "humoral immune response" (adjusted p value = $4.4e^{-5}$), "antimicrobial humoral response" (adjusted p value = $1.4e^{-5}$), and "antimicrobial humoral immune response mediated by antimicrobial peptide" (adjusted p value = 0.0003) (Table S3; Figure 2E) as highly significant pathways. Likewise, Reactome-ORA highlighted the involvement of various humoral responses in severe CAP, including the complement system ("regulation of complement cascade" [adjusted p value = 0.001] and "complement cascade" [adjusted p value = 0.003]), as well as humoral immunity mediated by "antimicrobial peptides" (adjusted p value = $1.2e^{-9}$) (Table S3; Figure 2E).

Most significant pathways from GO-ORA were those implicated in "defense response to bacterium" (adjusted p value = $2.3e^{-8}$),

Table 1. Clinical features of the pneumonia patient's cohort

	Pneumonia vs. healthy controls			DV vs. DB					DV + PV vs. DB + PB				
	Controls (*n* = 38)	Pneumonia (*n* = 154)	*n*	DB (*n* = 40)	*n*	DV (*n* = 9)	*n*	*p* value	DB + PB (*n* = 98)	*n*	DV + PV (*n* = 12)	*n*	*p* value
Gender (male)	68.4% (36/38)	51.3% (79/154)	154	62.5% (25/40)	40	55.6% (5/9)	9	0.72	53.1% (52/98)	98	58.3% (7/12)	12	0.769
Age (years)	3.3 (2.3–8.2)	3.36 (1.24–6.17)	153	2.77 (1.25–5.78)	40	2.75 (1.52–4.76)	8	0.923	3.78 (1.54–6.285)	98	1.93 (0.94–3.88)	11	0.096
Days hospitalized	–	5 (3–10)	149	6.5 (3.5–14)	38	6 (4.75–8.5)	8	0.621	5 (3–9)	95	6 (3.5–9)	11	0.536
PICU admission	–	24.7% (38/154)	154	30% (12/40)	40	22.2% (2/9)	9	1	22.4% (22/98)	98	25% (3/12)	12	1
PICU days	–	7 (3–11)	37	11 (6.5–16)	11	6.5 (4.25–8.75)	2	0.372	6 (3–11)	21	2 (2–6.5)	3	0.356
Death	–	3.2% (5/154)	154	7.5% (3/40)	40	11.1% (1/9)	9	0.569	4.1% (4/98)	98	8.3% (1/12)	12	0.445
Treatments													
Inotropes	–	9.7% (15/154)	154	5% (2/40)	40	11.1% (1/9)	9	0.464	6.1% (6/98)	98	8.3% (1/12)	12	0.565
IV	–	16.2% (25/154)	154	15% (6/40)	40	22.2% (2/9)	9	0.628	11.2% (11/98)	98	25% (3/12)	12	0.18
NIV	–	11% (17/154)	154	15% (6/40)	40	11.1% (1/9)	9	1	8.2% (8/98)	98	8.3% (1/12)	12	1
Oxygen	–	39% (60/154)	154	40% (16/40)	40	33.3% (3/9)	9	1	34.7% (34/98)	98	41.7% (5/12)	12	0.751
Analytical parameters													
CRP (mg/L)	–	81 (30.1–215.5)	59	146 (34.5–244)	11	28.5 (11.85–30.15)	6	0.062	190 (81–273)	33	27 (5.9–30.1)	7	0.001*
Platelets (10^9/L)	–	322 (259–433)	135	287.5 (175.5–405.5)	28	491 (283–507)	5	0.173	324 (262.75–421.5)	84	398 (273.5–495)	8	0.458
White cells (10^9/L)	–	17.2 (10.2–25)	133	18.75 (10.075–26.075)	28	7.2 (6.35–16.25)	3	0.423	18.65 (12.18–26.72)	84	6.85 (5.75–20.775)	6	0.089
Lactate (mmol/L)	–	1.9 (1.18–2.65)	30	2.5 (2.3–2.7)	5	2.5 (2.5–2.5)	1	1	2.2 (1.48–2.55)	16	2.5 (2.5–2.5)	1	0.539
Neutrophils (10^9/L)	–	11.87 (6.47–19.95)	128	13.4 (8.05–20.88)	26	12.06 (7.94–16.18)	2	0.762	14.61 (7.82–23.62)	82	4.16 (3.82–18.6)	5	0.111
Monocytes (10^9/L)	–	0.8 (0.17–1.7)	84	0.61 (0.125–1.2)	14	0.8 (0.4–1.15)	3	0.801	0.955 (0.168–1.92)	56	0.02 (0–0.8)	5	0.068
Lymphocytes (10^9/L)	–	3.17 (1.67–5.4)	130	1.95 (1.4–5.65)	27	4.1 (3.42–8.55)	3	0.213	2.6 (1.5–5.4)	83	3.42 (2.1475–6.35)	6	0.381

Wilcoxon and Fisher exact test were used to assess statistical significance between groups in numerical and categorical variables, respectively. IV, invasive ventilation; NIV, non-invasive ventilation. * indicates statistical significance.

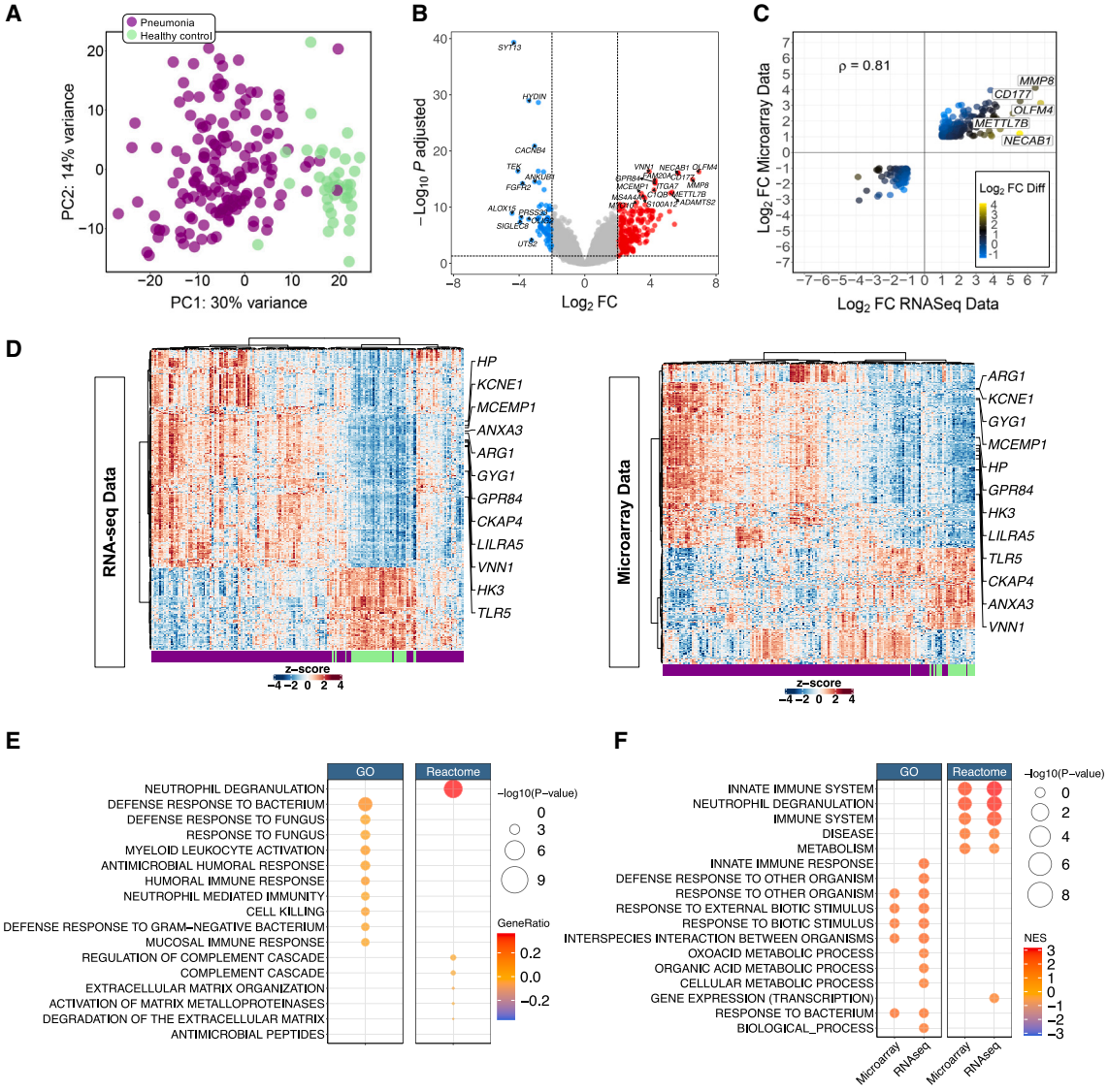


Figure 2. Pneumonia patients vs. healthy controls

(A) PCA of transcriptome profiles from pneumonia and healthy control samples. Two first principal components (PC1 and PC2) are shown.

(B) Volcano plot showing the DEGs between conditions: pneumonia vs. healthy control. Downregulated genes are colored blue, and upregulated genes are colored red (thresholds: adjusted p value = 0.05, $\log_2$FC = |2|).

(C) Correlation between $\log_2$FC of DEGs obtained from the comparison pneumonia vs. healthy controls in RNA-seq/microarray data. Only genes with adjusted p value < 0.05 and $\log_2$FC > |1| are displayed. The color scale represents the differences between the $\log_2$FC values of both analyses. The p value of the correlation is $2.2e^{-16}$, and only names of genes with a $\text{Log}_2\text{FC} > |5|$ are shown.

(D) Two-way hierarchical clustering analysis heatmap of DEGs between pneumonia and healthy control samples in RNA-seq and microarray validation cohorts. Each row represents one transcript; each column represents one patient. The bar at the bottom indicates the sample phenotype. Only genes with a $\log_2$FC > |1| and adjusted p value < 0.05 were represented in the heatmap, and only the genes that were common in the top 40 with the lower p value in both analysis were printed. Expression intensity is indicated by color (red, high expression; blue, low expression).

(E) Dot plot from over-representation analysis (ORA) pathway analysis of common DEGs between RNA-seq and microarray cohort with adjusted p value (FDR) < 0.05 and a $\log_2$FC > |1.5| for the comparison pneumonia patients vs. healthy controls using Gene Ontology (GO) and Reactome as reference. Size along the x axis indicates the number of genes in the input list that are annotated to the corresponding term/number of genes in the input list (gene ratio). Dot colors correspond to the different FDR p values associated with the pathways.

(F) Dot plot from GSEA pathway analysis of common DEGs between RNA-seq and microarray cohort with FDR p value < 0.05 and a $\log_2$FC > |1.5| for the comparison pneumonia patients vs. healthy controls using GO and Reactome as reference. Size of the dots corresponds to FDR p values associated with the pathways. Dot colors correspond to the pathway normalized enrichment scores (NESs) values.

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

"defense response to fungus" (adjusted p value = $1.0e^{-5}$), and related processes ("defense response to gram-negative bacterium" and "antibacterial humoral response") (Table S3; Figure 2E). It is not surprising that genes related to defense against bacteria were overrepresented in the overall analysis as most of the samples in the dataset correspond to bacterial-origin pneumonia.

Additionally, GO-ORA analysis highlighted a significant association of CAP with biological processes that play essential roles not only in innate protection against viral and bacterial invasion but also in disease pathogenesis. This encompasses regulatory pathways related to transcription factors, such as the pro-inflammatory nuclear factor κB (NF-κB) ("regulation of DNA-binding transcription factor activity," "positive regulation of DNA-binding transcription factor activity," and "positive regulation of NF-κB transcription factor activity"), as well as pathways indicating inflammatory response ("regulation of inflammatory response," "positive regulation of inflammatory response," and "acute inflammatory response") and activation of the cytokine signaling system ("positive regulation of cytokine production" and "cytokine-mediated signaling pathway") (Table S3; Figure 2E).

The results from the GSEA closely mirrored those obtained from both microarray[53] and our RNA-seq cohorts. Remarkably, all differentially regulated pathways exhibited significant overactivation in pneumonia cases (normalized enrichment score [NES] > 0) (Table S3; Figure 2F). The most significant pathways identified in the GO and Reactome analysis overlapped with those detected in the ORA, highlighting key processes such as the "response to other organism" and "response to bacterium" in GO. Additionally, pathways related to the "innate immune system" and processes like "neutrophil degranulation" in Reactome were also prominently featured.

Bacterial vs. viral pneumonia: Diagnostic signature

We conducted an additional PCA analysis to compare the transcriptomic profiles of pneumonia in DB and DV infections. Our goal was to investigate how these two groups cluster in relation to each other and in comparison to the control group. In the first PCA (Figure 3A), pneumonias separated distinctly from the control group along PC1, which accounted for 45% of the variation. Specifically, the DB profiles clustered at one pole of the plot, the controls in the opposite side, and the DV profiles in between. A separate PCA, including only DB (n = 40) and DV (n = 9) transcriptomic samples, revealed a relatively weak separation between the two groups. DV profiles primarily clustered to one side along its PC1 (28% of the variation), albeit with some slight mixing with other DB samples (Figure 3A).

Our differential expression analysis between DB and DV pneumonias identified 282 DEGs (FDR of 5%; Table S4). The majority showed under-expression in DB with respect to DV pneumonia (n = 217; 77%). However, when considering genes with more substantial expression changes (setting a threshold of adjusted p value < 0.05 and $|\log_2 FC| > 2$), we found that the proportion of over-expressed genes in DB pneumonia (n = 30; 55%) was slightly higher than the proportion of under-expressed genes (n = 25; 45%; Table S4; Figure 3B). Interestingly, *ALB* was the gene yielding the lowest adjusted p value ($1.88e^{-7}$) in the comparison between DB and DV patients (Table S4; Figures 3B and S1).

To further investigate, we selected a subset of DEGs candidates from the DV vs. DB differential expression in pneumonia samples (n = 36; see STAR Methods). We then employed the least absolute shrinkage and selection operator (LASSO) regression algorithm to identify the optimal transcriptomic signature for distinguishing between these two types of infection.

The best model for differentiation between viral and bacterial pneumonia comprised five genes (*FAM20A*, *BAG3*, *TDRD9*, *MXRA7*, and *KLF14*), forming the optimal signature (Table S5). Among these genes, four were over-expressed (*FAM20A*, *TDRD9*, *MXRA7*, and *KLF14*) while one (*BAG3*) showed under-expression in bacterial compared to viral pneumonia. This transcriptomic signature statistically discriminated between viral and bacterial pneumonia with high accuracy (p value = $3.0e^{-5}$; optimal cutoff value = 2.204; Figure 3C) and demonstrated excellent performance in the training set, generating an area under the curve (AUC) value of 0.95 [0.88–1.00] (sensitivity = 0.80 and specificity = 1.00) (Figure 3C; Table S6). Moreover, the signature reliably differentiated bacterial pneumonia from healthy controls (AUC: 0.91 [0.84–0.98]) (Figure S2).

Due to the limited number of samples classified as viral pneumonia (DV = 9; PV = 3), we were unable to split the dataset into training and test sets conventionally. Instead, we adopted an initial validation approach, evaluating the performance of the transcriptomic signature in a test dataset that included not only the samples confidentially classified as DB and DV, which had been already used in the training dataset for the signature discovery (DB = 40; DV = 9), but also all PB (n = 58) and PV (n = 3) pneumonias (DB + PB = 98; DV + PV = 12; Figure 1). This 5-transcripts signature identified in the training cohort exhibited robust discriminatory capacity when validated in this test set (p value = $2.8e^{-5}$, Figure 3D). In this scenario, the receiver operating characteristic (ROC) curve also indicates high accuracy, resulting in an AUC value of 0.87 [0.77–0.97] (sensitivity = 0.76 and specificity = 0.83; Figure 3D; Table S6). Considering only PV and PB samples, we obtained an expected decrease in the AUC value (0.69 [0.40–0.97]), as these samples are coded with a phenotype of lower clinical confidence and are therefore more likely to contain viral samples within the PB set, bacterial samples within the PV set, or mixed infections (see score values for classification in Figures 3C and 3D; see also Figure S3A).

The performance of our 5-transcriptomic signature was also compared to the previously described 2-transcript signature that generically differentiates viral from bacterial infections in febrile children using *IFI44L* and *FAM89A* transcripts.[30] Although the 2-transcript signature showed a good performance discriminating viral and bacterial pneumonia (AUC: 0.87 [0.76–0.96]) in our discovery dataset, the AUC value was lower than that obtained from the specific 5-transcript signature we propose (Figure S3B). When we tested the 2-transcripts viral/bacterial signature in the PB vs. PV subset, the AUC value was also lower than the AUC from the pneumonia signature (AUC: 0.44 [0.00–0.89]; Figure S3C). Finally, to check the best performance that we could achieve with these two transcripts in our discovery dataset, we re-trained the model using *IFI44L* and *FAM89A* transcripts. In this case we

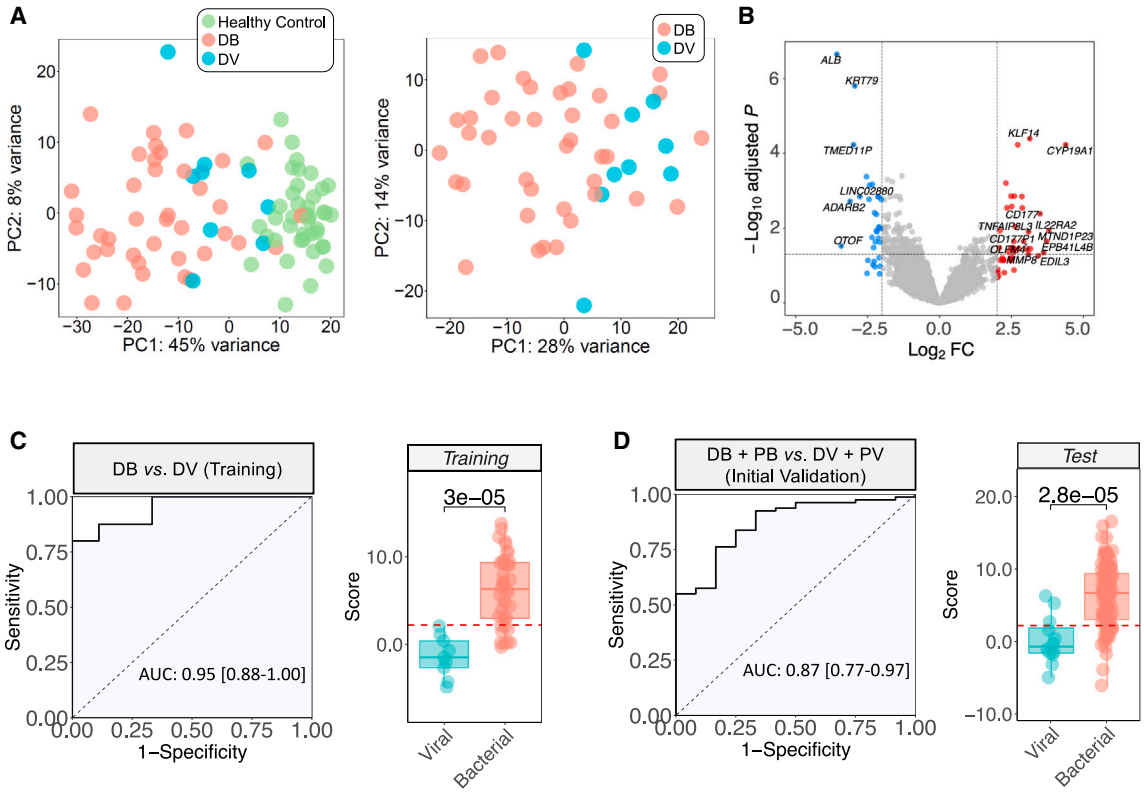


Figure 3. Viral and bacterial pneumonias

(A) PCA of transcriptome profiles of DV and DB pneumonias and healthy control samples. Two first principal components (PC1 and PC2) are shown.

(B) Volcano plot showing the DEGs between conditions: DB pneumonia vs. DV pneumonia (DV). Downregulated genes are colored blue, and upregulated genes are colored red (thresholds: adjusted *p* value = 0.05, $\log_2$FC = |2|).

(C) Receiver operating characteristic (ROC) curves based on the specific 5-transcript signature from the training cohort including the area under the curve (AUC) and 95% confident intervals (CIs) values (left). Boxplots of the predicted values using the optimal model in the training cohort. Wilcoxon *p* value is also displayed (right).

(D) Receiver operating characteristic (ROC) curves based on the specific 5-transcript signature from the test cohort including the area under the curve (AUC) and 95% confident intervals (CIs) values (left). Boxplots of the predicted values using the optimal model in the training cohort. Wilcoxon *p* value is also displayed (right). Red dashed line represents the optimal cutpoint. The boxes are defined by the upper and lower quartile (Q1 and Q3); whiskers extend to the most extreme data point, which is no more than 1.5 times the IQR from the box; the median is shown as a bold-colored horizontal line.

obtained an AUC value of 0.88 [0.77–0.98] (Figure S3D), which is still below the AUC from the specific pneumonia signature (AUC: 0.95 [0.88–1.00]).

Additionally, we tested the 5-transcripts signature and the 2-transcript signature from Herberg et al.[30] in the bacterial and viral pneumonia samples (excluding co-infections) from the dataset of Wallihan et al.[53] Both signatures yielded AUC values close to 0.5 for this dataset, with the 5-transcript signature achieving an AUC of 0.52 [0.40–0.64] and the 2-transcripts signature from Herberg et al. reaching an AUC of 0.56 [0.45–0.68]. These results indicate limited discrimination ability for these signatures, within this specific dataset.[30]

Validation of the 5-transcripts pneumonia's signature

We validated the diagnostic accuracy of the proposed signature using new gene expression data obtained from an independent cohort of children with pneumonia (Tables 2 and S7). The validation results confirm that the 5-transcript signature can accurately differentiate between DV and DB pneumonias in this new cohort, achieving an impressive AUC of 0.92 [0.83–1.00] (Figures 4A and 4B).

To ensure that the effectiveness of the signature was not influenced by disease severity or specific pathogens, we conducted a stratified analysis based on the causal pathogens of DB pneumonias—*S. pneumoniae*, *S. aureus*, and *S. pyogenes*—and by severity (mild/moderate vs. severe). The results affirmed the robustness of the signature across different bacterial pathogens: AUC values of 0.95 [0.87–1.00] for *S. pneumoniae*, 0.82 [0.54–1.00] for *S. aureus*, and a perfect AUC of 1.00 [1.00–1.00] for *S. pyogenes* (Figure 5A). Moreover, the signature consistently demonstrated high performance in distinguishing between viral and bacterial pneumonias across both mild/moderate and severe cases, with AUC values of 1.00 [1.00–1.00] for mild/moderate cases and 0.92 [0.77–1.00] for severe cases (Figure 5B).

Co-expression analysis of viral vs. bacterial pneumonia

To understand the distinct host mechanisms and key genes involved in the specific response to viral and bacterial

Table 2. Clinical features of the validation cohort

	Healthy controls	DB vs. DV				
	Controls (n = 35)	DB (n = 22)	n	DV (n = 10)	n	p value
Gender (male)	54.3% (19/35)	63.6% (14/22)	22	60% (6/10)	10	1
Age (years)	4.2 (1.24–6.17)	4.8 (0.1–17.5)	22	2.4 (0.1–15.6)	10	0.167
Days hospitalized	–	14.0 (3.5–14)	22	6 (4.75–8.5)	8	0.038
PICU admission	–	54.5% (12/22)	22	40%% (4/10)	10	0.703
PICU days	–	2.0 (0.0–14.0)	22	7,5 (2.00, 8.00)	4	0.843
Death	–	4.5% (1/22)	22	0% (0/9)	10	1
Treatments						
Inotropes	–	32% (7/22)	22	20% (2/10)	2	0.379
IV	–	31.8% (7/22)	22	30% (3/10)	9	0.750
NIV	–	13,6% (3/22)	22	30% (3/10)	9	1
Oxygen	–	54.5% (12/22)	22	60% (6/10)	9	0.637
Analytical parameters						
CRP (mg/L)	–	208.4 (46.2–375.4)	16	18.2 (1.3–52.5)	10	0.084
Platelets (10^9/L)	–	249.0 (17.0–1299.0)	19	271.5 (141.0–522.0)	10	0.449
White cells (10^9/L)	–	13.5 (4.8–28.3)	19	5.8 (2.8–11.5)	10	0.001*
Lactate (mmol/L)	–	1.4 (0.6–5.9)	8	0.7 (0.5–1.7)	4	0.147
Neutrophils (10^9/L)	–	9.6 (3.7–24.0)	15	3.0 (0.9–5.1)	9	0.001*
Monocytes (10^9/L)	–	0.8 (0.1–3.4)	15	0.4 (0.1–2.0)	8	0.258
Lymphocytes (10^9/L)	–	2.5 (0.2–6.0)	15	2.6 (0.9–4.7)	9	0.765

Wilcoxon and Fisher exact test were used to assess statistical significance between groups in numerical and categorical variables, respectively. IV, invasive ventilation; NIV, non-invasive ventilation. * indicates statistical significance.

pneumonia, we conducted a co-expression network analysis using the subset of pneumonia samples with DB or DV origin. The analysis detected 16 modules of co-expressed genes (Figure S4), with three of them significantly correlated (p value < 0.05) with viral/bacterial phenotypes (Figures 6A and 6B). The blue module (Enah//Vasp-like [EVL] module, $R = -0.4$; p value = 0.0043; 661 genes) and darked module (Transforming Growth Factor Beta 3 [TGFB3] module, $R = -0.32$; p-value = 0.025; 73 genes) exhibited negative correlation with bacterial pneumonia. In contrast, the salmon module (mitogen-activated protein kinase [MAPK]14]) showed a positive correlation (R = 0.32; p value = 0.026; 1,132 genes) with bacterial pneumonia (Figures 6A and 6B). The EVL and TGFBR3 modules clustered together on the dendrogram, indicating their involvement in similar biological processes globally activated during viral pneumonia (Figure 6A). Genes within these significant modules displayed a strong correlation between trait significance (GS) and module membership (MM), suggesting that highly interconnected genes (higher MM) within the module are closely related to the causal pathogen of pneumonia (Figure 6B). These driver genes, located in the upper right side of the plots, are pivotal for predicting viral or bacterial pneumonia.

Functional assessment of the significant modules unveiled different immune responses for viral and bacterial pneumonia. The modules over-regulated in viral pneumonia showed enrichment in terms associated with both innate and adaptative immune responses. For example, the EVL module was linked to adaptative T cell response (T cell selection; T cell receptor signaling; and T cell activation, differentiation, and proliferation) and interferon production (interferon gamma production) (Figure 6C; Table S8), while the TGBR3 module participated in "natural killer mediated immunity" (innate response) and, as in the case of the EVL module, in an adaptative response involving "leukocyte mediated cytotoxicity" and "T cell mediated immunity" (Table S8).

Conversely, the significant pathways related to genes from the bacterial MAPK14 module primarily represent a different innate response characterized by "positive regulation of cytokine production," "cytokine-mediated signaling pathway," "acute inflammatory response," "pattern recognition receptor signaling," and "myeloid leukocytes activation" (granulocytes, monocytes, macrophages, and dendritic cells). These findings were further supported by an ORA using the Reactome database, emphasizing the role of MAPK14 module genes in innate immune responses, including neutrophil degranulation, various Toll-like receptor (TLR) cascades, and signaling by interleukins (Figure 6D; Table S8).

DISCUSSION

Pneumonia is a complex disease characterized by diversity in its etiology, symptomatology, and clinical management. This complexity extends to decisions regarding the prescription of antibiotics and hospitalization for children, which remains a significant issue in terms of medical cost.[13] Some efforts have been made to identifying molecular biomarkers to assist in diagnosing and predicting pneumonia outcomes. However, there is still not an optimal test to accurately diagnose pneumonia in children or

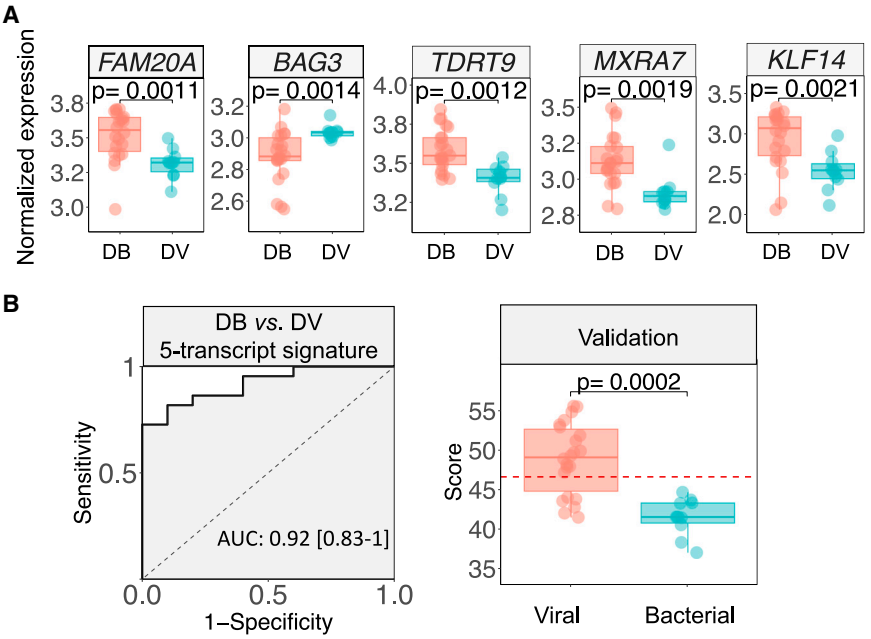


Figure 4. Validation cohort

(A) Boxplots showing the expression values of the genes included in the 5-transcript signature in the validation cohort. The boxes are defined by the upper and lower quartile (Q1 and Q3); the median is shown as a bold-colored horizontal line; whiskers extend to the most extreme data point, which is no more than 1.5 times the IQR from the box.

(B) ROC curve, AUC value with confidence interval, and boxplot of the predicted value obtained from applying the 5-transcript viral/bacterial signature coefficients in the validation cohort. Wilcoxon test *p* values are displayed in the boxplots.

to distinguish between bacterial and viral pneumonia. Consequently, antibiotics are often prescribed when a bacterial infection cannot be ruled out leading to unnecessary cost and contributes to the development of antimicrobial resistance.[9,54]

Our current analysis reveals that pneumonia in children induces extensive modifications in the host's transcriptome. Our findings indicate that pneumonia alters the expression of more than 5,000 genes in blood of affected children. Subsequent validation of these differential expression results in an independent cohort of pediatric pneumonia cases confirmed these findings. Additionally, two independent enrichment approaches (ORA and GSEA) using GO and Reactome as reference databases revealed a convergence of altered pathways within the immune system. Specifically, these pathways involve the activation and migration of leukocytes, with significant focus on neutrophil transcripts. During pneumonia, there is recruitment of neutrophils into the lungs. In cases of acute lung injury, a distinct pattern of activated neutrophils, referred to as the neutrophil activation signature,[55] emerges. This pattern is unique to the process of neutrophils migrating toward infected lung tissue, leading to various responses, including the release of bactericidal enzymes through degranulation, the discharge of cytokines and reactive oxygen species, and the generation of neutrophil extracellular traps. While these activities facilitate the elimination of causative pathogens, they may also result in tissue damage and exacerbate acute inflammation.[56–58]

Most prominent overexpression changes detected in our RNA-seq cohort, which were subsequently validated in an independent dataset, were represented by genes that produce proteins secreted by secondary granules released from neutrophils. These genes include *OLFM4*, *MMP8*, and *CD177*, and some of them have been previously reported as candidate biomarkers for the diagnosis/prognosis of CAP and predictors of sepsis.[59–61] MMP8 is a collagenase derived from neutrophils that has implications in a wide variety of pulmonary pathologies. It can degrade all components of the extracellular matrix and plays a role in neutrophil migration.[62] Its expression has already been shown to be significantly higher in pneumonia patients compared to healthy controls, representing one of the most over-expressed gene in adult pneumonia samples from various studies,[27,63] and also in our pediatric cohort. Serum and plasma levels of MMP8 protein have been proposed for diagnosing CAP, and as prognostic biomarker of fatal outcome in septic patients and predictive of serious bacterial infection, suggesting that could be an inflammatory modulator in sepsis.[64–66] However, its poor specificity limits its use as a single protein diagnostic biomarker. Our findings support the evidence that *MMP8* gene is also over-expressed in children with pneumonia. Similarly, we observed a significant overexpression of the *OLFM4* gene in pneumonia patients. *OLFM4* is a matrix glycoprotein of neutrophil-specific granules and has been described as a marker of severity in infectious diseases. In children, it has been identified as an independent risk factor for poor outcomes in sepsis,[67,68] and its inhibition has been proposed as a possible therapeutic approach in infected patients.[68] *CD177* is a glycosylphosphatidylinositol-anchored protein that plays a crucial role in regulating neutrophils by modulating their migration and activation. *CD177* has been identified as the most dysregulated molecule in purified neutrophils from patients with septic shock and severe influenza infection.[69] Recently, it has also been associated with worse outcome and higher mortality in patients infected with SARS-CoV-2[70] and included in a diagnostic four-gene signature for pediatric sepsis.[59]

ABL was the gene yielding the lowest significant value when comparing DB and DV patients. *ALB* encodes for the albumin protein, the most abundant protein in human blood, but it is primarily expressed in the liver and pancreas (https://www.gtexportal.org/). Although the overall expression of ALB in our blood samples was low (median $\log_2$ values of 2.07, 5, and 3.09 for DB, DV, and healthy controls, respectively), it showed a significant overexpression in DV compared to DB patients. Some studies have showed that decreased levels of albumin in serum and plasma may indicate a higher risk of ICU admission and death in sepsis,[71,72] serve as a diagnostic biomarker for infectious diseases,[73] and act as a risk

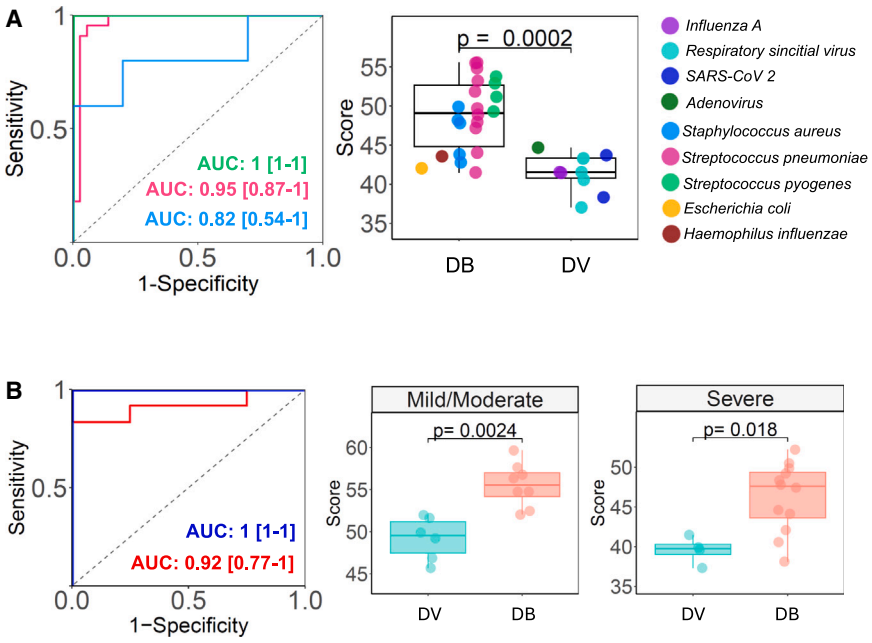


Figure 5. Performance of the 5-transcript signature

(A) ROC curves, AUC values, and boxplots of the performance of the 5-transcript signature in pneumonias with different causal pathogens. Only bacterial groups with more than 3 samples were considered for the AUC and ROC analysis. The boxes are defined by the upper and lower quartile (Q1 and Q3); the median is shown as a bold-colored horizontal line; whiskers extend to the most extreme data point, which is no more than 1.5 times the IQR from the box.

(B) ROC curves, AUC values, and boxplots of the performance of the 5-transcript signature in pneumonias with different severities (severe or mild/moderate).

factor for more severe clinical manifestations in COVID-19 patients.[74,75]

In addition to the neutrophil-related innate response, our analysis identified a common enrichment of terms related to humoral immunity processes, consistent with previous results from pediatric pneumonia cases,[76] as well as an enrichment of innate processes commonly activated in response to viral and bacterial infections, such as the cytokines release and inflammation. At the same time, however, the over-activation of these processes can also be responsible for the clinical consequences of severe pneumonia, increasing the risk of severe sepsis and death.[77,78]

Since pneumonia can be caused by a broad range of viral or bacterial pathogens, it is essential to investigate new approaches to differentiate the pathogenic origin of the disease, which may support decisions regarding antibiotic prescriptions. Molecular methods for detecting causative bacteria remain less effective in pneumonia patients, mainly due to challenges in obtaining samples from the infection site and the frequent absence of positive results from accessible sites such as blood.[79] Moreover, detecting viruses in nasopharyngeal secretions from pneumonia patients is common due to colonization, but this does not exclude the possibility of a concurrent bacterial infection.[80,81] A comparison of gene expression between DV vs. DB pneumonia revealed 282 DEGs between both categories. Functional studies of the co-expression modules driving the host-specific responses to viral or bacterial pneumonia highlighted differentially activated immunological pathways in both pneumonia etiologies. We identified two significant modules associated with viral pneumonia (EVL and TGBR3) and one associated with bacterial pneumonia (MAPK14). Genes from the MAPK14 module are primarily involved in innate responses related to cytokine production, acute inflammation, neutrophil degranulation, deficiencies in TLR cascade, and interleukin signaling. TLR receptors are critical in constraining the proliferation and dissemination of invading bacteria. Activation of TLRs triggers innate immune signaling pathways and transcriptional upregulation of pro-inflammatory mediators that recruit neutrophils to pulmonary tissue during bacterial infection.[82] An effective and timely inflammatory response is key to pathogen clearance and control the infection in pneumonia. However, an exacerbated and prolonged release of pro- and anti-inflammatory cytokines, followed by neutrophil hyper-responsiveness, is associated with an increased risk of severe sepsis and poor outcomes in CAP.[77,83] TLRs can detect various conserved patterns from both viral and bacterial pathogens. Interestingly, significant pathways involving TLRs detected in the bacterial module are among the TLR cascades commonly activated by bacterial pathogen-associated molecular patterns (PAMPs; e.g., lipoproteins TLR1/TLR2/TLR6 and lipopolysaccharide TLR4).[84] Most TLRs, except TLR3, signal via myeloid differentiation primary response 88 (MyD88) and interleukin-1 receptor-associated kinase 4 (IRAK-4). IRAK-4 is critical for MyD88-dependent TLR signaling, and patients with IRAK-4 mutations are susceptible to recurrent bacterial infections.[85,86] Patients with both IRAK-4 and MyD88 deficiencies are predisposed to severe bacterial infection,[87] and defects in MyD88-mediated TLR signaling make children much more susceptible to pneumonia.[88] The canonical TLR pathway can activate NF-κB and MAPKs through different mediators; in this regard, the hub gene of our bacterial module, *MAPK14*, is a key gene in the development of ventilator-associated bacterial pneumonia in adults.[89]

Collectively, these findings suggest that immune responses to severe bacterial pneumonia are primarily characterized by dysfunction in the innate TLR signaling. This dysfunction leads to an exacerbated inflammatory response, driven by cytokine release and neutrophil degranulation, which may play a pivotal role in the development of the severe clinical symptoms observed in these patients.

In the case of viral pneumonia-correlated modules (EVL and TGBR3), we identified a significant representation of adaptative immunity mediated by T cells. Genes within the EVL module were associated with processes such as T cell selection; T cell receptor signaling; and T cell activation, differentiation, and

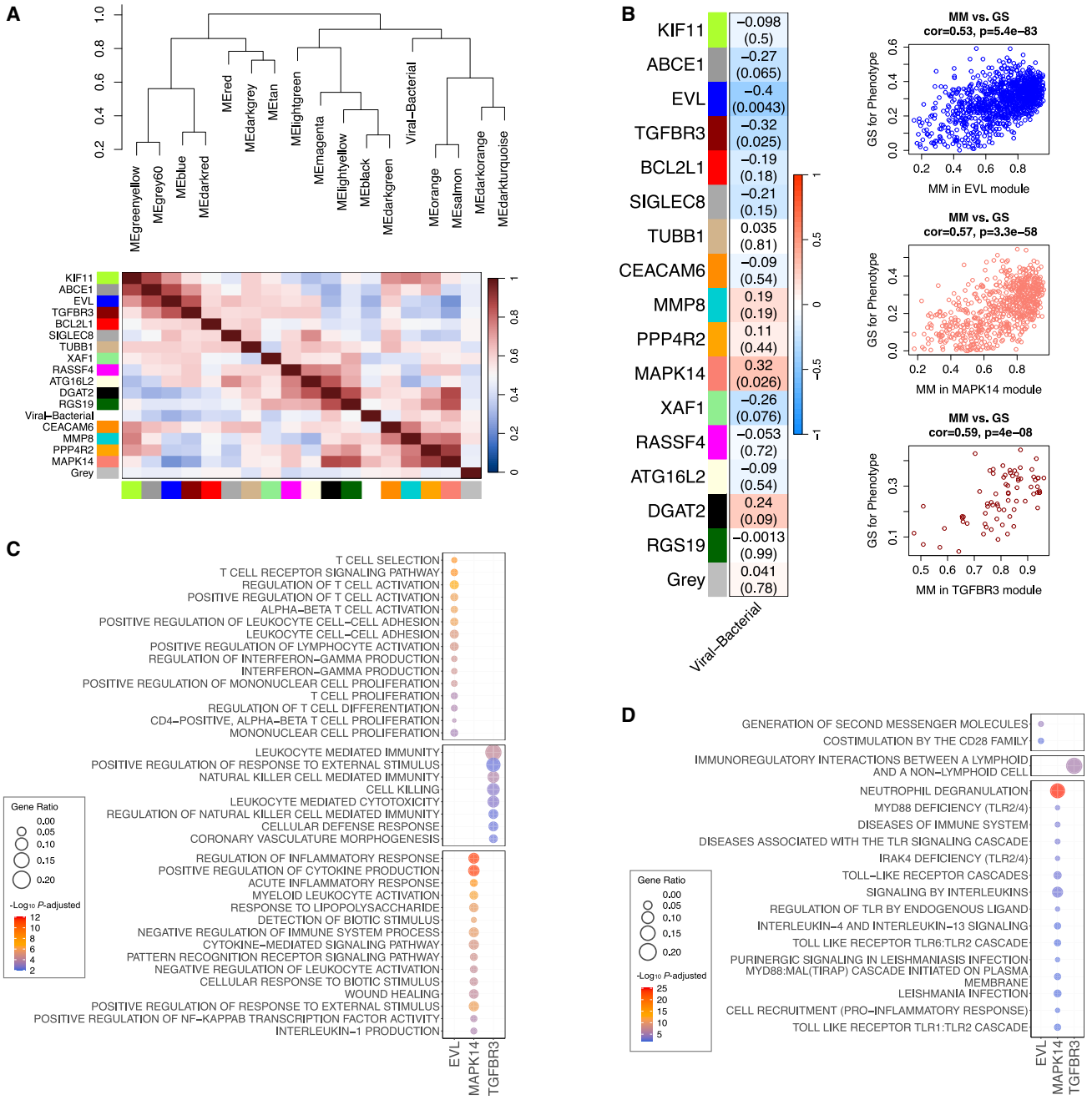


Figure 6. Co-expression analysis of viral vs. bacterial pneumonia

(A) Hierarchical clustering eigengene dendrogram and heatmap showing relationships among the modules and pneumonia etiology (DB and DV pneumonias). Gene names on the left of the heatmap are the hub genes of each module.

(B) Correlation values heatmap between co-expression modules and DB and DV conditions (DV was considered the reference group). Values show individual correlations of the module with DB phenotype (left). Correlation values are also indicated with different colors (legend gradient color bar). *p* values of these correlations are represented in brackets (lower values). Plots showing comparison between MM (module membership) and GS (correlation with bacterial/viral condition) of genes from the most significant modules detected.

(C) Top 15 statistically associated GO biological processes detected for each of the significantly correlated modules.

(D) Top 15 statistically associated Reactome pathways detected for each of the significantly correlated modules.

USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

proliferation. The TGFBR3 module also reveals pathways related to adaptative immunity, such as "leukocyte mediated cytotoxicity" and "T cell mediated immunity." T cells are essential in combating viral infections, by inducing cell death through cytokines release.[90] Notably, both $CD4^+$ and $CD8^+$ T cells increase in population during the first days to weeks of an acute viral infection.[91] $CD8^+$ cells have been described to be critical in clearing various viral infections in mouse models,[92] while $CD4^+$ T cells serve as helpers for B cells and $CD8^+$ T cells, also possessing independent antiviral function through cytokine secretion.[93] Of interest, polymorphisms in the hub gene *TGFBR3* were previously associated with a history of pneumonia in sickle cell disease patients.[94] *EVL* expression was associated with poor prognosis in sepsis.[95]

We have identified a 5-transcript blood signature that can accurately differentiate bacterial from viral pneumonias with an AUC of 0.95 (sensitivity = 0.80; specificity = 1.00), demonstrating superior discriminant power between viral and bacterial pneumonia compared with previously published non-syndrome-specific signatures for differentiating viral from bacterial infections in children (AUC: 0.87 [0.75–0.96] and 0.88 [0.78–0.98] in the re-trained model).[30] However, as the general signature also performs effectively in this cohort, both methods could be implemented complementarily in a clinical context for the definitive confirmation of the diagnosis. The 5-transcript signature's performance remains high when including pneumonia samples of both PB and PV origin (AUC = 0.87; sensitivity = 0.76; specificity = 0.83), affirming the robustness and discriminatory power of the transcriptomic signature. Among the genes selected in this signature, some are associated with antiviral or bacterial activity. For instance, *BAG3* is a cochaperone recently linked to intracellular bacterial proliferation[96]. *FAM20A*, a pseudokinase, exhibits increased expression in the lung and liver,[97] in sepsis,[98] and in blood neutrophils of acute respiratory distress syndrome patients.[55] The *TDRD9* gene was previously described in a transcriptomic signature for sepsis derived from CAP.[99] *MXRA7* is highly expressed in murine and human ocular tissues and may play a role in pathological processes involving injury, neovascularization, and wound healing, although its function remains mainly unknown.[100] Finally, the Krüppel-like transcription factor *KLF14* exhibits a modulatory role in macrophages-driven inflammation[101] and has been found to be upregulated in septic patients.[102] Its potential regulatory function in sepsis positions this gene as a therapeutic target candidate for sepsis.

We have also observed that the 5-transcript signature identified in the present study and the signature identified by Herberg et al.[30] for differentiating viral from bacterial infection did not prove effective in the pneumonia cohort from Wallihan et al..[53] This outcome can be explained by the fact that the vast majority of patients with bacterial pneumonia in the Wallihan et al. cohort have infections caused by *Mycoplasma pneumoniae* (30/35; 86%). These results suggest that the host gene expression response to this pathogen is specific and distinct from that triggered by other pneumonia-causing bacteria. The fact that *M. pneumoniae* is generating worrying outbreaks in the post-COVID-19 pandemic era[103,104] deserves further dedicated investigation.

The validation of the transcriptomic signature in an independent cohort of children with pneumonia further reinforces its high specificity and sensitivity in accurately distinguishing bacterial from viral infections. Importantly, our analysis confirmed that the effectiveness of the signature is independent of both the causal bacterial pathogen responsible and the severity of the disease. This robust performance across diverse clinical contexts underscores the potential of this signature as a reliable diagnostic tool for pediatric pneumonia, offering a significant advancement in precision and reliability over current diagnostic methods.

Our study involved hospitals from different countries, reducing the potential bias from specific populations or areas. It has characterized and independently validated transcriptomic changes caused by CAP in hospitalized children.

We have uncovered various immune mechanisms and genes that are distinctly activated in viral and bacterial pneumonia, making a significant advance in the field. Most notably, our 5-transcript signature stands out for its ability to accurately differentiate between bacterial and viral pneumonia, even in cases of uncertain origin. Crucially, this signature might make such distinctions before the full clinical symptoms emerge. This early diagnostic capability holds the potential to drastically reduce the time to a definitive diagnosis compared to traditional culture-based methods, ultimately shortening hospital stays, and mitigate the consequences of delayed or missed diagnoses, such as unnecessary hospital investigations and antibiotic prescriptions when a bacterial infection is suspected. These benefits position our findings as a major step forward in improving pneumonia management.

Limitations of the study

The primary limitation of our study is the limited number of samples from children with viral pneumonia. Therefore, further validation is required, utilizing additional datasets that encompass well-defined pneumonia infections of both viral and bacterial origins. Despite this limitation, our results reaffirm the classification of pediatric patients into viral and bacterial pneumonia categories based on the clinical algorithm developed by Herberg et al..[30]

CONSORTIA

GENDRES network

Miriam Cebey López, Antonio Salas Ellacuriaga, Ana Vega Gliemmo, José Peña Guitián, Alexa Regueiro, Antonio Justicia Grande, Leticia Pías Peleteiro, María López Sousa, María Jose de Castro, Carmen Curros Novo, Elena Rodrigo, Miriam Puente Puig, Rosaura Leis Trabazo, Nazareth Martinón Torres, Alberto Gómez Carballa, Jacobo Pardo Seco, Sara Pischedda, José María Martinón Sánchez, Belén Mosquera Pérez, Isabel Villanueva González, Lorenzo Redondo Collazo, Carmen Rodríguez-Tenreiro, María del Sol Porto Silva and Federico Martinón Torres, Máximo Francisco Fraga Rodríguez, Orlando Fernández Lago, José Ramón Antúnez, Enrique Bernaola Iturbe, Laura Moreno Galarraga, Jorge Álvarez, Mercedes Herranz, Francisco Gil, Eva Gembero, Jorge Rodríguez, Teresa González López, Delfina Suarez Vázquez, Ángela Vázquez Vázquez, Susana Rey García, Nathalie Carreira Sande, Ana López Fernández, Nuria Romero Pérez, José Antonio Couceiro Gianzo, Nazareth Fuentes Perez, Francisco Giménez Sánchez, Miguel Sánchez Forte, Cristina Calvo Rey, María Luz García García, Iciar Olabarrieta Arnal,

Adelaida Fernández Rincón, Ignacio Oulego Erroz, David Naranjo Vivas, Santiago Lapeña, Paula Alonso Quintela, Jorge Martínez Sáenz de Jubera, Estibaliz Garrido García, Cristina Calvo Monge, Eider Oñate Vergara, Jesús de la Cruz Moreno, Mª Carmen Martínez Padilla, Beatriz Jiménez Jurado, Carmen Santiago Gutiérrez, María Esther Vidaurreta del Castillo, Manuel Baca Cots, David Moreno Pérez, Ana Cordón Martínez, Antonio Urda Cardona, José Miguel Ramos Fernández, Esmeralda Núñez Cuadros, Susana Beatriz Reyes, María Cruz León León, Santiago Alfayate, Cristina Calvo, Carlos Grasa, Cristian Quintana Ortega, Leticia La Banda Montalvo, María Lopez Cerdán, Ana Dominguez Castells, Francisco Giménez Sánchez, Andrés J. Alcaraz Romero, Diego Bautista Lozano, Sara Uillen Martin, Roi Piñeiro, Juan Ignacio Sánchez Díaz, Alba Palacios Cuesta, Elvira González Salas, Sira Fernández De Miguel, Belen Joyanes Abancens, Esther Aleo Lujan, Alfredo Tagarro García, María Luisa Herreros, Rut del Valle, Libertad Latorre Navarro, María Concepción Zazo Sanchidrián, Mariano Esteban, Marta González Lorenzo, Mª Carmen Vicent Castello, Lorena Moreno Requena, Juan Luis Santos Pérez, César Gavilán Martín, Lucía González-Moro Azorín, Monterrat López Franco, Manuel Silveira Cancela, María José Cilleruelo Ortega, Luz Golmayo, Francisca Portero Azorín, Andrés Concha Torre, Lucía Rodríguez García, Carlos Rodrigo Gonzalo de Liria, Andrés Antón Pagarolas, Maria Méndez, Cristina Prat, Jesús López-Herce, Miriam García Samprudencio, Gema Manrique Martín, Paula García Casas, Débora Sanz Álvarez, María Jesús Cabero, Miguel Lillo Lillo, Marta Pareja, Pablo Rojo, Cristina Epalza (Hospital 12 Octubre); Adriana Navas Carretero, Estefanía Barral, Miriam Herrera, Elvira Cobo Vázquez, Elena del Castillo Navío, Patricia Flores Perez, Paula García Casas, Esteban Gómez Sánchez, Juan Valencia Ramos, Francisco Javier Pilar Orive, Elva Rodríguez Merino, Ana Pérez Aragón, Mª Yolanda Ruiz del Prado, David Moreno, Beatriz Carazo, Jordi Antón, and María teresa Rives Ferreiro. Further details may be consulted at http://www.gendres.org, and www.genvip.eu.

EUCLIDS consortium

Michael Levin, Lachlan Coin, Stuart Gormley, Shea Hamilton, Jethro Herberg, Bernardo Hourmat, Clive Hoggart, Myrsini Kaforou, Vanessa Sancho-Shimizu, Victoria Wright, Amina Abdulla, Paul Agapow, Maeve Bartlett, Evangelos Bellos, Hariklia Eleftherohorinou, Rachel Galassini, David Inwald, Meg Mashbat, Stefanie Menikou, Sobia Mustafa, Simon Nadel, Rahmeen Rahman, Clare Thakker, Sumit Bokhandi, Sue Power, Heather Barham, Nazima Pathan, Jenna Ridout, Deborah White, Sarah Thurston, Saul Faust, Sanjay Patel, Jenni McCorkell, Patrick Davies, Lindsey Crate, Helen Navarra, Stephanie Carter, Ramesh Ramaiah, Rekha Patel, Catherine Tuffrey, Andrew Gribbin, Sharon McCready, Mark Peters, Katie Hardy, Fran Standing, Lauran O'Neill, Eugenia Abelake, Akash Deep, Eniola Nsirim, Andrew Pollard, Louise Willis, Zoe Young, C. Royad, Sonia White, PM. Fortune, Phil Hudnott, Federico Martinón Torres, Antonio Salas Ellacuriaga, Fernando Álvez González, Ruth Barral-Arca, Miriam Cebey-López, María José Curras-Tuala, Natalia García, Luisa García Vicente, Alberto Gómez-Carballa, Jose Gómez Rial, Andrea Grela Beiroa, Antonio Justicia Grande, Pilar Leboráns Iglesias, Alba Elena Martínez Santos, Nazareth Martinón-Torres, José María Martinón Sánchez, Beatriz Morillo Gutiérrez, Belén Mosquera Pérez, Pablo Obando Pacheco, Jacobo Pardo-Seco, Sara Pischedda, Irene Rivero Calle, Carmen Rodríguez-Tenreiro, Lorenzo Redondo-Collazo, Sonia Serén Fernández, María del Sol Porto Silva, Ana Vega, Jose Manuel Fernández García, María Elena Gamborino Caramés, María Sol Rodríguez Calvo, Marta Aldonza Torres, Vanesa Álvarez Iglesias, Carmen Curros Novo, Isabel Ferreirós Vidal, Narmeen Mallah, Laura Navarro Ramón, Isabel Rego Lijo, Alba Camino Mera, Lúa Castelo Martínez, Ana Isabel Dacosta Urbieta, Wiktor Dominik Nowak, Julia García Currás, Nour El Zaharaa Mallah, Julián Montoto Louzao, Sara Rey Vázquez, Sandra Viz Lasheras, Patricia Regueiro Casuso, Miriam Taboada Puga, Lucía Vilanova Trillo, Susana Beatriz Reyes, María Cruz León León, Álvaro Navarro Mingorance, Xavier Gabaldó Barrios, Eider Oñate Vergara, Andrés Concha Torre, Ana Vivanco, Reyes Fernández, Francisco Giménez Sánchez, Miguel Sánchez Forte, Pablo Rojo, J. Ruiz Contreras, Alba Palacios, Cristina Epalza Ibarrondo, Elizabeth Fernandez Cooke, Marisa Navarro, Cristina Álvarez Álvarez, María José Lozano, Eduardo Carreras, Sonia Brió Sanagustín, Olaf Neth, Mª del Carmen Martínez Padilla, Luis Manuel Prieto Tato, Sara Guillén, Laura Fernández Silveira, David Moreno, Ronald de Groot, A. Marceline Tutu van Furth, Michiel van der Flier, Navin P. Boeddha, Gertjan J. A. Driessen, Jan A. Hazelzet, Taco W. Kuijpers, Dasja Pajkrt, Elisabeth A. M. Sanders, Diederik van de Beek, A. van der Ende, H.L.A. Philipsen, A.O.A. Adeel, M.A. Breukels, D.M.C. Brinkman, C.C.M.M. de Korte, E. de Vries, W.J. de Waal, R. Dekkers, A. Dings-Lammertink, R.A. Doedens, A.E. Donker, M. Dousma, T.E. Faber, G.P.J.M. Gerrits, J.A.M. Gerver, J. Heidema, J. Homan-van der Veen, M.A.M. Jacobs, N.J.G. Jansen, P. Kawczynski, K. Klucovska, M.C.J. Kneyber, Y. Koopman-Keemink, V.J. Langenhorst, J. Leusink, B.F. Loza, I.T. Merth, C.J. Miedema, C. Neeleman, J.G. Noordzij, C.C. Obihara, A.L.T. van Overbeek – van Gils, G.H. Poortman, S.T. Potgieter, J. Potjewijd, P.P.R. Rosias, T. Sprong, G.W. ten Tussher, B.J. Thio, G.A. Tramper-Stranders, M. van Deuren, H. van der Meer, A.J.M. van Kuppevelt, A.M. van Wermeskerken, W.A. Verwijs, T.F.W. Wolfs, Luregn J Schlapbach, Philipp Agyeman, Christoph Aebi, Christoph Berger, Eric Giannoni, Martin Stocker, Klara M Posfay-Barbe, Ulrich Heininger, Sara Bernhard-Stirnemann, Anita Niederer-Loher, Christian Kahlert, Paul Hasters, Christa Relly, Walter Baer, Enitan Carrol, Stéphane Paulus, Hannah Frederick, Rebecca Jennings, Joanne Johnston, Rhian Kenwright, Colin G Fink, Elli Pinnock, Marieke Emonts, Rachel Agbeko, Suzanne Anderson, Fatou Secka, Kalifa Bojang, Isatou Sarr, Ngane Kebbeh, Gibbi Sey, Momodou Saidykhan, Fatoumatta Cole, Gilleh Thomas, Martin Antonio, Werner Zenz, Daniela S. Klobassa, Alexander Binder, Nina A. Schweintzger, Manfred Sagmeister, Hinrich Baumgart, Markus Baumgartner, Uta Behrends, Ariane Biebl, Robert Birnbacher, Jan-Gerd Blanke, Carsten Boelke, Kai Breuling, Jürgen Brunner, Maria Buller, Peter Dahlem, Beate Dietrich, Ernst Eber, Johannes Elias, Josef Emhofer, Rosa Etschmaier, Sebastian Farr, Ylenia Girtler, Irina Grigorow, Konrad Heimann, Ulrike Ihm, Zdenek Jaros, Hermann Kalhoff, Wilhelm Kaulfersch, Christoph Kemen, Nina Klocker, Bernhard Köster, Benno Kohlmaier, Eleni Komini, Lydia Kramer, Antje Neubert, Daniel Ortner, Lydia Pescollderungg, Klaus Pfurtscheller, Karl Reiter, Goran Ristic, Siegfried Rödl,

Andrea Sellner, Astrid Sonnleitner, Matthias Sperl, Wolfgang Stelzl, Holger Till, Andreas Trobisch, Anne Vierzig, Ulrich Vogel, Christina Weingarten, Stefanie Welke, Andreas Wimmer, Uwe Wintergerst, Daniel Wüller, Andrew Zaunschirm, Ieva Ziuraite, and Veslava Žukovskaja.

DIAMONDS consortium

Michael Levin, Lachlan Coin, Stuart Gormley, Shea Hamilton, Jethro Herberg, Bernardo Hourmat, Clive Hoggart, Myrsini Kaforou, Vanessa Sancho-Shimizu, Victoria Wright, Amina Abdulla, Paul Agapow, Maeve Bartlett, Evangelos Bellos, Hariklia Eleftherohorinou, Rachel Galassini, David Inwald, Meg Mashbat, Stefanie Menikou, Sobia Mustafa, Simon Nadel, Rahmeen Rahman, Clare Thakker, Sumit Bokhandi, Sue Power, Heather Barham, Nazima Pathan, Jenna Ridout, Deborah White, Sarah Thurston, Saul Faust, Sanjay Patel, Jenni McCorkell, Patrick Davies, Lindsey Crate, Helen Navarra, Stephanie Carter, Ramesh Ramaiah, Rekha Patel, Catherine Tuffrey, Andrew Gribbin, Sharon McCready, Mark Peters, Katie Hardy, Fran Standing, Lauran O'Neill, Eugenia Abelake, Akash Deep, Eniola Nsirim, Andrew Pollard, Louise Willis, Zoe Young, C. Royad, Sonia White, PM. Fortune, Phil Hudnott, Federico Martinón Torres, Fernando Álvez González, Ruth Barral-Arca, Miriam Cebey-López, María José Curras-Tuala, Natalia García, Luisa García Vicente, Alberto Gómez-Carballa, Jose Gómez Rial, Andrea Grela Beiroa, Antonio Justicia Grande, Pilar Leboráns Iglesias, Alba Elena Martínez Santos, Nazareth Martinón-Torres, José María Martinón Sánchez, Beatriz Morillo Gutiérrez, Belén Mosquera Pérez, Pablo Obando Pacheco, Jacobo Pardo-Seco, Sara Pischedda, Irene Rivero Calle, Carmen Rodríguez-Tenreiro, Lorenzo Redondo-Collazo, Sonia Serén Fernández, María del Sol Porto Silva, Ana Vega, Jose Manuel Fernández García, María Elena Gamborino Caramés, María Sol Rodríguez Calvo, Marta Aldonza Torres, Vanesa Álvarez Iglesias, Carmen Curros Novo, Isabel Ferreirós Vidal, Narmeen Mallah, Laura Navarro Ramón, Isabel Rego Lijo, Alba Camino Mera, Lúa Castelo Martínez, Ana Isabel Dacosta Urbieta, Wiktor Dominik Nowak, Julia García Currás, Nour El Zaharaa Mallah, Julián Montoto Louzao, Sara Rey Vázquez, Sandra Viz Lasheras, Patricia Regueiro Casuso, Miriam Taboada Puga, Lucía Vilanova Trillo, Susana Beatriz Reyes, María Cruz León León, Álvaro Navarro Mingorance, Xavier Gabaldó Barrios, Eider Oñate Vergara, Andrés Concha Torre, Ana Vivanco, Reyes Fernández, Francisco Giménez Sánchez, Miguel Sánchez Forte, Pablo Rojo, J. Ruiz Contreras, Alba Palacios, Cristina Epalza Ibarrondo, Elizabeth Fernandez Cooke, Marisa Navarro, Cristina Álvarez Álvarez, María José Lozano, Eduardo Carreras, Sonia Brió Sanagustín, Olaf Neth, M[a] del Carmen Martínez Padilla, Luis Manuel Prieto Tato, Sara Guillén, Laura Fernández Silveira, David Moreno, Ronald de Groot, A. Marceline Tutu van Furth, Michiel van der Flier, Navin P. Boeddha, Gertjan J. A. Driessen, Jan A. Hazelzet, Taco W. Kuijpers, Dasja Pajkrt, Elisabeth A. M. Sanders, Diederik van de Beek, A. van der Ende, H.L.A. Philipsen, A.O.A. Adeel, M.A. Breukels, D.M.C. Brinkman, C.C.M.M. de Korte, E. de Vries, W.J. de Waal, R. Dekkers, A. Dings-Lammertink, R.A. Doedens, A.E. Donker, M. Dousma, T.E. Faber, G.P.J.M. Gerrits, J.A.M. Gerver, J. Heidema, J. Homan-van der Veen, M.A.M. Jacobs, N.J.G. Jansen, P. Kawczynski, K. Klucovska, M.C.J. Kneyber, Y. Koopman-Keemink, V.J. Langenhorst, J. Leusink, B.F. Loza, I.T. Merth, C.J. Miedema, C. Neeleman, J.G. Noordzij, C.C. Obihara, A.L.T. van Overbeek – van Gils, G.H. Poortman, S.T. Potgieter, J. Potjewijd, P.P.R. Rosias, T. Sprong, G.W. ten Tussher, B.J. Thio, G.A. Tramper-Stranders, M. van Deuren, H. van der Meer, A.J.M. van Kuppevelt, A.M. van Wermeskerken, W.A. Verwijs, T.F.W. Wolfs, Luregn J Schlapbach, Philipp Agyeman, Christoph Aebi, Christoph Berger, Eric Giannoni, Martin Stocker, Klara M Posfay-Barbe, Ulrich Heininger, Sara Bernhard-Stirnemann, Anita Niederer-Loher, Christian Kahlert, Paul Hasters, Christa Relly, Walter Baer, Enitan Carrol, Stéphane Paulus, Hannah Frederick, Rebecca Jennings, Joanne Johnston, Rhian Kenwright, Colin G Fink, Elli Pinnock, Marieke Emonts, Rachel Agbeko, Suzanne Anderson, Fatou Secka, Kalifa Bojang, Isatou Sarr, Ngane Kebbeh, Gibbi Sey, Momodou Saidykhan, Fatoumatta Cole, Gilleh Thomas, Martin Antonio, Werner Zenz, Daniela S. Klobassa, Alexander Binder, Nina A. Schweintzger, Manfred Sagmeister, Hinrich Baumgart, Markus Baumgartner, Uta Behrends, Ariane Biebl, Robert Birnbacher, Jan-Gerd Blanke, Carsten Boelke, Kai Breuling, Jürgen Brunner, Maria Buller, Peter Dahlem, Beate Dietrich, Ernst Eber, Johannes Elias, Josef Emhofer, Rosa Etschmaier, Sebastian Farr, Ylenia Girtler, Irina Grigorow, Konrad Heimann, Ulrike Ihm, Zdenek Jaros, Hermann Kalhoff, Wilhelm Kaulfersch, Christoph Kemen, Nina Klocker, Bernhard Köster, Benno Kohlmaier, Eleni Komini, Lydia Kramer, Antje Neubert, Daniel Ortner, Lydia Pescollderungg, Klaus Pfurtscheller, Karl Reiter, Goran Ristic, Siegfried Rödl, Andrea Sellner, Astrid Sonnleitner, Matthias Sperl, Wolfgang Stelzl, Holger Till, Andreas Trobisch, Anne Vierzig, Ulrich Vogel, Christina Weingarten, Stefanie Welke, Andreas Wimmer, Uwe Wintergerst, Daniel Wüller, Andrew Zaunschirm, Ieva Ziuraite, and Veslava Žukovskaja.

PERFORM consortium

Michael Levin, Aubrey Cunnington, Tisham De, Jethro Herberg, Myrsini Kaforou, Victoria Wright, Lucas Baumard, Evangelos Bellos, Giselle D'Souza, Rachel Galassini, Dominic Habgood-Coote, Shea Hamilton, Clive Hoggart, Sara Hourmat, Heather Jackson, Ian Maconochie, Stephanie Menikou, Naomi Lin, Samuel Nichols, Ruud Nijman, Ivonne Pena Paz, Oliver Powell, Priyen Shah, Ching-Fen Shen, Clare Wilson, Amina Abdulla, Ladan Ali, Sarah Darnell, Rikke Jorgensen, Sobia Mustafa, Salina Persand, Molly Stevens, Eunjung Kim, Benjamin Pierce, Katy Fidler, Julia Dudley, Vivien Richmond, Emma Tavliavini, Ching-Chuan Liu, Shih-Min Wang, Federico Martinón-Torres, Antonio Salas, Fernando Álvez González, Cristina Balo Farto, Ruth Barral-Arca, María Barreiro Castro, Xabier Bello, Mirian Ben García, Sandra Carnota, Miriam Cebey-López, María José Curras-Tuala, Carlos Durán Suárez, Luisa García Vicente, Alberto Gómez-Carballa, Jose Gómez Rial, Pilar Leboráns Iglesias, Nazareth Martinón-Torres, José María Martinón Sánchez, Belén Mosquera Pérez, Jacobo Pardo-Seco, Lidia Piñeiro Rodríguez, Sara Pischedda, Sara Rey Vázquez, Irene Rivero Calle, Carmen Rodríguez-Tenreiro, Lorenzo Redondo-Collazo, Miguel Sadiki Ora, Sonia Serén Fernández, Cristina Serén Trasorras, Marisol Vilas Iglesias, Dace Zavadska, Anda Balode, Arta Bārzdiņa, Dārta Deksne, Dace Gardovska, Dagne Grāvele, Ilze Grope, Anija Meiere, Ieva Nokalna, Jana Pavāre, Zanda Pučuka, Katrīna Selecka, Aleksandra Sidorova, Dace Svile, Urzula Nora Urbāne,

Effua Usuf, Kalifa Bojang, Syed M. A. Zaman, Fatou Secka, Suzanne Anderson, Anna Rocalsatou Sarr, Momodou Saidykhan, Saffiatou Darboe, Samba Ceesay, Umberto D'alessandro, Henriëtte A. Moll, Dorine M. Borensztajn, Nienke N. Hagedoorn, Chantal Tan, Clementien L. Vermont, Joany Zachariasse, W Dik, Philipp Agyeman, Luregn J Schlapbach, Eric Giannoni, Martin Stocker, Klara M Posfay-Barbe, Ulrich Heininger, Sara Bernhard-Stirnemann, Anita Niederer-Loher, Christian Kahlert, Giancarlo Natalucci, Christa Relly, Thomas Riedel, Christoph Aebi, Christoph Berger, Enitan D Carrol, Stéphane Paulus, Elizabeth Cocklin, Rebecca Jennings, Joanne Johnston, Simon Leigh, Karen Newall, Sam Romaine, Andrew J. Pollard, Rama Kandasamy, Michael J. Carter, Daniel O'Connor, Sagida Bibi, Dominic F. Kelly, Meeru Gurung, Stephen Thorson, Imran Ansari, David R. Murdoch, Shrijana Shrestha, Marieke Emonts, Emma Lim, Lucille Valentine, Karen Allen, Kathryn Bell, Adora Chan, Stephen Crulley, Kirsty Devine, Daniel Fabian, Sharon King, Paul McAlinden, Sam McDonald, Anne McDonnell, Ailsa Pickering, Evelyn Thomson, Amanda Wood, Diane Wallia, Frances Baxter, Ashley Bell, Mathew Rhodes, Rachel Agbeko, Christine Mackerness, Bryan Baas, Lieke Kloosterhuis, Wilma Oosthoek, Tasnim Arif, Joshua Bennet, Kalvin Collings, Ilona van der Giessen, Alex Martin, Aqeela Rashid, Emily Rowlands, Gabriella de Vries, Fabian van der Velden, Mike Martin, Ravi Mistry, Ulrich von Both, Laura Kolberg, Manuela Zwerenz, Judith Buschbeck, Christoph Bidlingmaier, Vera Binder, Katharina Danhauser, Nikolaus Haas, Matthias Griese, Tobias Feuchtinger, Julia Keil, Matthias Kappler, Eberhard Lurz, Georg Muench, Karl Reiter, Carola Schoen, François Mallet, Karen Brengel-Pesce, Alexandre Pachot, Marine Mommert, Marko Pokorn, Mojca Kolnik, Katarina Vincek, Tina Plankar Srovin, Natalija Bahovec, Petra Prunk, Veronika Osterman, Tanja Avramoska, Taco Kuijpers, Ilse Jongerius, J. Merlijn van den Berg, Dieneke Schonenberg, Anouk M. Barendregt, Dasja Pajkrt, Martijn van der Kuip, A. Marceline Tutu van Furth, Evelien Sprenkeler, Judith Zandstra, Guido van Mierlo, Jan Geissler, Sebastián J. Vastert, Benjamin Evans, Jake Stevens, Peter Matthews, Kyle Billing, Werner Zenz, Alexander Binder, Benno Kohlmaier, Daniela S. Kohlfürst, Nina A. Schweintzger, Christoph Zurl, Susanne Hösele, Manuel Leitner, Lena Pölz, Glorija Rajic, Bianca Stoiser, Martina Strempfl, Manfred G. Sagmeister, Sebastian Bauchinger, Martin Benesch, Astrid Ceolotto, Ernst Eber, Siegfried Gallistl, Harald Haidl, Almuthe Hauer, Christa Hude, Andreas Kapper, Markus Keldorfer, Sabine Löffler, Tobias Niedrist, Heidemarie Pilch, Andreas Pfleger, Klaus Pfurtscheller, Siegfried Rödl, Andrea Skrabl-Baumgartner, Volker Strenger, Elmar Wallner, Dennie Tempel, Danielle van Keulen, Annelieke M Strijbosch, Maike K. Tauchert, Patricia Schmied, Irene Alba-Alejandre, Florian Hoffmann, Sabrina Juranek, Esther Maier, Karl Reiter, Sebastian Schroepf, Manuel Dewez, David Bath, Elizabeth Fitchett, and Fiona Cresswell.

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Antonio Salas (antonio.salas@usc.es).

Materials availability

This study did not generate new unique reagents, and all materials utilized for the study are described in the STAR Methods section.

Data and code availability

- RNA-seq data supporting the findings of the present study are publicly available at Gene Expression Omnibus (GEO) under accession number GSE261482 and EBI ArrayExpress under accession number E-MTAB-11671,[105] E-MTAB-12793,[106] and E-MTAB-14564. Additional details are provided in Table S7.
- Original R code has been deposited at GitHub (https://github.com/Sandraviz/Pneumonia_Transcriptomic-Analysis).
- Any additional information required to analyze the data reported in this paper is available upon request.

ACKNOWLEDGMENTS

The authors would like to express their appreciation to the study investigators of GENDRES network (www.gendres.org) (Annex), as well as the nursery and laboratory service at the Hospital Clínico Universitario de Santiago de Compostela, for their invaluable dedication and support. This research project was made possible through the access granted by the Galician Supercomputing Center (CESGA) to its supercomputing infrastructure. The supercomputer FinisTerrae III and its permanent data storage system have been funded by the Spanish Ministry of Science and Innovation, the Galician Government, and the European Regional Development Fund (ERDF). This work was supported by the European Seventh Framework Programme for Research and Technological Development (FP7) under the EUCLIDS project (grant agreement number 279185) and the European Union's Horizon 2020 research and innovation program under grant agreement nos. 668303 (PERFORM) and 848196 (DIAMONDS). This study also received support by (1) ISCIII: TRINEO: PI22/00162, DIAVIR: DTS19/00049, Resvi-Omics: PI19/01039 (to A.S.), ReSVinext: PI16/01569, Enterogen: PI19/01090, OMI-COVI-VAC: PI22/00406 (to F.M.-T.), and cofinanciados FEDER; (2) GAIN: IN607B 2020/08 and IN607A 2023/02 (to A.S.) and GEN-COVID (IN845D 2020/23 [to F.M.-T.], IIN607A2021/05 [to F.M.-T.], and IN677D 2024/06 [to A.G.-C.]; (3) ACIS: BI-BACVIR (PRIS-3, to A.S.) and CovidPhy (SA 304 C, to A.S.); (4) Spanish Ministry of Science and Innovation (MCIN)/Spanish Research Agency (AEI) (PID2022-142156OB-I00, to A.G.-C.); and (5) consorcio Centro de Investigación Biomédica en Red de Enfermedades Respiratorias (CB21/06/00103, to A.S. and F.M.-T.). A.G.-C. is supported by the Miguel Servet programme (CP23/00080) contract, funded by the Instituto de Salud Carlos III (ISCIII) and co-funded by the European Union. L.J.S. was supported by the NOMIS Foundation. The funders were not involved in the study design; collection, analysis, and interpretation of data; the writing of this article; or the decision to submit it for publication.

AUTHOR CONTRIBUTIONS

A.S., F.M.-T., A.G.-C., and S.V.-L. conceptualized and designed the study. A.S., F.M.-T., M.L., M.V.d.F., C.F., M.E., W.Z., A.J.P., T.T.K., S.A., M.P., H.A.M., and R.d.G. provided key resources. I.R.-C., A.I.D., J.A.H., A.J.C., and E.D.C. were responsible for clinical classification of the patients and collection of clinical metadata. C.L.V., U.v.B., D.S.K., M.T., D.Z., S.C.P., L.J.S., P.A., C.C., M.d.C.M.-P., A.P.-A., E.G.-S., J.V.-R., F.G.-S., P.A.-Q., and L.M.-G. contributed to patients' recruitment. V.J.W. coordinated sample processing and biobanking. S.V.-L., X.B., J.P.-S., A.G.-C., M.K., and D.H.-C. conducted bioinformatics and computational biology analysis. Computational models and machine learning approaches used to analyze the data were developed by A.S., S.V.-L., and A.G.-C.. A.S., S.V.-L., and A.G.-C. wrote the first draft of the manuscript. All authors contributed to discussions regarding the interpretation of results and reviewed and approved the final manuscript.

DECLARATION OF INTERESTS

The authors have a European patent application related to this work under the identification number EP24382240.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS
- METHOD DETAILS
  - RNA-seq analysis
  - Differential expression analysis
  - Pathway analysis
  - Signature discovery using LASSO
  - Signature validation
  - Weighted Gene Correlation Network Analysis
- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2025.111747.

Received: February 26, 2024
Revised: October 24, 2024
Accepted: January 2, 2025
Published: January 4, 2025

REFERENCES

1. Vos, T., Lim, S.S., Abbafati, C., Abbas, K.M., Abbasi, M., Abbasifard, M., Abbasi-Kangevari, M., Abbastabar, H., Abd-Allah, F., and Abdelalim, A. (2020). Global burden of 369 diseases and injuries in 204 countries and territories, 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019. Lancet *396*, 1204–1222.
2. World Health Organization (WHO) (2022). Pneumonia in children.
3. Haq, I.J., Battersby, A.C., Eastham, K., and McKean, M. (2017). Community acquired pneumonia in children. Br. Med. J. *356*, j686.
4. Shoar, S., and Musher, D.M. (2020). Etiology of community-acquired pneumonia in adults: a systematic review. Pneumonia *12*, 11.
5. Lynch, J.P., III. (2001). Hospital-acquired pneumonia: risk factors, microbiology, and treatment. Chest *119*, 373S–384S.
6. Torres, A., Catia, C., Niederman, M.S., Rosario, M., Chalmers, J.D., and Wunderink, R.G. (2021). Pneumonia (Primer). Nat. Rev. Dis. Prim. *7*, 25.
7. Jain, S., Williams, D.J., Arnold, S.R., Ampofo, K., Bramley, A.M., Reed, C., Stockmann, C., Anderson, E.J., Grijalva, C.G., Self, W.H., et al. (2015). Community-acquired pneumonia requiring hospitalization among US children. N. Engl. J. Med. *372*, 835–845.
8. Le Roux, D.M., and Zar, H.J. (2017). Community-acquired pneumonia in children—a changing spectrum of disease. Pediatr. Radiol. *47*, 1392–1398.
9. Schot, M.J.C., Dekker, A.R.J., Giorgi, W.G., Hopstaken, R.M., de Wit, N.J., Verheij, T.J.M., and Cals, J.W.L. (2018). Diagnostic value of signs, symptoms and diagnostic tests for diagnosing pneumonia in ambulant children in developed countries: a systematic review. NPJ Prim. Care Respir. Med. *28*, 40.
10. Rambaud-Althaus, C., Althaus, F., Genton, B., and D'Acremont, V. (2015). Clinical features for diagnosis of pneumonia in children younger than 5 years: a systematic review and meta-analysis. Lancet Infect. Dis. *15*, 439–450.
11. Scott, J.A.G., Wonodi, C., Moïsi, J.C., Deloria-Knoll, M., DeLuca, A.N., Karron, R.A., Bhat, N., Murdoch, D.R., Crawley, J., Levine, O.S., et al. (2012). The definition of pneumonia, the assessment of severity, and clinical standardization in the Pneumonia Etiology Research for Child Health study. Clin. Infect. Dis. *54*, S109–S116.
12. Eslamy, H.K., and Newman, B. (2011). Pneumonia in normal and immunocompromised children: an overview and update. Radiol. Clin. *49*, 895–920.
13. Prina, E., Ceccato, A., and Torres, A. (2016). New aspects in the management of pneumonia. Crit. Care *20*, 267.
14. Cillóniz, C., Torres, A., and Niederman, M.S. (2021). Management of pneumonia in critically ill patients. Br. Med. J. *375*, e065871.
15. Thomas, J., Pociute, A., Kevalas, R., Malinauskas, M., and Jankauskaite, L. (2020). Blood biomarkers differentiating viral versus bacterial pneumonia aetiology: a literature review. Ital. J. Pediatr. *46*, 4–10.
16. Christ-Crain, M., and Opal, S.M. (2010). Clinical review: the role of biomarkers in the diagnosis and management of community-acquired pneumonia. Crit. Care *14*, 203.
17. van Aerde, K.J., de Haan, L., van Leur, M., Gerrits, G.P., Schers, H., Moll, H.A., Hagedoorn, N.N., Herberg, J.A., Levin, M., Rivero-Calle, I., et al. (2021). Respiratory tract infection management and antibiotic prescription in children: a unique study comparing three levels of healthcare in the Netherlands. Pediatr. Infect. Dis. J. *40*, e100–e105.
18. Berti, E., Galli, L., de Martino, M., and Chiappini, E. (2013). International guidelines on tackling community-acquired pneumonia show major discrepancies between developed and developing countries. Acta Paediatr. *102*, 4–16.
19. Izadnegahdar, R., Cohen, A.L., Klugman, K.P., and Qazi, S.A. (2013). Childhood pneumonia in developing countries. Lancet Respir. Med. *1*, 574–584.
20. Lynch, T., Platt, R., Gouin, S., Larson, C., and Patenaude, Y. (2004). Can we predict which children with clinically suspected pneumonia will have the presence of focal infiltrates on chest radiographs? Pediatrics *113*, e186–e189.
21. Grief, S.N., and Loza, J.K. (2018). Guidelines for the Evaluation and Treatment of Pneumonia. Prim. Care *45*, 485–503.
22. Torres, A., Niederman, M.S., Chastre, J., Ewig, S., Fernandez-Vandellos, P., Hanberger, H., Kollef, M., Li Bassi, G., Luna, C.M., Martin-Loeches, I., et al. (2018). Summary of the international clinical guidelines for the management of hospital-acquired and ventilator-acquired pneumonia. ERJ Open Res. *4*, 00028-2018.
23. Cilloniz, C., Liapikou, A., and Torres, A. (2020). Advances in molecular diagnostic tests for pneumonia. Curr. Opin. Pulm. Med. *26*, 241–248.
24. Sungurlu, S., and Balk, R.A. (2018). The role of biomarkers in the diagnosis and management of pneumonia. Clin. Chest Med. *39*, 691–701.
25. Zar, H.J., Andronikou, S., and Nicol, M.P. (2017). Advances in the diagnosis of pneumonia in children. Br. Med. J. *358*, j2739.
26. Bhuiyan, M.U., Blyth, C.C., West, R., Lang, J., Rahman, T., Granland, C., de Gier, C., Borland, M.L., Thornton, R.B., Kirkham, L.-A.S., et al. (2019). Combination of clinical symptoms and blood biomarkers can improve discrimination between bacterial or viral community-acquired pneumonia in children. BMC Pulm. Med. *19*, 71–79.
27. Scicluna, B.P., Klein Klouwenberg, P.M.C., van Vught, L.A., Wiewel, M.A., Ong, D.S.Y., Zwinderman, A.H., Franitza, M., Toliat, M.R., Nürnberg, P., Hoogendijk, A.J., et al. (2015). A molecular biomarker to diagnose community-acquired pneumonia on intensive care unit admission. Am. J. Respir. Crit. Care Med. *192*, 826–835.
28. Verbakel, J.Y., Lemiengre, M.B., De Burghgraeve, T., De Sutter, A., Aertgeerts, B., Bullens, D.M.A., Shinkins, B., Van den Bruel, A., and Buntinx, F. (2018). Point-of-care C reactive protein to identify serious infection in acutely ill children presenting to hospital: prospective cohort study. Arch. Dis. Child. *103*, 420–426.
29. Eccles, S., Pincus, C., Higgins, B., and Woodhead, M.; Guideline Development Group (2014). Diagnosis and management of community and hospital acquired pneumonia in adults: summary of NICE guidance. Br. Med. J. *349*, g6722.
30. Herberg, J.A., Kaforou, M., Wright, V.J., Shailes, H., Eleftherohorinou, H., Hoggart, C.J., Cebey-López, M., Carter, M.J., Janes, V.A., Gormley, S.,

et al. (2016). Diagnostic test accuracy of a 2-transcript host RNA signature for discriminating bacterial vs viral infection in febrile children. JAMA *316*, 835–845.

31. Barral-Arca, R., Gómez-Carballa, A., Cebey-López, M., Currás-Tuala, M.J., Pischedda, S., Viz-Lasheras, S., Bello, X., Martinón-Torres, F., and Salas, A. (2020). Rna-seq data-mining allows the discovery of two long non-coding rna biomarkers of viral infection in humans. Int. J. Mol. Sci. *21*, 2748.
32. Mejias, A., Dimo, B., Suarez, N.M., Garcia, C., Suarez-Arrabal, M.C., Jartti, T., Blankenship, D., Jordan-Villegas, A., Ardura, M.I., Xu, Z., et al. (2013). Whole blood gene expression profiles to assess pathogenesis and disease severity in infants with respiratory syncytial virus infection. PLoS Med. *10*, e1001549.
33. Bloom, C.I., Graham, C.M., Berry, M.P.R., Rozakeas, F., Redford, P.S., Wang, Y., Xu, Z., Wilkinson, K.A., Wilkinson, R.J., Kendrick, Y., et al. (2013). Transcriptional blood signatures distinguish pulmonary tuberculosis, pulmonary sarcoidosis, pneumonias and lung cancers. PLoS One *8*, e70630.
34. Barral-Arca, R., Pardo-Seco, J., Martinón-Torres, F., and Salas, A. (2018). A 2-transcript host cell signature distinguishes viral from bacterial diarrhea and it is influenced by the severity of symptoms. Sci. Rep. *8*, 8043.
35. Gómez-Carballa, A., Cebey-López, M., Pardo-Seco, J., Barral-Arca, R., Rivero-Calle, I., Pischedda, S., Currás-Tuala, M.J., Gómez-Rial, J., Barros, F., and Martinón-Torres, F. (2019). A qPCR expression assay of IFI44L gene differentiates viral from bacterial infections in febrile children. Sci. Rep. *9*, 1–12.
36. Kaforou, M., Wright, V.J., Oni, T., French, N., Anderson, S.T., Bangani, N., Banwell, C.M., Brent, A.J., Crampin, A.C., Dockrell, H.M., et al. (2013). Detection of tuberculosis in HIV-infected and-uninfected African adults using whole blood RNA expression signatures: a case-control study. PLoS Med. *10*, e1001538.
37. Sweeney, T.E., Braviak, L., Tato, C.M., and Khatri, P. (2016). Genome-wide expression for diagnosis of pulmonary tuberculosis: a multicohort analysis. Lancet Respir. Med. *4*, 213–224.
38. Anderson, S.T., Kaforou, M., Brent, A.J., Wright, V.J., Banwell, C.M., Chagaluka, G., Crampin, A.C., Dockrell, H.M., French, N., Hamilton, M.S., et al. (2014). Diagnosis of childhood tuberculosis and host RNA expression in Africa. N. Engl. J. Med. *370*, 1712–1723.
39. Herberg, J.A., Kaforou, M., Gormley, S., Sumner, E.R., Patel, S., Jones, K.D.J., Paulus, S., Fink, C., Martinon-Torres, F., Montana, G., et al. (2013). Transcriptomic profiling in childhood H1N1/09 influenza reveals reduced expression of protein synthesis genes. J. Infect. Dis. *208*, 1664–1668.
40. Barral-Arca, R., Gómez-Carballa, A., Cebey-López, M., Bello, X., Martinón-Torres, F., and Salas, A. (2020). A meta-analysis of multiple whole blood gene expression data unveils a diagnostic host-response transcript signature for respiratory syncytial virus. Int. J. Mol. Sci. *21*, 1831.
41. DeBerg, H.A., Zaidi, M.B., Altman, M.C., Khaenam, P., Gersuk, V.H., Campos, F.D., Perez-Martinez, I., Meza-Segura, M., Chaussabel, D., Banchereau, J., et al. (2018). Shared and organism-specific host responses to childhood diarrheal diseases revealed by whole blood transcript profiling. PLoS One *13*, e0192082.
42. Mahajan, P., Kuppermann, N., Suarez, N., Mejias, A., Casper, C., Dean, J.M., and Ramilo, O.; Febrile Infant Working Group for the Pediatric Emergency Care Applied Research Network PECARN (2015). RNA transcriptional biosignature analysis for identifying febrile infants with serious bacterial infections in the emergency department: a feasibility study. Pediatr. Emerg. Care *31*, 1–5.
43. Gómez-Carballa, A., Barral-Arca, R., Cebey-López, M., Bello, X., Pardo-Seco, J., Martinón-Torres, F., and Salas, A. (2021). Identification of a minimal 3-transcript signature to differentiate viral from bacterial infection from best genome-wide host RNA biomarkers: a multi-cohort analysis. Int. J. Mol. Sci. *22*, 3148.
44. Parnell, G.P., McLean, A.S., Booth, D.R., Armstrong, N.J., Nalos, M., Huang, S.J., Manak, J., Tang, W., Tam, O.-Y., Chan, S., and Tang, B.M. (2012). A distinct influenza infection signature in the blood transcriptome of patients with severe community-acquired pneumonia. Crit. Care *16*, R157.
45. Viasus, D., Simonetti, A.F., Nonell, L., Vidal, O., Meije, Y., Ortega, L., Arnal, M., Bódalo-Torruella, M., Sierra, M., Rombauts, A., et al. (2023). Whole-Blood Gene Expression Profiles Associated with Mortality in Community-Acquired Pneumonia. Biomedicines *11*, 429.
46. Zhao, J., He, X., Min, J., Yao, R.S.Y., Chen, Y., Chen, Z., Huang, Y., Zhu, Z., Gong, Y., Xie, Y., et al. (2023). A multicenter prospective study of comprehensive metagenomic and transcriptomic signatures for predicting outcomes of patients with severe community-acquired pneumonia. EBioMedicine *96*, 104790.
47. Hopp, L., Loeffler-Wirth, H., Nersisyan, L., Arakelyan, A., and Binder, H. (2018). Footprints of sepsis framed within community acquired pneumonia in the blood transcriptome. Front. Immunol. *9*, 1620.
48. Anahtar, M., Chan, L.W., Ko, H., Rao, A., Soleimany, A.P., Khatri, P., and Bhatia, S.N. (2022). Host protease activity classifies pneumonia etiology. Proc. Natl. Acad. Sci. USA *119*, e2121778119.
49. Siljan, W.W., Sivakumaran, D., Ritz, C., Jenum, S., Ottenhoff, T.H., Ulvestad, E., Holter, J.C., Heggelund, L., and Grewal, H.M. (2022). Host transcriptional signatures predict etiology in community-acquired pneumonia: Potential antibiotic stewardship tools. Biomark. Insights *17*, 11772719221099130.
50. Principi, N., and Esposito, S. (2017). Biomarkers in pediatric community-acquired pneumonia. Int. J. Mol. Sci. *18*, 447.
51. Virkki, R., Juven, T., Rikalainen, H., Svedström, E., Mertsola, J., and Ruuskanen, O. (2002). Differentiation of bacterial and viral pneumonia in children. Thorax *57*, 438–441.
52. Silterra, J., Gillette, M.A., Lanaspa, M., Pellé, K.G., Valim, C., Ahmad, R., Acácio, S., Almendinger, K.D., Tan, Y., Madrid, L., et al. (2017). Transcriptional categorization of the etiology of pneumonia syndrome in pediatric patients in malaria-endemic areas. J. Infect. Dis. *215*, 312–320.
53. Wallihan, R.G., Suárez, N.M., Cohen, D.M., Marcon, M., Moore-Clingenpeel, M., Mejias, A., and Ramilo, O. (2018). Molecular distance to health transcriptional score and disease severity in children hospitalized with community-acquired pneumonia. Front. Cell. Infect. Microbiol. *8*, 382.
54. Kolberg, L., Khanijau, A., van der Velden, F.J., Herberg, J., De, T., Galassini, R., Cunnington, A.J., Wright, V.J., Shah, P., and Kaforou, M. (2023). Raising AWaRe-ness of antimicrobial stewardship challenges in pediatric emergency care: results from the PERFORM study assessing consistency and appropriateness of antibiotic prescribing across Europe. Clin. Infect. Dis. *78*, 526–534.
55. Juss, J.K., House, D., Amour, A., Begg, M., Herre, J., Storisteanu, D.M.L., Hoenderdos, K., Bradley, G., Lennon, M., Summers, C., et al. (2016). Acute respiratory distress syndrome neutrophils have a distinct phenotype and are resistant to phosphoinositide 3-kinase inhibition. Am. J. Respir. Crit. Care Med. *194*, 961–973.
56. Potey, P.M., Rossi, A.G., Lucas, C.D., and Dorward, D.A. (2019). Neutrophils in the initiation and resolution of acute pulmonary inflammation: understanding biological function and therapeutic potential. J. Pathol. *247*, 672–685.
57. Soehnlein, O., Oehmcke, S., Rothfuchs, A., Rothfuchs, A.G., Frithiof, R., Mörgelin, M., Mörgelin, M., Herwald, H., and Lindbom, L. (2008). Neutrophil degranulation mediates severe lung damage triggered by streptococcal M1 protein. Eur. Respir. J. *32*, 405–412.
58. Sande, C.J., Njunge, J.M., Mwongeli Ngoi, J., Mutunga, M.N., Chege, T., Gicheru, E.T., Gardiner, E.M., Gwela, A., Green, C.A., Drysdale, S.B., et al. (2019). Airway response to respiratory syncytial virus has incidental antibacterial effects. Nat. Commun. *10*, 2218.

59. Yao, Y., Zhao, J., Hu, J., Song, H., Wang, S., and Wang, Y. (2022). Identification of a Four-Gene Signature for Diagnosing Paediatric Sepsis. BioMed Res. Int. *2022*, 5217885.

60. Lu, J., Li, Q., Wu, Z., Zhong, Z., Ji, P., Li, H., He, C., Feng, J., and Zhang, J. (2020). Two gene set variation indexes as potential diagnostic tool for sepsis. Am. J. Transl. Res. *12*, 2749–2759.

61. Banerjee, S., Mohammed, A., Wong, H.R., Palaniyar, N., and Kamaleswaran, R. (2021). Machine learning identifies complicated sepsis course and subsequent mortality based on 20 genes in peripheral blood immune cells at 24 H post-ICU admission. Front. Immunol. *12*, 592303.

62. Cabral-Pacheco, G.A., Garza-Veloz, I., Castruita-De la Rosa, C., Ramirez-Acuña, J.M., Perez-Romero, B.A., Guerrero-Rodriguez, J.F., Martinez-Avila, N., and Martinez-Fierro, M.L. (2020). The roles of matrix metalloproteinases and their inhibitors in human diseases. Int. J. Mol. Sci. *21*, 9739.

63. Uhel, F., Scicluna, B.P., van Vught, L.A., Cremer, O.L., Bonten, M.J., Schultz, M.J., and van der Poll, T. (2019). Matrix metalloproteinase-8: a useful biomarker to refine the diagnosis of community-acquired pneumonia upon intensive care unit admission? Crit. Care *23*, 226.

64. Lauhio, A., Hästbacka, J., Pettilä, V., Tervahartiala, T., Karlsson, S., Varpula, T., Varpula, M., Ruokonen, E., Sorsa, T., and Kolho, E. (2011). Serum MMP-8,-9 and TIMP-1 in sepsis: high serum levels of MMP-8 and TIMP-1 are associated with fatal outcome in a multicentre, prospective cohort study. Hypothetical impact of tetracyclines. Pharmacol. Res. *64*, 590–594.

65. Fang, X., Duan, S.-F., Hu, Z.-Y., Wang, J.-J., Qiu, L., Wang, F., and Chen, X.-L. (2022). Inhibition of Matrix Metalloproteinase-8 Protects Against Sepsis Serum Mediated Leukocyte Adhesion. Front. Med. *9*, 814890.

66. Herberg, J., Huang, H., Thezenas, M.L., Janes, V., Carter, M., Gormley, S., Hamilton, M.S., Kessler, B., Levin, M., and Casals-Pascual, C. (2019). Lipocalin-2 is a sensitive and specific marker of bacterial infection in children. Preprint at bioRxiv. https://doi.org/10.1016/j.cmi.2020.07.006.

67. Liu, W., and Rodgers, G.P. (2022). Olfactomedin 4 is a biomarker for the severity of infectious diseases (Oxford University Press US), p. ofac061.

68. Alder, M.N., Opoka, A.M., Lahni, P., Hildeman, D.A., and Wong, H.R. (2017). Olfactomedin 4 is a candidate marker for a pathogenic neutrophil subset in septic shock. Crit. Care Med. *45*, e426–e432.

69. Demaret, J., Venet, F., Plassais, J., Cazalis, M.-A., Vallin, H., Friggeri, A., Lepape, A., Rimmelé, T., Textoris, J., and Monneret, G. (2016). Identification of CD177 as the most dysregulated parameter in a microarray study of purified neutrophils from septic shock patients. Immunol. Lett. *178*, 122–130.

70. Levy, Y., Wiedemann, A., Hejblum, B.P., Durand, M., Lefebvre, C., Surenaud, M., Lacabaratz, C., Perreau, M., Foucat, E., and Dechenaud, M. (2021). CD177, a specific marker of neutrophil activation, is associated with coronavirus disease 2019 severity and death. iScience *24*, 102711.

71. Wang, Z., Zhang, L., Li, S., Xu, F., Han, D., Wang, H., Huang, T., Yin, H., and Lyu, J. (2022). The relationship between hematocrit and serum albumin levels difference and mortality in elderly sepsis patients in intensive care units—a retrospective study based on two large database. BMC Infect. Dis. *22*, 629.

72. Takegawa, R., Kabata, D., Shimizu, K., Hisano, S., Ogura, H., Shintani, A., and Shimazu, T. (2019). Serum albumin as a risk factor for death in patients with prolonged sepsis: an observational study. J. Crit. Care *51*, 139–144.

73. Dai, D.-M., Wang, D., Hu, D., Wan, W.-L., Su, Y., Yang, J.-L., Wang, Y.-P., Wang, F., Yang, L., Sun, H.-M., et al. (2020). Difference in hematocrit and plasma albumin levels as an additional biomarker in the diagnosis of infectious disease. Arch. Med. Sci. *16*, 522–530.

74. Turcato, G., Zaboli, A., Kostic, I., Melchioretto, B., Ciccariello, L., Zaccaria, E., Olivato, A., Maccagnani, A., Pfeifer, N., and Bonora, A. (2022). Severity of SARS-CoV-2 infection and albumin levels recorded at the first emergency department evaluation: a multicentre retrospective observational study. Emerg. Med. J. *39*, 63–69.

75. Chen, C., Zhang, Y., Zhao, X., Tao, M., Yan, W., and Fu, Y. (2021). Hypoalbuminemia–An indicator of the severity and prognosis of COVID-19 patients: a multicentre retrospective analysis. Infect. Drug Resist. *14*, 3699–3710.

76. Cheng, J., Ji, D., Yin, Y., Wang, S., Song, K., Pan, Q., Zhang, Q., and Yang, L. (2022). Proteomic profiling of serum small extracellular vesicles reveals immune signatures of children with pneumonia. Transl. Pediatr. *11*, 891–908.

77. Kellum, J.A., Kong, L., Fink, M.P., Weissfeld, L.A., Yealy, D.M., Pinsky, M.R., Fine, J., Krichevsky, A., Delude, R.L., and Angus, D.C.; GenIMS Investigators (2007). Understanding the inflammatory cytokine response in pneumonia and sepsis: results of the Genetic and Inflammatory Markers of Sepsis (GenIMS) Study. Arch. Intern. Med. *167*, 1655–1663.

78. Nguyen Thi Dieu, T., Pham Nhat, A., Craig, T.J., and Duong-Quy, S. (2017). Clinical characteristics and cytokine changes in children with pneumonia requiring mechanical ventilation. J. Int. Med. Res. *45*, 1805–1817.

79. Feikin, D.R., Hammitt, L.L., Murdoch, D.R., O'Brien, K.L., and Scott, J.A.G. (2017). The enduring challenge of determining pneumonia etiology in children: considerations for future research priorities. Clin. Infect. Dis. *64*, S188–S196.

80. Murdoch, D.R., O'Brien, K.L., Driscoll, A.J., Karron, R.A., and Bhat, N.Pneumonia Methods Working Group; PERCH Core Team (2012). Laboratory methods for determining pneumonia etiology in children. Clin. Infect. Dis. *54*, S146–S152.

81. Shah, P., Voice, M., Calvo-Bado, L., Rivero-Calle, I., Morris, S., Nijman, R., Broderick, C., De, T., Eleftheriou, I., Galassini, R., et al. (2023). Relationship between molecular pathogen detection and clinical disease in febrile children across Europe: a multicentre, prospective observational study. Lancet Reg. Health. Eur. *32*, 100682.

82. Arora, S., Ahmad, S., Irshad, R., Goyal, Y., Rafat, S., Siddiqui, N., Dev, K., Husain, M., Ali, S., Mohan, A., and Syed, M.A. (2019). TLRs in pulmonary diseases. Life Sci. *233*, 116671.

83. Chollet-Martin, S., Montravers, P., Gibert, C., Elbim, C., Desmonts, J.M., Fagon, J.Y., and Gougerot-Pocidalo, M.A. (1992). Subpopulation of hyperresponsive polymorphonuclear neutrophils in patients with adult respiratory distress syndrome. Am. Rev. Respir. Dis. *146*, 990–996.

84. Fitzgerald, K.A., and Kagan, J.C. (2020). Toll-like receptors and the control of immunity. Cell *180*, 1044–1066.

85. Medvedev, A.E., Lentschat, A., Kuhns, D.B., Blanco, J.C.G., Salkowski, C., Zhang, S., Arditi, M., Gallin, J.I., and Vogel, S.N. (2003). Distinct mutations in IRAK-4 confer hyporesponsiveness to lipopolysaccharide and interleukin-1 in a patient with recurrent bacterial infections. J. Exp. Med. *198*, 521–531.

86. Ku, C.-L., Picard, C., Erdös, M., Jeurissen, A., Bustamante, J., Puel, A., von Bernuth, H., Filipe-Santos, O., Chang, H.-H., Lawrence, T., et al. (2007). IRAK4 and NEMO mutations in otherwise healthy children with recurrent invasive pneumococcal disease. J. Med. Genet. *44*, 16–23.

87. Picard, C., Casanova, J.-L., and Puel, A. (2011). Infectious diseases in patients with IRAK-4, MyD88, NEMO, or IκBα deficiency. Clin. Microbiol. Rev. *24*, 490–497.

88. Von Bernuth, H., Picard, C., Jin, Z., Pankla, R., Xiao, H., Ku, C.-L., Chrabieh, M., Mustapha, I.B., Ghandil, P., Camcioglu, Y., et al. (2008). Pyogenic bacterial infections in humans with MyD88 deficiency. Science *321*, 691–696.

89. Xu, X., Yuan, B., Liang, Q., Huang, H., Yin, X., Sheng, X., Nie, N., and Fang, H. (2015). Gene expression profile analysis of ventilator-associated pneumonia. Mol. Med. Rep. *12*, 7455–7462.

90. Jagadeesh, A., Prathyusha, A., Sheela, G.M., and Bramhachari, P.V. (2020). T cells in viral infections: the myriad flavours of antiviral immunity.

Dynam. Immune Activation Viral Dis., 139–148. https://doi.org/10.1007/978-981-15-1045-8_9.

91. Kervevan, J., and Chakrabarti, L.A. (2021). Role of CD4+ T cells in the control of viral infections: Recent advances and open questions. Int. J. Mol. Sci. *22*, 523.
92. Schmidt, M.E., and Varga, S.M. (2018). The CD8 T cell response to respiratory virus infections. Front. Immunol. *9*, 678.
93. Swain, S.L., McKinstry, K.K., and Strutt, T.M. (2012). Expanding roles for CD4+ T cells in immunity to viruses. Nat. Rev. Immunol. *12*, 136–148.
94. Santiago, R.P., Figueiredo, C.V.B., Fiuza, L.M., Yahouédéhou, S.C.M.A., Oliveira, R.M., Aleluia, M.M., Carvalho, S.P., Fonseca, C.A., Nascimento, V.M.L., Rocha, L.C., et al. (2020). Transforming growth factor beta receptor 3 haplotypes in sickle cell disease are associated with lipid profile and clinical manifestations. Mediat. Inflamm. *2020*, 3185015.
95. Ning, J., Sun, K., Wang, X., Fan, X., Jia, K., Cui, J., and Ma, C. (2023). Use of machine learning-based integration to develop a monocyte differentiation-related signature for improving prognosis in patients with sepsis. Mol. Med. *29*, 37.
96. Basu, S., Naha, A., Veeraraghavan, B., Ramaiah, S., and Anbarasu, A. (2022). In silico structure evaluation of BAG3 and elucidating its association with bacterial infections through protein–protein and host-pathogen interaction analysis. J. Cell. Biochem. *123*, 115–127.
97. Nalbant, D., Youn, H., Nalbant, S.I., Sharma, S., Cobos, E., Beale, E.G., Du, Y., and Williams, S.C. (2005). FAM20: an evolutionarily conserved family of secreted proteins expressed in hematopoietic cells. BMC Genom. *6*, 11–21.
98. Lu, X., Xue, L., Sun, W., Ye, J., Zhu, Z., and Mei, H. (2018). Identification of key pathogenic genes of sepsis based on the Gene Expression Omnibus database. Mol. Med. Rep. *17*, 3042–3054.
99. Davenport, E.E., Burnham, K.L., Radhakrishnan, J., Humburg, P., Hutton, P., Mills, T.C., Rautanen, A., Gordon, A.C., Garrard, C., Hill, A.V.S., et al. (2016). Genomic landscape of the individual host response and outcomes in sepsis: a prospective cohort study. Lancet Respir. Med. *4*, 259–271.
100. Jia, C., Zhang, F., Zhu, Y., Qi, X., and Wang, Y. (2017). Public data mining plus domestic experimental study defined involvement of the old-yet-uncharacterized gene matrix-remodeling associated 7 (MXRA7) in physiopathology of the eye. Gene *632*, 43–49.
101. Wang, H., Guo, Y., Lu, H., Luo, Y., Hu, W., Liang, W., Garcia-Barrio, M.T., Chang, L., Schwendeman, A., Zhang, J., and Chen, Y.E. (2022). Krüppel-like factor 14 deletion in myeloid cells accelerates atherosclerotic lesion development. Cardiovasc. Res. *118*, 475–488.
102. Yuan, Y., Fan, G., Liu, Y., Liu, L., Zhang, T., Liu, P., Tu, Q., Zhang, X., Luo, S., Yao, L., et al. (2022). The transcription factor KLF14 regulates macrophage glycolysis and immune function by inhibiting HK2 in sepsis. Cell. Mol. Immunol. *19*, 504–515.
103. Larcher, R., Boudet, A., Roger, C., Villa, F., and Loubet, P. (2024). Mycoplasma pneumoniae is back! Is it the next pandemic? Anaesth. Crit. Care Pain Med. *43*, 101338.
104. Sauteur, P.M.M., Beeton, M.L., Pereyre, S., Bébéar, C., Gardette, M., Hénin, N., Wagner, N., Fischer, A., Vitale, A., and Lemaire, B. (2023). Mycoplasma pneumoniae: delayed re-emergence after COVID-19 pandemic restrictions. Lancet Microbe *5*, e100.
105. Habgood-Coote, D., Wilson, C., Shimizu, C., Barendregt, A.M., Philipsen, R., Galassini, R., Calle, I.R., Workman, L., Agyeman, P.K.A., Ferwerda, G., et al. (2023). Diagnosis of childhood febrile illness using a multi-class blood RNA molecular signature. Med *4*, 635–654. https://doi.org/10.1016/j.medj.2023.06.007.
106. Jackson, H.R., Miglietta, L., Habgood-Coote, D., D'Souza, G., Shah, P., Nichols, S., Vito, O., Powell, O., Davidson, M.S., Shimizu, C., et al. (2023). Diagnosis of Multisystem Inflammatory Syndrome in Children by a Whole-Blood Transcriptional Signature. J. Pediatric Infect. Dis. Soc. *12*, 322–331. https://doi.org/10.1093/jpids/piad035.
107. Martinón-Torres, F., Salas, A., Rivero-Calle, I., Cebey-López, M., Pardo-Seco, J., Herberg, J.A., Boeddha, N.P., Klobassa, D.S., Secka, F., Paulus, S., et al. (2018). Life-threatening infections in children in Europe (the EUCLIDS Project): a prospective cohort study. Lancet. Child Adolesc. Health *2*, 404–414.
108. Habgood-Coote, D., Wilson, C., Shimizu, C., Barendregt, A.M., Philipsen, R., Galassini, R., Calle, I.R., Workman, L., Agyeman, P.K.A., Ferwerda, G., et al. (2023). Diagnosis of childhood febrile illness using a multi-class blood RNA molecular signature. Med *4*, 635–654.
109. Andrews, S. (2010). FastQC: A Quality Control Tool for High Throughput Sequence Data.
110. Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics *29*, 15–21.
111. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R.; 1000 Genome Project Data Processing Subgroup (2009). The sequence alignment/map format and SAMtools. Bioinformatics *25*, 2078–2079.
112. Liao, Y., Smyth, G.K., and Shi, W. (2014). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics *30*, 923–930.
113. Zhang, Y., Parmigiani, G., and Johnson, W.E. (2020). ComBat-seq: batch effect adjustment for RNA-seq count data. NAR Genom. Bioinform. *2*, lqaa078. https://doi.org/10.1093/nargab/lqaa078.
114. Risso, D., Ngai, J., Speed, T.P., and Dudoit, S. (2014). Normalization of RNA-seq data using factor analysis of control genes or samples. Nat. Biotechnol. *32*, 896–902.
115. Robinson, M.D., and Oshlack, A. (2010). A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol. *11*, R25.
116. Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W., and Smyth, G.K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. *43*, e47.
117. Yu, G., Wang, L.-G., Han, Y., and He, Q.-Y. (2012). clusterProfiler: an R package for comparing biological themes among gene clusters. OMICS A J. Integr. Biol. *16*, 284–287.
118. Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. J. Stat. Software *33*, 1–22.
119. Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., and Müller, M. (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinf. *12*, 77–78.
120. López-Ratón, M., Rodríguez-Álvarez, M.X., Cadarso-Suárez, C., and Gude-Sampedro, F. (2014). OptimalCutpoints: an R package for selecting optimal cutpoints in diagnostic tests. J. Stat. Software *61*, 1–36.
121. Langfelder, P., and Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. BMC Bioinf. *9*, 559.

USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Biological samples		
Human whole blood samples	This paper	Available in Table S7
Critical commercial assays		
TruSeq RNA Sample Preparation Kit	Illumina	# RS-122-2001
Illumina® Ribo-Zero Gold kit	Illumina	# 20037135
Deposited data		
RNA-Seq data	This paper	GEO database: GSE261482
RNA-Seq data	This paper	ArrayExpress database: E-MTAB-14564
RNA-Seq data	Jackson et al.[106]	ArrayExpress database: E-MTAB-12793
RNA-Seq data	Habgood-Coote et al.[105]	ArrayExpress database: E-MTAB-11671
Microarray data	Wallihan et al.[68]	GEO accession number: GSE103119
R code	This paper	https://github.com/Sandraviz/Pneumonia_Transcriptomic-Analysis
Software and algorithms		
R software	https://www.r-project.org/	Version 4.2.0
FastQC	Andrews et al.[55]	https://github.com/s-andrews/FastQC
STAR	Dobin et al.[56]	https://github.com/alexdobin/STAR
FeatureCounts	Liao et al.[58]	https://www.bioconductor.org/packages/release/bioc/html/Rsubread.html
DESeq2	Love et al.[59]	https://bioconductor.org/packages/release/bioc/html/DESeq2.html
RUVSeq	Risso et al.[60]	https://bioconductor.org/packages/release/bioc/html/RUVSeq.html
limma	Ritchie et al.[62]	https://bioconductor.org/packages/release/bioc/html/limma.html
Clusterprofiler	Yu et al.[63]	https://bioconductor.org/packages/release/bioc/html/clusterProfiler.html
glmnet	Friedman et al.[64]	https://cran.r-project.org/web/packages/glmnet/index.html
pROC	Robin et al.[65]	https://cran.r-project.org/web/packages/pROC/index.html
OptimalCutPoints	López-Ratón et al.[66]	https://cran.r-project.org/web/packages/OptimalCutpoints/index.html
WGCNA	Langfelder et al.[67]	https://cran.r-project.org/web/packages/WGCNA/index.html

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

This study utilized blood samples from paediatric patients recruited in the European Union Childhood Life-threatening Infectious Diseases Study (EUCLIDS- https://www.euclids-project.eu/).[107] The EUCLIDS study enrolled children and young adolescents aged 1 month to 18 years who were admitted to hospitals with suspected sepsis or severe focal infection. The study encompassed a network of 194 hospitals across nine European countries, as well as two hospitals located in The Gambia, Africa. Patient recruitment occurred between July 1, 2012, and December 31, 2015.

Pneumonia phenotype was defined as clinical symptoms compatible with acute respiratory infection and inflammation of one or both lungs with lobar or segmental or multilobar collapse/consolidation on chest X-ray. More specifically, following radiological findings of consolidation/pleural effusion must be met: alveolar consolidation (defined as a dense or fluffy opacity that occupies a portion or whole of a lobe or of the entire lung that may or may not contain air-bronchograms) or pleural effusion (defined as fluid

in the lateral pleural space and not just in the minor or oblique fissure) that was spatially associated with a pulmonary parenchymal infiltrate (including other infiltrate) or obliterated enough of the hemithorax to obscure an opacity. It does not include just perihilar consolidation or patchy consolidation. Children with pneumonia (n = 154) and age and sex matched healthy controls (n = 38; Table 1) were selected from the EUCLIDS database[108] for transcriptome analysis in whole blood samples collected in PAXgene™ and Tempus™ RNA tubes. In a subsequent analysis, pneumonia samples were further classified using the algorithm developed by Herberg et al.[30] C-reactive protein (CRP) and neutrophil threshold criteria were employed where available, with 72.5% and 33.3% of missing CRP values in definitive bacterial (DB) and definitive viral (DV), respectively (see[30] for details on clinical characteristics of the cohort and definitions on DB and DV). Thus, DB (n = 40) and DV (n = 9) pneumonias were selected to study the differences in transcriptome depending on the etiology of the infection. Pneumonias categorized as Probable Bacterial (PB) and Probable Viral (PV) were excluded from this specific analysis to minimize possible confounding factors in stablishing a transcriptomic signature for differentiating bacterial from viral pneumonia in children but were used post-hoc to test the signature in a more realistic clinical scenario (Figure 1).

In order to validate the proposed signature, a new independent cohort comprising blood samples obtained from children with viral (n = 10) and bacterial (n = 22) pneumonias was employed (Figure 1). Clinical and demographic information are provided in Table 2. These samples were recruited by the PErsonalised Risk assessment in Febrile illness to Optimise Real-life Management across the European Union (PERFORM - https://www.perform2020.org/) and the Diagnosis and Management of Febrile Illness using RNA Personalised Molecular Signature Diagnosis (DIAMONDS - https://www.diamonds2020.eu) consortiums.

All newly generated data from human patients were obtained through harmonized procedures for patient recruitment, classification, clinical data collection, and sample acquisition, processing, and storage across the participating centers. All necessary ethical approvals were obtained from each participating country's Ethics Committee (EC).

METHOD DETAILS

RNA-seq analysis

For the discovery cohort, RNA was isolated from blood samples collected in PAXgene™ and in Tempus™ RNA tubes. RNA sequencing was conducted on a HiSeq 4000 (Illumina) platform, with library preparation and sequencing of 30 million 75 or 100 bp paired-end reads. The Illumina's TruSeq RNA Sample Preparation Kit was used for library preparation, and ribosomal and globin RNA depletion was performed using the Illumina® Ribo-Zero Gold kit. For the validation cohort, RNA was isolated using PAXgene blood miRNA isolation kit according to the manufacturer's instructions (Qiagen). A DNAse treatment was carried out with the RNA clean & concentrator kit (Zymo Research) prior to sequencing. RNA was quantified using RiboGreen (Invitrogen) on the FLUOstar OPTIMA plate reader (BMG Labtech) and the integrity was analyzed on the TapeStation 2200(Agilent, RNA ScreenTape). After a normalization step, a strand specific library preparation was completed using NEBNext® Ultra™ II mRNA kit (NEB) and NEB rRNA/globin depletion probes following manufacturer's recommendations. Individual libraries were normalized using Qubit, pooled together and diluted. The sequencing was performed using a 150 paired-end configuration in a Novaseq6000 platform.

Quality control of all raw data was carried out using *FastQC*,[109] alignment and read counting were performed using *STAR*,[110] alignment filtering was done with *SAMtools*[111] and read counting was carried out using *FeatureCounts*.[112]

Differential expression analysis

RNAseq data was processed for batch correction using control samples and *COMBAT-Seq* R package.[113] The *RUVg* method implemented in the *RUVSeq* package of R[114] was employed to identified and adjust for other factors of undesirable variation present in the dataset, selecting a $k = 2$ as a number of variables to include. After that, data normalization, differential expression (DE) analysis and covariates correction were carried out using *DESeq* package.[115] Differentially expressed genes (DEGs) were identified through the Negative Binomial distribution implemented in the *DESeq* package. The fitted model included known covariates such as sex, batch, and type of collection tube as well as the factors of undesirable variation, to correct the differences in gene expression related to these factors. A generalized linear model was fitted, and a t-statistic was calculated for each gene. P values were corrected for multiple testing using the Benjamini-Hochberg False Discovery Rate (FDR).

As a validation cohort for the pneumonia *vs*. control study, we downloaded the GSE103119 microarray dataset from the Gene Expression Omnibus (GEO) database. This dataset comprises whole blood gene expression data from children with pneumonia (n = 152) as well as healthy controls (n = 20). Data were normalized, processed and analyzed using the *limma* R package.[116] A linear model was fitted, considering the sex as a categorical covariate. Multiple testing correction was performed using the FDR.

Principal Component Analysis (PCA) was used to visualize the global transcriptome patterns of both RNA-seq and microarray data. A Spearman test was computed for the correlation indices (ρ) and the associated p values. Wilcoxon test was performed to assess statistical significance between groups.

Pathway analysis

We used the Reactome and GO (Gene Ontology) databases as references to examine biological pathways and processes associated with the DEGs in pneumonia patients. We followed both an over-representation analysis (ORA) and a gene-set enrichment (GSEA) approaches, considering DEGs with a $\log_2 FC \geq |1.5|$ and an adjusted p value < 0.05. The analysis was performed using the

Clusterprofiler[117] R package. To account for multiple test correction, we applied the FDR procedure, establishing a significant threshold of 0.05.

Signature discovery using LASSO

We employed a Least Absolute Shrinkage and Selection Operator (LASSO) regression model to identify a subset of genes that could serve as a predictive transcriptomic signature differentiating viral from bacterial pneumonia in the RNA-seq cohort. A predictive transcriptomic signature was computed using the R package *glmnet*[118] and the *cv.glmnet* function. A logistic LASSO regression model was fitted with the alpha parameter set to 1. We selected 36 DEGs filtered based on $|\log_2 FC| > 1.5$, adjusted p value < 0.05, and a Base Mean > 50 as input for the model. To determine the optimal parameter selection for the LASSO regression model, we conducted a 10-fold cross-validation, which helps evaluate the model's performance and effectively tune the parameters.

To assess the accuracy of the predictive transcriptomic signature, we calculated the area under the receiver operating characteristic curve (AUC) with 95% confidence intervals (CI) using the *pROC* package[119] in R. Receiver Operating Characteristic (ROC) curves were generated to graphically represent the model's true positive rate against the false positive rate. The optimal cut-point value, maximizing sensitivity and specificity, was calculated using the Youden method included in the *OptimalCutPoints* R package.[120]

Signature validation

We assessed the performance and accuracy of the 5-transcript signatures using RNAseq data from an additional pediatric cohort (Table 2). Coefficients and intercepts from the LASSO model were applied to the new data to perform ROC analysis and calculate the AUC. We tested different data comparisons: DB vs. DV pneumonias, a disaggregation of different DB pneumonias by the causal pathogen (*S.pneumoniae*, *S.aureus* and *S.pyogenes*) *vs*. DV pneumonias and a disaggregation by severity groups based on PICU admission of the patients (mild/moderate *vs.* severe disease).

Weighted Gene Correlation Network Analysis

We conducted a co-expression network analysis using the Weighted Gene Correlation Network Analysis (*WGCNA)* package[121] to detect clusters of co-expressed genes specifically correlated to viral or bacterial pneumonia. We selected all patients with DV ($n = 9$) and DB ($n = 40$) pneumonia for the analysis. Only protein-coding genes exhibiting the highest expression variance among samples (the top 25%) were included, totaling 4,831 genes. Normalized gene expression data were used to construct a signed weighted correlation network. A matrix of correlations between all pairs of selected genes was generated from the expression values and converted into an adjacency matrix with a power function. We determined a soft-thresholding power of 20 (maximum model fitting index), selected on the criterion of scale-free topology after evaluating several candidate powers (Figure S5).

We calculated the Topological Overlap Matrix (TOM) and the corresponding dissimilarity (1–TOM) values. For module detection and merging, we showed a minimum module size of 30, and a cut height threshold of 20, respectively. Correlation between module eigengenes and the viral/bacterial pneumonia was calculated to identify modules of co-expressed genes significantly associated with the phenotype (gene significance, or GS) and using the viral category as reference. Module Membership (MM) measured intramodular connectivity, and the most interconnected gene (hub gene) named the modules.

The biological pathways represented by each of the significantly correlated modules were investigated through an over-representation analysis with the *Clusterprofiler* R package,[117] with terms from GO and Reactome databases as references.

All graphics were created using R software v.4.2.0 (www-r-project.org).

QUANTIFICATION AND STATISTICAL ANALYSIS

Wilcoxon and Fisher exact test were used to assess statistical significance on clinical and demographic features between groups in numerical and categorical variables respectively. (*) indicates statistical significance. "*n*" represent number of patients or controls in each case and median values are presented with either proportion on their group or interquartile ranges with the observed minimum and maximum values.

USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

5.5. ARTICLE 5: A DIAGNOSTIC HOST-SPECIFIC TRANSCRIPTOME RESPONSE FOR MYCOPLASMA PNEUMONIAE PNEUMONIA TO GUIDE PEDIATRIC PATIENT TREATMENT

Sandra Viz-Lasheras[1,2,3,#], Alberto Gómez-Carballa[1,2,3,#], Jacobo Pardo-Seco[1,2,3], Xabier Bello[1,2,3], Irene Rivero-Calle[1,3,4], Ana Isabel Dacosta[1,3,4], Myrsini Kaforou[5], Dominic Coote[5], Aubrey J. Cunnington[5], Marieke Emonts[6,7,8], Jethro A. Herberg[5], Victoria J. Wright[5], Enitan D. Carrol[9,10], Stephane C. Paulus[11], Werner Zenz[12], Daniela S. Kohlfürst[12], Michiel Van der Flier[13,14], Ronald de Groot[14], Luregn J. Schlapbach[15], Philipp Agyeman[16], Andrew J. Pollard[11], Colin Fink[17], Taco T. Kuijpers[18], Suzanne Anderson[19], Cristina Calvo[20,21,22,23,24], María del Carmen Martínez-Padilla[25], Ana Pérez-Aragón[26], Esteban Gómez-Sánchez[27], Juan Valencia-Ramos[27], Francisco Giménez-Sánchez[28], Paula Alonso-Quintela[29], Laura Moreno-Galarraga[30,31], Ulrich von Both[32], Marko Pokorn[33], Dace Zavadska[34], María Tsolia[35], Clementien L. Vermont[36], Henriëtte A. Moll[36], Michael Levin[5], Federico Martinón-Torres[1,3,4], Antonio Salas[1,2,3,&], on behalf of EUCLIDS, DIAMONDS, GENDRES, and PERFORM consortia.

Affiliations

1- Genetics, Vaccines and Infections Research Group (GenViP), Instituto de Investigación Sanitaria de Santiago, 15706 Universidade de Santiago de Compostela, Santiago de Compostela, Galicia, Spain

2- Unidade de Xenética, Instituto de Ciencias Forenses, Facultade de Medicina, Universidade de Santiago de Compostela, and Genética de Poblaciones en Biomedicina (GenPoB) Research Group, Instituto de Investigación Sanitaria (IDIS), 15706 Hospital Clínico Universitario de Santiago (SERGAS), Galicia, Spain

3- Centro de Investigación Biomédica en Red de Enfermedades Respiratorias (CIBER-ES), Madrid, Spain

4- Translational Pediatrics and Infectious Diseases, Department of Pediatrics, 15706 Hospital Clínico Universitario de Santiago de Compostela, Santiago de Compostela, Galicia, Spain

5- Department of Infectious Disease, Imperial College London, London, W2 1PG, United Kingdom

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

6- Great North Children's Hospital, Paediatric Immunology, Infectious Diseases & Allergy, Newcastle upon Tyne Hospitals NHS Foundation Trust, Newcastle upon Tyne, NE1 4LP, United Kingdom

7- Translational and Clinical Research Institute, Newcastle University, Newcastle upon Tyne, NE2 4HH, United Kingdom

8- NIHR Newcastle Biomedical Research Centre based at Newcastle upon Tyne Hospitals NHS Trust and Newcastle University, Newcastle upon Tyne, NE4 5PL, United Kingdom

9- Department of Infectious Diseases, Alder Hey Children's NHS Foundation Trust, Liverpool, L12 2AP, United Kingdom

10- Department of Clinical Infection, Microbiology and Immunology, Institute of Infection, Veterinary and Ecological Sciences, University of Liverpool, Liverpool, L69 7BE, United Kingdom

11- Department of Paediatrics, University of Oxford and the NIHR Oxford Biomedical Research Centre, Oxford, OX3 9DU, United Kingdom

12- Department of General Paediatrics, Medical University of Graz, Graz, Auenbruggerplatz 34/2, 8036, Graz, Austria

13- Pediatric Infectious Diseases and Immunology, Wilhelmina Children's Hospital, University Medical Center Utrecht, Utrecht, 3508 AB, The Netherlands

14- Pediatric Infectious Diseases and Immunology, Amalia Children's Hospital, and Section Pediatric Infectious Diseases, Laboratory of Medical Immunology, Department of Laboratory Medicine, Radboud Institute for Molecular Life Sciences, Radboud University Medical Center, Nijmegen, 6500 HB, The Netherlands

15- Department of Intensive Care and Neonatology, and Children's Research Center, University Children's Hospital Zürich, University of Zürich, Zürich, Switzerland

16- Department of Pediatrics, Inselspital, Bern University Hospital, University of Bern, Bern, Switzerland

17- Micropathology Ltd, University of Warwick, Warwick, CV4 7EZ, United Kingdom

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

18- Division of Pediatric Immunology, Rheumatology and Infectious diseases, Emma Children's Hospital, Amsterdam Univiersyt Medical Center (Amsterdam UMC) Amsterdam, 1105 AZ, The Netherlands

19- Medical Research Council Unit at the London School of Hygiene & Tropical Medicine, Banjul, The Gambia

20- General Pediatrics, Infectious and Tropical Diseases Department, Hospital La Paz, 28046 Madrid, Spain

21- La Paz Research Institute (IdiPAZ), 28029 Madrid, Spain

22- Faculty of Medicine, Universidad Autónoma de Madrid (UAM), 28049 Madrid, Spain

23- Centro de Investigación Biomédica en Red en Enfermedades Infecciosas (CIBERINFEC), Madrid, Spain

24- Red de Investigación Traslacional en Infectología Pediátrica (RITIP), Madrid, Spain

25- Unidad de Cuidados Intensivos Pediátricos, Complejo Hospitalario de Jaen, Spain

26- Hospital Universitario Virgen de las Nieves, Servicio de Pediatría, Granada, Spain

27- Department of Pediatric Intensive Care Unit, Hospital Universitario de Burgos, Burgos, Spain

28- Instituto Hispalense de Pediatría, Instituto Balmis de Vacunas, Almería, Spain

29- Neonatal Intensive Care Unit, Complejo Asistencial Universitario de León, León, Spain

30- Department of Pediatrics, Complejo Hospitalario de Navarra, Servicio Navarro de Salud, Pamplona, Spain

31- IdiSNA (Instituto de Investigación Sanitaria de Navarra), Navarra Institute for Health Research, Pamplona, Spain

32- Infectious Diseases, Department of Pediatrics, Dr von Hauner Children's Hospital, University Hospital, LMU Munich, Munich, Germany

33- Division of Paediatrics, University Medical Centre Ljubljana and Medical Faculty, University of Ljubljana, Ljubljana, Slovenia

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

34- Children's Clinical University Hospital, Rīga Stradins University, Rïga, Latvia

35- Second Department of Paediatrics, National and Kapodistrian University of Athens (NKUA), School of Medicine, Panagiotis & Aglaia, Kyriakou Children's Hospital, Athens, Greece

36- Department of Pediatrics, Erasmus MC, Rotterdam, Netherlands

Link: https://www.nature.com/articles/s41467-025-55932-9

Abstract

Mycoplasma pneumoniae causes atypical pneumonia in children and young adults. Its lack of a cell wall makes it resistant to beta-lactams, which are the first-line treatment for typical pneumonia. Current diagnostic tests are time-consuming and have low specificity, leading clinicians to administer empirical antibiotics. Using a lasso regression simulation approach and blood microarray data from 107 children with pneumonia (including 30 *M. pneumoniae*) we identified 8 different transcriptomic signatures, ranging from 3-10 transcripts, that differentiate mycoplasma pneumonia from other bacterial/viral pneumonias with high accuracy (AUC: 0.84–0.95). Additionally, we demonstrated that existing signatures for broadly distinguishing viral/bacterial infections and viral/bacterial pneumonias were ineffective in distinguishing *M. pneumoniae* from viral pneumonia. The new signatures were successfully validated in an independent RNA-seq cohort of children with pneumonia, demonstrating their robustness. The high sensibility of these signatures presents a valuable opportunity to guide the treatment and management of *M. pneumoniae* pneumonia patients.

DOI: 10.1038/s41467-025-55932-9

JOURNAL: Nature Communications (ISSN: 2041-1723)

IMPACT FACTOR: **14.7** (2023)

CATHEGORY: MULTIDISCIPLINARY SCIENCE

RELATIVE POSSITION IN THE CATHEGORY: **8/134** in MULTIDISCIPLINARY SCIENCES (**Q1 and D1**).

OTHER QUALITY EVIDENCE: Nature Communication (ISSN: 2041-1723) It is an open-access, peer-reviewed scientific journal published by Nature Publishing Group since 2010. *Nature Communications* covers a wide range of natural sciences, including physics, chemistry, earth sciences, medicine, and biology. Rank by Journal Citation Indicator situate this journal in the **8/135** position **(D1)** for MULTIDISCIPLINARY SCIENCES. The journal had **21.622** citations in 2023.

PATENT APPLICATTION: There is a European Patent application derived from this work under application number No. 24 382 675.7.

AUTHOR CONTRIBUTION: In the present study, the doctoral student took part in the study conceptualization and design, computational model and machine learning approaches development, data analysis and writing process.

USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

nature communications

Article

https://doi.org/10.1038/s41467-025-55932-9

A diagnostic host-specific transcriptome response for *Mycoplasma pneumoniae* pneumonia to guide pediatric patient treatment

Received: 24 July 2024

Accepted: 5 January 2025

Published online: 15 January 2025

Check for updates

A list of authors and their affiliations appears at the end of the paper

Mycoplasma pneumoniae causes atypical pneumonia in children and young adults. Its lack of a cell wall makes it resistant to beta-lactams, which are the first-line treatment for typical pneumonia. Current diagnostic tests are time-consuming and have low specificity, leading clinicians to administer empirical antibiotics. Using a *LASSO* regression simulation approach and blood microarray data from 107 children with pneumonia (including 30 *M. pneumoniae*) we identify eight different transcriptomic signatures, ranging from 3-10 transcripts, that differentiate mycoplasma pneumonia from other bacterial/viral pneumonias with high accuracy (AUC: 0.84–0.95). Additionally, we demonstrate that existing signatures for broadly distinguishing viral/bacterial infections and viral/bacterial pneumonias are ineffective in distinguishing *M. pneumoniae* from viral pneumonia. The new signatures are successfully validated in an independent RNAseq cohort of children with pneumonia, demonstrating their robustness. The high sensibility of these signatures presents a valuable opportunity to guide the treatment and management of *M. pneumoniae* pneumonia patients.

Mycoplasma pneumoniae commonly causes respiratory infections in children and young adults, presenting with clinical manifestations ranging from asymptomatic to pneumonia. *M. pneumoniae* is one of the most common causes of atypical pneumonia. The incidence of *M. pneumoniae* pneumonia is uncertain and is typically described as endemic, marked by cyclic epidemics occurring every 3 to 5 years. The COVID-19 pandemic, along with associated non-pharmaceutical interventions, had a significant impact on the circulation of respiratory pathogens, including *M. pneumoniae*[1]. A global surveillance network noted a decrease in *M. pneumoniae* detection from April 2020 to March 2021[2]. Despite a resurgence in other pathogens, *M. pneumoniae* infections remained low until March 2022-23. Concerns about a potential upsurge in mycoplasma infections have arisen due to waning herd immunity[3,4]. Recent outbreaks of *M. pneumoniae* infections have been reported during the last months in different European countries, including Spain[4], Denmark[5], France[6], and the Netherlands[7] (see also https://edition.cnn.com/2023/12/01/health/mycoplasma-pneumonia/index.html). In November 2023, the World Health Organization (WHO) reported an increase in outpatient consultations and hospitalizations for *M. pneumoniae* pneumonia in China since May, along with rises in respiratory syncytial virus (RSV), adenovirus, and influenza virus cases since October[1,3,8].

Establishing the microbiological etiology of pneumonia remains a challenge for clinicians within healthcare facilities, primarily due to the complexity of obtaining direct microbiological samples, which often require invasive procedures. Additionally, there is a delay in obtaining results from other indirect tests, such as pneumococcal antigen in urine, PCR in respiratory swabs, or serology, when available. In the particular case of *M. pneumoniae*, cultures are considered the gold-standard for bacterial diagnosis; however, due to the difficulty in

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

e-mail: antonio.salas@usc.es

obtaining representative samples, slow growth rates, and low sensitivity, culture methods are impractical for timely clinical and treatment decision-making[9]. Serology is another commonly used approach for diagnosing *M. pneumoniae*[9], but tests show low specificity. IgM antibodies can remain elevated for months after infection, with frequent cross-reactivity with other pathogens but also potential for generating false positives[10]. As a result, serology has gradually been replaced by PCR[11], a rapid method offering higher sensitivity and specificity compared to serology and culture. PCR is now widely used in developed countries for the diagnosis of *M. pneumoniae*. Specific diagnosis of *M. pneumoniae* is important since, due to its lack of a cell wall, it does not respond to beta-lactams antibiotics such as amoxicillin, which is the first-line treatment for typical community acquired pneumonia (CAP). Azithromycin is the first line of treatment instead[12]. Thus, clinicians often initiate empirical antibiotic treatment and adjust it later based on microbiological test results[13].

Recent advances in transcriptomics have allowed for the investigation of specific signatures triggered by the host in response to infections by viruses, bacteria[14,15] and different pathogens[16]. Given the outbreaks of *M. pneumoniae* and the lack of precise and clinically useful diagnostic tools, it is essential to explore the specificity of the gene expression response against *M. pneumoniae* and other microorganisms responsible for pneumonia. While there have been a few attempts to characterize the gene expression profile in pneumonia patients infected by different microorganisms, none of them have focused specifically on the atypical pneumonia caused by *M. pneumoniae*. However, we noted that a transcriptomic microarray study by Wallihan et al.[17] included 30 children infected by *M. pneumoniae*, although a specific blood transcriptomic signature for this infection was not investigated.

Considering the urgent need prompted by recent outbreaks, it may be helpful to explore host gene expression mechanisms specifically triggered by *M. pneumoniae* and evaluate them in the context of those generated from pneumonia caused by other pathogens. Delivering a host-transcriptome signature specific for *M. pneumoniae* infection would be desirable to overcome the limitations of current clinical and non-clinical procedures.

In this study, we identify and validate eight distinct transcriptomic signatures, consisting of 3 to 10 genes each, that can accurately differentiate *M. pneumoniae* pneumonia from other bacterial and viral pneumonias in the pediatric population. These signatures represent a significant advancement over previous transcriptomic models, enhancing diagnostic accuracy and offering valuable insights for clinical decision-making.

Results

A robust multi-cohort study design incorporating discovery and validation phases across different platforms was implemented to derive optimal transcriptomic signatures for the specific diagnosis of *M. pneumoniae*. The overall design is summarized in Fig. 1. Initially, differential gene expression was analyzed between *M. pneumoniae* and viral datasets from[17] (step 1 in Fig. 1). The most promising differentially expressed genes (DEGs) were then used in downstream analyses to identify the best transcriptomic signatures through a systematic resampling approach and penalized regression methods (steps 2–7 in Fig. 1). The top *M. pneumoniae* signatures were rigorously evaluated for their ability to discriminate between *M. pneumoniae* and viruses, as well as other pathogens, coinfections, and healthy controls[17] (step 8 in Fig. 1). Additionally, the performance of these signatures was benchmarked against other published signatures to compare their discriminative power (step 9 in Fig. 1). Finally, the leading signatures were validated in an independent RNAseq cohort of pneumonia patients, confirming their robustness and clinical utility (step 10 in Fig. 1). Further methodological details are provided in Fig. 1 and the Materials and Methods section.

Transcriptomic analysis

We conducted a principal component analysis (PCA) on a subset of the 500 most variable genes with the entire pneumonia cohort, which comprised four different groups: bacterial infection, viral infection, *M. pneumoniae* infection and co-infections. The first principal component (PC1), explaining 16.85% of the variation, clearly shows the segregation of the *M. pneumoniae* samples from pneumonias caused by other bacteria. As expected, co-infected samples are distributed evenly between the two groups. Viral profiles intermingle with other profiles in the plot, including *M. pneumoniae* and bacterial transcriptomes; Fig. 2A. PC2 (accounting for 8.16% of the variation) makes a subtle distinction between bacterial infections on one pole of the component, with most co-infections included in this cluster. *M. pneumoniae* samples are slightly displayed towards the center of the component, while viral samples are evenly distributed along the whole component, reproducing the same behavior as in PC1 (Fig. 2A).

In a comparative analysis of transcriptomes in children with pneumonia caused by *M. pneumoniae* vs. viral pneumonia, we identified 3783 DEGs, using a False Discovery Rate (FDR) significance threshold of 5% (Supplementary Data 1). Among these DEGs, 2288 were found to be upregulated, while 1495 were downregulated (Fig. 2B).

Diagnostic signature discovery

To identify the best minimal signatures for differentiating *M. pneumoniae* infection from viral infections, we constructed a *LASSO* model using the top 276 DEGs, applying the criteria of adjusted *P*-value < 0.01, $Log_2FC > |1|$, and Log_2 Average Expression > 2 (Supplementary Data 1). Subsequently, we generated 1000 different transcriptomic signatures using the 1000 randomized training datasets (TA). We then calculated the area under the receiver operating characteristic curve (AUC) for each signature in the TA dataset, TE dataset and the EUCLIDS cohort (non-mycoplasma bacterial and viral pneumonias). Among the 1000 signatures, we computed the Signature Scores (SS) to obtain the best signatures of size 3 to 10. Eight transcriptomic signatures emerged as the best candidates for distinguishing pneumonia caused by *M. pneumoniae* from viral pneumonia (Table 1). These signatures include genes in common, totaling 18 different transcripts, with 15 over-expressed and 3 under-expressed in mycoplasma pneumonias compared to pneumonias from viral etiology (Supplementary Fig. 1). Notably, all these signatures demonstrated reliable performance when employed to compare *M. pneumoniae* samples against all non-mycoplasma samples, including cases of co-infections and other bacterial pneumonias (Supplementary Data 2).

We assessed the performance for each of the eight selected signatures to differentiate between *M. pneumoniae* and viral pneumonia across all randomized datasets (TA), excluding those used to generate each signature (therefore, 999 subsets). The median AUC values ranged from 0.84 (for the 3-transcript signature) to 0.95 (for the 10-transcript signature) (Fig. 3; Supplementary Data 2), demonstrating the overall high-accuracy of the signatures to discriminate between both phenotypes. We set the optimal cut-off for each signature to optimize the discrimination of the two categories. Additionally, we evaluated each transcriptomic signature in the complete dataset to test their performance in differentiating mycoplasma pneumonias from all other pneumonias, represented with a receiver operating characteristic curve (ROC) and their respective AUC values. (Fig. 3; Supplementary Data 2). Equally, the AUC values from this analysis aligned with the AUC median values observed across all TA datasets.

We further tested the accuracy of the signatures in discriminating between *M. pneumoniae* pneumonia from other clinically relevant groups, including: *i*) *M. pneumoniae* pneumonia vs. pneumonias without a detected pathogen, *ii*) *M. pneumoniae* pneumonia in

https://doi.org/10.1038/s41467-025-55932-9

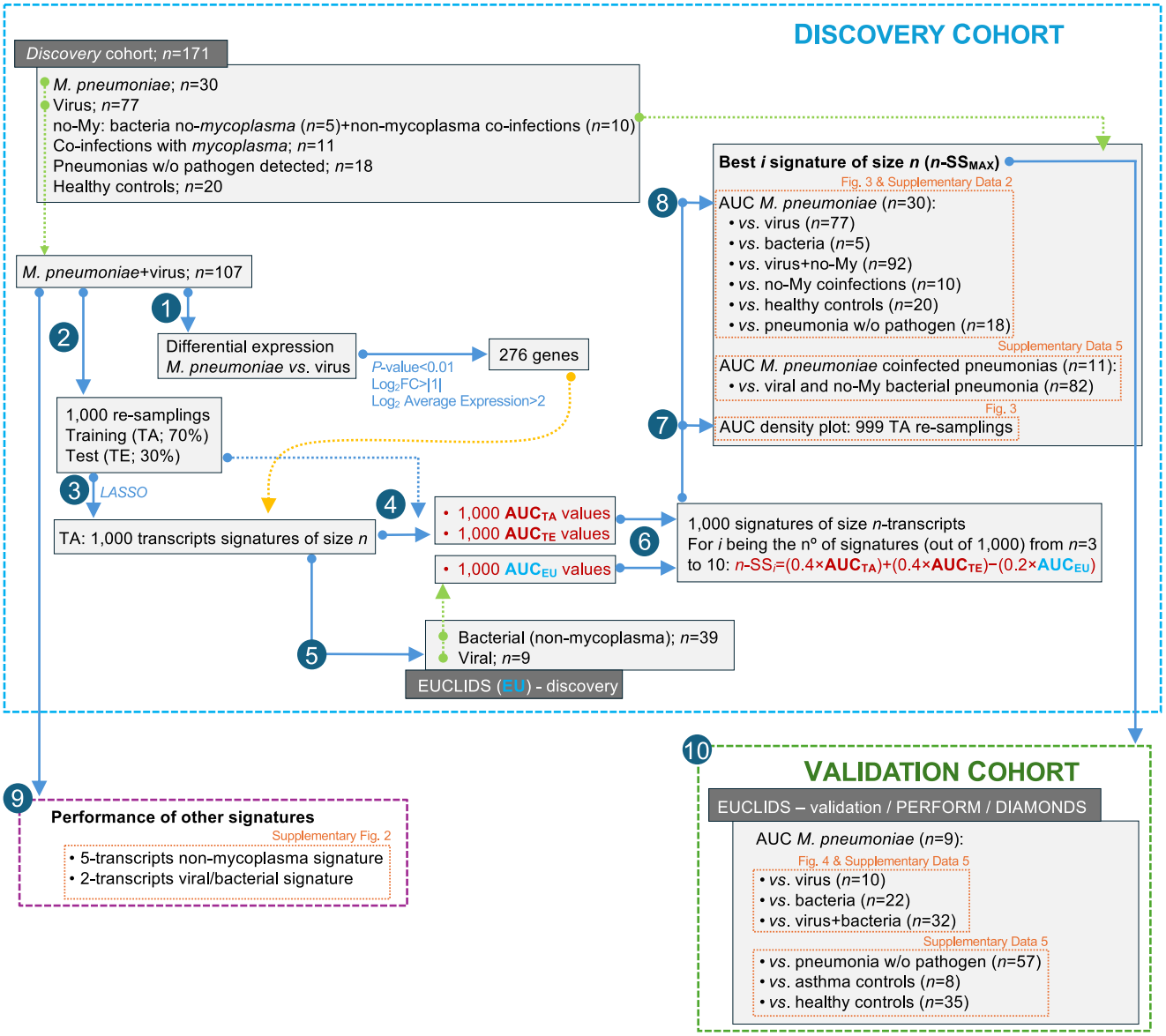


Fig. 1 | Overall study design. Differential expression was investigated between *M. pneumoniae* and virus datasets from[17] (step 1). We employed a re-sampling approach (step 2) to generate 1000 transcript signatures of size *n* (step 3). These signatures were utilized to compute 1000 AUC values on both the training (TA) and test (TE) datasets (step 4), as well as on the non-mycoplasma bacteria and virus datasets from the EUCLIDS (discovery) cohort (step 4 and 5). Next, Signature Scores (SS) were obtained for the different *n*-transcript signatures (from $n = 3$ to 10) based on AUC_{TA}, AUC_{TE}, and AUC_{EU} values (step 6), allowing to derive the best signature of size *n* (step 7); $n\text{-}SS_{MAX}$. AUC values were then computed with the best *n*-transcript signatures for the comparison between *M. pneumoniae* vs. virus and *M. pneumoniae* vs. virus + other pathogens from ref. 17 (step 8). For comparative purposes, we also assessed the performance of other available signatures in the literature (step 9). Finally, the best signatures were subsequently validated in the EUCLIDS (validation cohort), DIAMONDS, and PERFORM cohorts (step 10).

coinfection with other pathogens vs. all other pneumonias; and *iii*) *M. pneumoniae* pneumonia vs. healthy controls. Notably, the signatures performed exceptionally well in distinguishing *M. pneumoniae* from healthy controls, with AUC values ranging from 0.87 to 0.98, and from pneumonias without a detected pathogen, with AUC values ranging from 0.81 to 0.90; Supplementary Data 2. As expected, the accuracy of the signatures decreases when tested against pneumonias caused by *M. pneumoniae* in co-infection with other pathogens, compared to viral and bacterial pneumonias; Supplementary Data 2. However, the 9-transcript signature still yielded AUC values of 0.77 or higher for all comparisons: *M. pneumoniae* in co-infection vs. viral + bacterial pneumonias (AUC = 0.77 [CI:0.65–0.89]), vs. bacterial pneumonias (AUC = 0.80 [CI:0.57–1.00]) and vs. viral pneumonias (AUC = 0.77 [CI:0.65–0.89]). Interestingly, the 3-transcript signature also showed AUC values above 0.70 in *M. pneumoniae* co-infection scenarios, achieving its highest performance when distinguishing from bacterial pneumonias (AUC = 0.83 [CI:0.63–1.00]).

Validation of *M. pneumoniae* signatures in an independent cohort

The diagnostic accuracy of the proposed RNA signatures was evaluated using additional gene expression data generated from a new pediatric cohort of pneumonia samples and a different technology (RNAseq); see details in Supplementary Data 3 and Supplementary Data 4 The results confirmed that all RNA signatures can discriminate between viral and *M. pneumoniae* pneumonias with AUCs higher than 0.68 (7-transcript), being the 9-transcript signature the most predictive (AUC = 1.00 [CI: 1.00–1.00]; sensitivity: 1.00; specificity: 1.00); Fig. 4, Supplementary Data 5. Unexpectedly, the third most predictive model was the 3-transcript signature (AUC = 0.86 [CI: 0.65–1.00]; sensitivity: 0.89;

https://doi.org/10.1038/s41467-025-55932-9

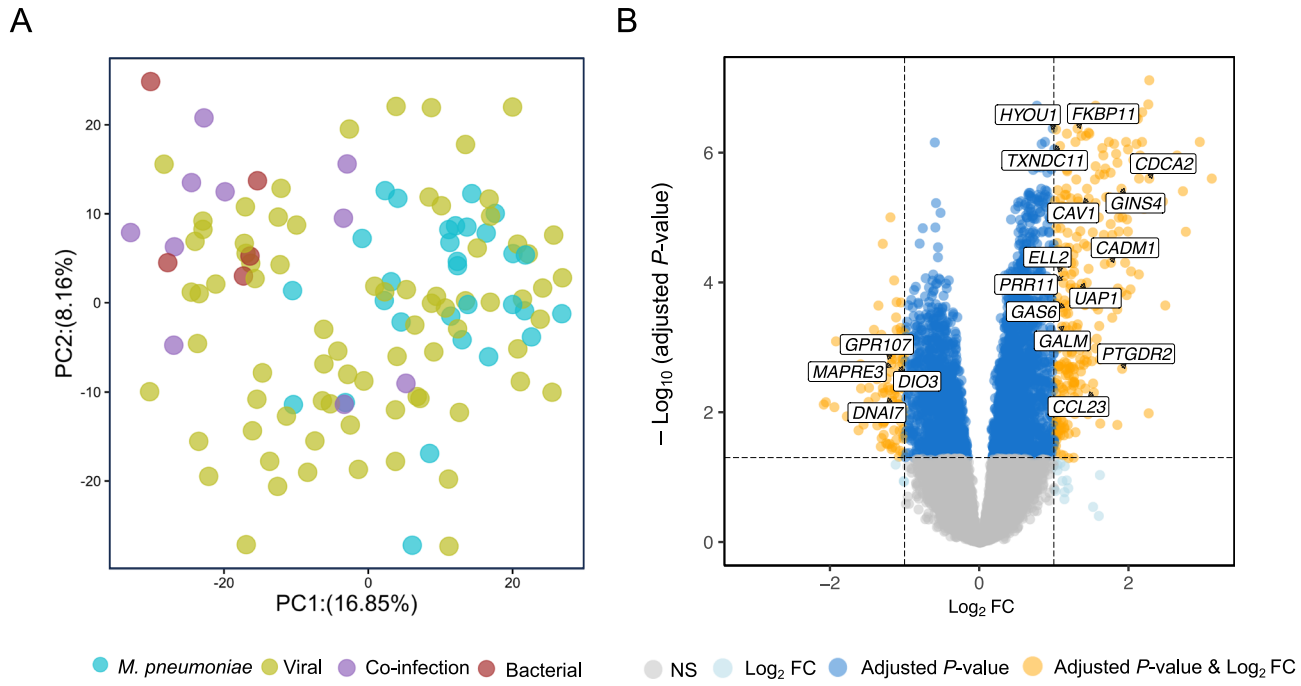


Fig. 2 | PCA of transcriptomic profiles of different pneumonia cohorts and differential expression analysis. A PCA of transcriptome profiles from blood samples of viral, *M. pneumoniae*, co-infections or other bacteria pneumonia infections. The two first principal components (PC1 and PC2) are shown. **B** Volcano plot showing the DEGs between conditions: *M. pneumoniae* pneumonia vs. viral pneumonia. The statistical test used to obtain *P*-values was a two-sided moderated t-test implemented in the *limma* package. Adjustments for multiple comparisons were performed using the Benjamini-Hochberg procedure to control the false discovery rate (FDR). The plot shows the $\log_2$ fold change (Log_2FC) on the x-axis and the $-\log_{10}$ adjusted *P*-value on the y-axis. Genes with an adjusted *P*-value < 0.05 and a $\log_2\text{FC} > |1|$ were colored in orange, genes with an adjusted *P*-value < 0.05 but a $\log_2\text{FC} < |1|$ were colored in dark blue, genes with an adjusted *P*-value > 0.05 but a $\log_2\text{FC} > |1|$ were colored in light blue and non-significant genes with $\log_2$ FC < |1| were colored in grey. Only genes included in the final signatures were labeled in the graphic.

specificity: 0.90) (Fig. 4, Supplementary Data 5). Remarkably, all signatures can also accurately differentiate *M. pneumoniae* pneumonia from other bacterial pneumonias, with AUCs ranging from 0.74 (CI: 0.55–0.92) for the 4-transcript signature to 0.85 (CI: 0.73–0.99) for the 10-transcript signature. Considering all viral and non-mycoplasma bacterial pneumonias against *M. pneumoniae* pneumonias, the AUC values were similar to those obtained from the comparison *M. pneumoniae* vs. bacterial pneumonia. In this case, the 9-transcript signature yielded the best performance of all the signatures tested (AUC = 0.89 [CI: 0.79–0.99]); Fig. 4, Supplementary Data 5.

As in the case of the discovery dataset, we determined the optimal cut-off for each signature in the validation that establishes the threshold to discriminate between pneumonias from different etiologies in the three comparisons groups (Fig. 4; Supplementary Data 5).

Additionally, we also validated the eight signatures in other groups, some of which are clinically relevant, including pneumonias diagnosed without a detected pathogen, controls with asthma and healthy controls. The AUC values obtained were: *i*) from 0.76 (4-transcripts signature) to 0.85 (9-transcripts signature) for the contrast *M. pneumoniae* pneumonia vs. pneumonias diagnosed without a detected pathogen, *ii*) from 0.68 (7-transcript signature) to 0.86 (6-transcripts signature) for the contrast *M. pneumoniae* pneumonia vs. controls with asthma, and *iii*) from 0.83 (4-transcripts signature and 9-transcripts signature) to 0.96 (8-transcripts signature) for the contrast *M. pneumoniae* pneumonia vs. healthy controls (Supplementary Data 5).

Performance of other available signatures in *M. pneumoniae* samples

As *M. pneumoniae* pediatric pneumonias appear to induce a different alteration in the transcriptome compared with other bacterial infections causing pneumonia in children, we opted to examine the performance of two available signatures designed to differentiate viral and bacterial pediatric infections in samples from children with viral pneumonia and pneumonia caused by *M. pneumoniae*: the two-transcript signature (*IFI44L*/*FAM89A*), capable of differentiating between pediatric viral from bacterial infections[14]; and the novel 5-transcript signature[18], designed to specifically distinguish between viral and bacterial CAP in children. We observed that neither of these two tested signatures could effectively diagnose *M. pneumoniae* as a bacterial infection or a bacterial pneumonia, yielding AUC values of 0.56 and 0.52, respectively (Supplementary Fig. 2).

Differentially regulated pathways between *M. pneumoniae* and viral pneumonia

To detect biological pathways responsible for the different response to *M. pneumoniae* and viral pneumonia in our study cohort, we performed a GSVA analysis directly on gene expression data. We identify 525 significantly (adjusted *P*-value < 0.05) differentially regulated pathways (DRPs) between both categories, with over half upregulated in *M. pneumoniae* pneumonia (358/525; 68%); Supplementary Data 6; Fig. 5A. Among the top 20 pathways, the most notable changes in pathways activity were mainly represented by up-regulated processes in atypical pneumonia (18/20; 90%); Supplementary Fig. 3.

The most DRP was "SRP dependent co-translational protein targeting to membrane signal sequence recognition" (adjusted *P*-value = 1×10^{-06}) followed by "interleukin-12 mediated signaling" pathways (adjusted *P*-value = 2×10^{-08}), and a group of biological routes related to polysaccharides and glycolipids synthesis, including "nucleotide sugar biosynthetic process" (adjusted *P*-value = 9×10^{-05}), "UDP N-Acetylglucosamine metabolic process" (adjusted *P*-value = 9×10^{-05}), "amino sugar biosynthetic process" (adjusted *P*-value = 2×10^{-05}), "UDP N-Acetylglucosamine biosynthetic process"

Table 1 | Model coefficients and weights (%) for the genes included in each of the signatures

n	Gene Signatures (coefficients; weight in %)
3	*FKBP11* (1.246; 32.8%), ***TXNDC11*** (1.380; 36.3%), ***DIO3*** (−1.172, 30.9%). Intercept: −16.662
4	*GAS6* (0.387,11.76%), *HYOU1* (1.354, 41.13%), *CAV1* (0.623, 18.93%), ***DIO3*** (−0.928, 28.19%). Intercept: −10.676
5	*DNAI7* (−0.595; 13.54%), ***GALM*** (0.714; 16.25%), *GINS4* (0.419; 9.54%), ***TXNDC11*** (2.038; 46.39%), ***DIO3*** (−0.627; 14.27%). Intercept: −13.747
6	*PRR11* (0.905; 29.94%), *DNAI7* (−0.298; 9.87%), ***GALM*** (0.793; 26.25%), *GINS4* (0.260; 8.61%), ***TXNDC11*** (0.491; 16.26%), *CAV1* (0.274; 9.06%). Intercept: −14.197
7	*CCL23* (0.624; 14.16%), ***GALM*** (0.910; 20.65%), ***TXNDC11*** (1.400; 31.77%), *ELL2* (−0.009; 0.20%), ***DIO3*** (−0.839; 19.04%), *UAP1* (−0.211; 4.79%), *CDCA2* (0.414; 9.39%). Intercept: −13.914
8	*CCL23* (0.545; 10.06%), *PRR11* (0.974; 17.99%), *GAS6* (0.016; 0.29%), ***GALM*** (1.012; 18.69%), ***TXNDC11*** (0.980; 18.10%), *CAV1* (0.001; 0.02%), ***DIO3*** (−1.196; 22.09%), *CADM1* (0.691; 12.76%). Intercept: −0.187
9	*CCL23* (0.396; 8.25%), *PRR11* (0.926; 19.30%), *DNAI7* (−0.412; 8.59%), ***GALM*** (0.319; 6.65%), *GINS4* (0.315; 6.56%), ***TXNDC11*** (0.053; 1.10%), *CAV1*(0.952; 19.84%), ***DIO3*** (−1.163; 24.24%), *PTGDR2* (0.262; 5.46%). Intercept: −10.163
10	*CCL23* (0.374; 2.52%), *PRR11*(1.316; 8.86%), *DNAI7* (−1.146; 7.72%), ***GALM*** (2.091; 14.08%), *GINS4* (1.136; 7–65%), ***TXNDC11*** (2.901; 19.54%), *MAPRE3* (−0.823; 5.54%), ***DIO3*** (−3.411; 22.97%), *GPR107* (−0.876; 5.90%), *CADM1* (0.773; 5.21%). Intercept: −28.067

Genes *DIO3*, *GALM*, and *TXNDC11* (shown in bold) appear in at least five *n*-transcript signatures.
n number of transcripts in the optimal signature.

(adjusted *P*-value = 4×10^{-05}) and "GDP-mannose metabolic process" (adjusted *P*-value = 9×10^{-05}); Supplementary Fig. 3; Supplementary Data 6.

Afterwards, we investigated the functional involvement of the genes included in the transcriptomic signatures by examining the DRPs related to these genes. We found that 10 out of the 18 genes (*UAP1*, *PTGDR2*, *CAV1*, *HYOU1*, *GALM*, *GAS6*, *GINS4*, *CDCA2*, *MAPRE3*, *DIO3*; Supplementary Data 7, Fig. 5B) participate in 9 significantly DRPs. Notably, among these pathways, "the nucleotide sugar biosynthetic process" emerged as one of the most significant pathways in the overall analysis (adjusted *P*-value = 9×10^{-05}), Fig. 5B. Additionally, the involvement of two of the predictive genes in the same DRP (*CAV1* and *HYOU1*) was only detected for the "response to endoplasmic reticulum stress process" (adjusted *P*-value = 0.002).

Discussion

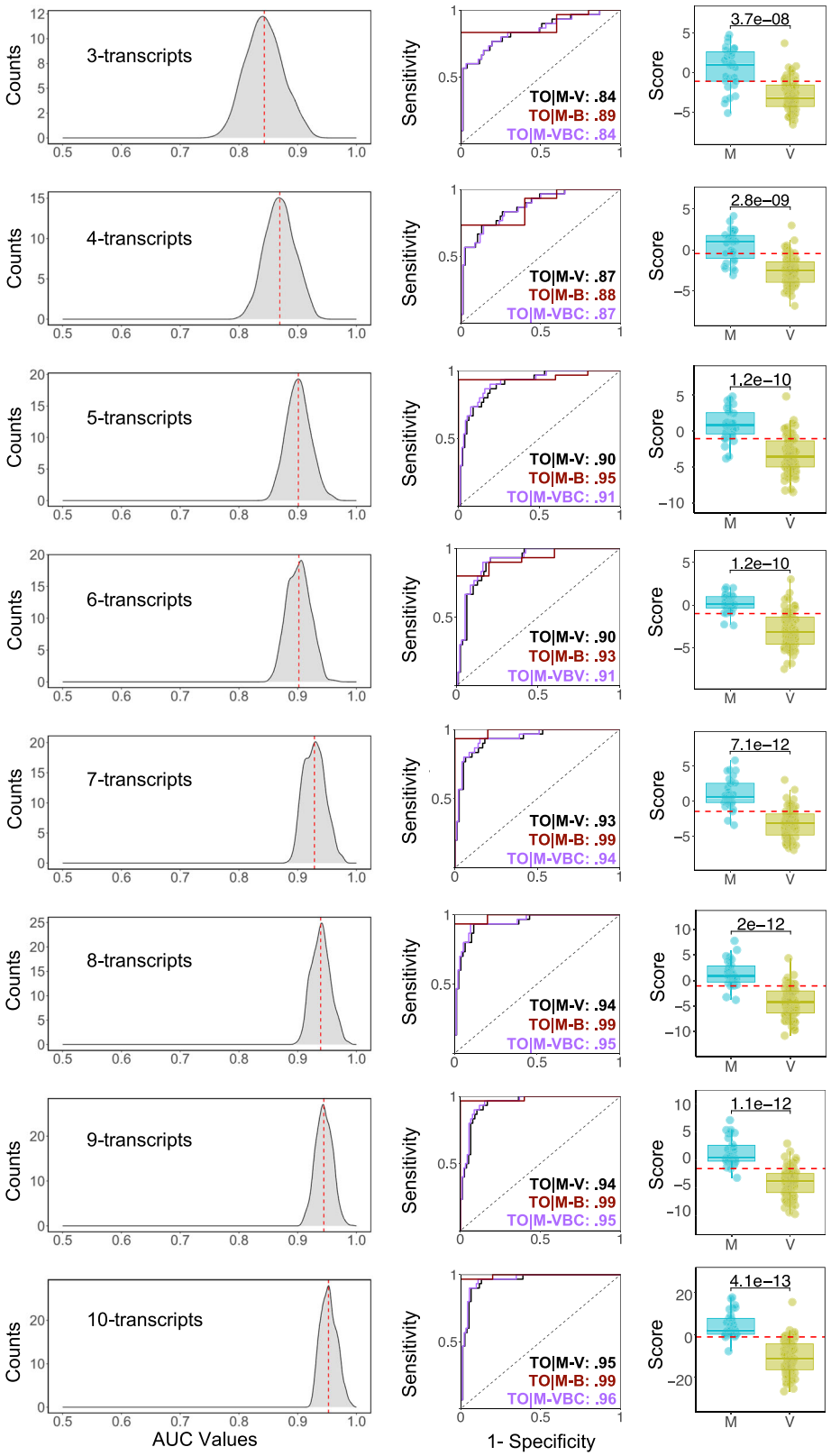
Pneumonia etiology in children is highly diverse and typically challenging to effectively assess in the clinical practice. The increase of cases of *M. pneumoniae* from May 2023 raised concerns within the clinical community. Infections caused by this pathogen require specific treatment protocols, as they do not respond to standard antibiotic prescriptions typically employed when bacterial pneumonia is suspected[12,13]. Current diagnostic tests can suffer from long turnaround times and may have limited sensitivity and specificity. PCR-based diagnostic tests for mycoplasma detection, while highly sensitive and specific, have become more widely used in developed countries, especially following the SARS-CoV-2 pandemic. However, PCR tests are ineffective to discern between a *M. pneumoniae* active infection and asymptomatic carriage, which is common in children's upper respiratory tract[19]. They also struggle in cases of coinfections, where *M. pneumoniae* may not be the primary pathogen. Given that *M. pneumoniae* pneumonia requires tailored treatment, misdiagnosis can have serious consequences for patient care. A host-based diagnosis using specific transcriptomic signatures can help to address the limitations of the pathogen-based tests by improving the ability to distinguish between carriage and active infection in *M. pneumoniae* pneumonia cases. Our analysis shows that transcriptomic-based signatures available in the literature, previously proven to be effective to broadly differentiate between viral and bacterial infections or specifically pneumonia from viral and bacterial etiology, do not function properly for discriminating *M. pneumoniae* pneumonia cases from viral pneumonias in children. Our findings also indicate that *M. pneumoniae* pneumonia alters the expression of >3000 genes in blood of affected children, enabling the inference of eight new specific signatures capable of specifically differentiating pneumonia caused by *M. pneumoniae* from viral pneumonia. Notably, the performance of these signatures does not decline when used to discriminate between *M. pneumoniae* infections from other bacterial pneumonias or non-mycoplasma co-infections, ensuring high specificity in detecting *M. pneumoniae* pneumonias. The identification of these specific transcriptomic signatures capable of diagnosing *M. pneumoniae* pneumonias in children holds promise for enhancing the diagnosis and treatment for these patients, who cannot be identified using the currently available signatures. Most notably, these signatures could guide the decision to administer a specific macrolide instead of a beta-lactam antibiotic. As all the *n*-transcript signatures exhibit high sensitivities, the test would ensure the detection of most *M. pneumoniae* infected patients and improve their management.

Employing an innovative simulation-driven theoretical framework, we uncovered eight distinct transcriptomic signatures, each representing the optimal combinations of 3 to 10 transcripts, capable of differentiating *M. pneumoniae* pneumonia from other bacterial and viral pneumonias with high accuracy (AUC: 0.84–0.95). These also signatures demonstrated strong discriminatory power in other clinically relevant groups, such as coinfections (with or without *M. pneumoniae*), pneumonias without a detected pathogen, and healthy controls. This broad applicability makes the signatures valuable not only for distinguishing *M. pneumoniae* pneumonia from viral pneumonia but also for a variety of common clinical scenarios.

The robustness of these signatures was confirmed in an independent validation cohort, which included viral and bacterial pneumonias, pneumonias without a detected pathogen, controls with asthma and healthy controls. While our theoretical model (Fig. 3) predicted that signature performance would improve with a greater number of predictors, in the validation cohort, the 9-transcript signature exhibited the best performance, but the 3-transcript signature also displayed unexpectedly high accuracy.

The choice and usefulness of these signatures would completely depend on different factors of the clinical context: available equipment, budget, number of samples, urgency for results, etc. Few technologies can accurately measure that number of transcripts simultaneously at a reasonable cost and in a timely fashion. For instance, RNA sequencing is not a suitable choice because, while it is time- and cost-effective for high-throughput gene expression, it is impractical when studying only a limited number of target genes. However, NanoString (www.NanoString.com/) offers high precision and is much more convenient when targeting a limited number of genes. Unfortunately, these platforms are often reserved for research purposes and are not typically found in the daily laboratory routine of medical centers. Conventional RT-qPCR is the gold standard for these

https://doi.org/10.1038/s41467-025-55932-9



cases, but can be time-consuming, laborious, and less cost-effective, especially when the number of target genes and the sample size increase. New PCR approaches, such as 1-step digital PCR (dPCR), allow for the combination of up to 5 genes, with a high sensibility and bypassing certain steps[20], and would be an option to study the biomarkers of some of the proposed signatures. On the other hand, reducing the number of predictors increases the likelihood of transferability to a point-of-care (POC) device, and the signatures with the lower number of transcripts are the most promising candidates for implementation in this type of tests, specially the 3-transcript signature, which shows remarkably accurate performance compared to other signatures with a larger number of predictors. The use of

https://doi.org/10.1038/s41467-025-55932-9

Fig. 3 | Transcriptomic signatures obtained from the discovery cohort. AUCs and ROC curves from the density plots (left panels) of the AUC values computed on the 999 training resamples and using the optimal *n*-transcript signature; red dashed vertical lines in the left panels indicate the median values. ROC curves and AUC values (central panels) for the total cohort of *i) M. pneumoniae* ($n = 30$) vs. viral pneumonia ($n = 77$) (black line; $AUC_{TO|M\text{-}V}$), *ii) M. pneumoniae* pneumonias ($n = 30$) vs. Bacterial pneumonias ($n = 5$) (red line; $AUC_{TO|M\text{-}B}$) and *iii) M. pneumoniae* pneumonias ($n = 30$) vs. all non-mycoplasma pneumonia infections (including, virus, non-mycoplasma bacteria, and non-mycoplasma co-infections; $n = 92$) (violet line; $AUC_{TO|M\text{-}VBC}$). Boxplots of the predicted values using each optimal model in the total cohort with two-sided Wilcoxon rank-sum test *P*-values (right panels). Red dashed line represents the optimal cutpoint. The boxes are defined by the upper and lower quartile (Q1 and Q3); the median is shown as a bold-colored horizontal line; whiskers extend to the most extreme data point which is no more than 1.5 times the interquartile range (IQR) from the box. VBC all non-mycoplasma pneumonias (including, virus, non-mycoplasma bacteria, and non-mycoplasma co-infections), B bacterial, M *M. pneumoniae*, TO total sample, V viral.

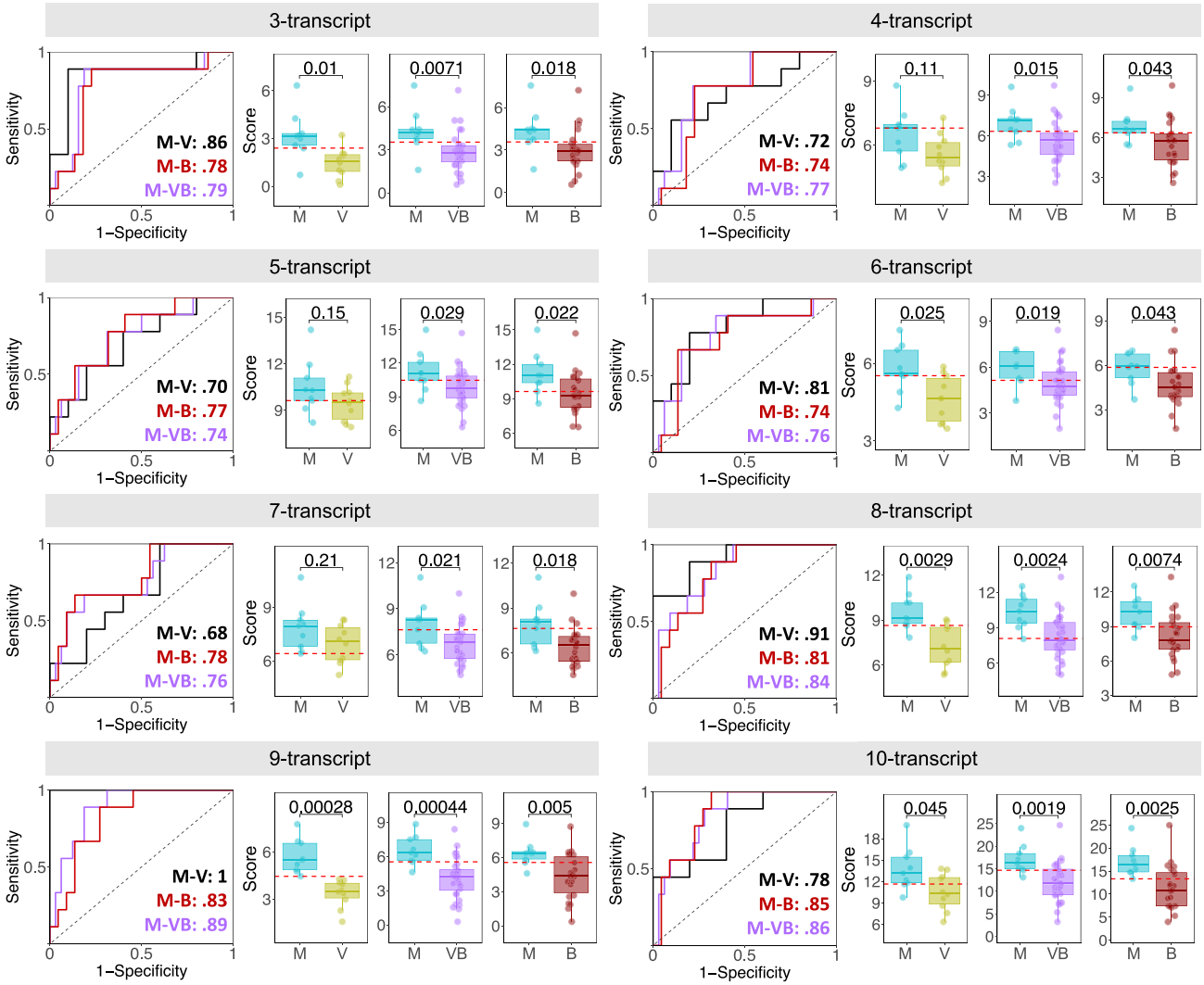


Fig. 4 | Validation of the transcriptomic signatures. ROC curves with AUC values for the three tested validation subsets: *M. pneumoniae* ($n = 9$) vs. viral pneumonias ($n = 10$) (black line; $AUC_{M\text{-}V}$), *M. pneumoniae* ($n = 9$) vs. non-mycoplasma pneumonias including viral and other bacterial pneumonias ($n = 32$) (violet line; $AUC_{M\text{-}VB}$) and *M. pneumoniae* vs. non-mycoplasma bacterial pneumonias ($n = 22$) (red line; $AUC_{M\text{-}B}$) (left panels). Boxplots of the predicted values using each model in the three validation subsets (right panels). Two-sided Wilcoxon *P*-values are also displayed. The red dashed line represents the optimal cut off. The boxes represent the distribution of scores for *M. pneumoniae*, viral and bacterial pneumonia samples. They are defined by the upper and lower quartiles (Q1 and Q3), with the median displayed as a bold-colored horizontal line. The whiskers extend to the most extreme data points within 1.5 times the interquartile range (IQR) from the box. Abbreviations as in legend of Fig. 3. VB all non-mycoplasma pneumonias including viral and other bacterial pneumonias, B bacterial, M *M. pneumoniae*, V viral.

compact signatures would facilitate the translation to a rapid diagnostic POC test enhancing its widespread adoption and leading to reduced costs and time.

Differential pathway analysis also reveals different processes that are over-activated in *M. pneumoniae* compared to viral pneumonias causing pneumonia. Many of these significantly altered pathways are related to nucleotide sugar metabolism, which is closely associated with human infections due to its involvement in surface carbohydrate biosynthesis, protein glycosylation, and glycosylated lipid biosynthesis; all essential processes for pathogens' ability to infect and survive within the human host[21–23]. Nucleotide sugars function as glycosyl donors for glycosylation reactions in proteins catalyzed by the

https://doi.org/10.1038/s41467-025-55932-9

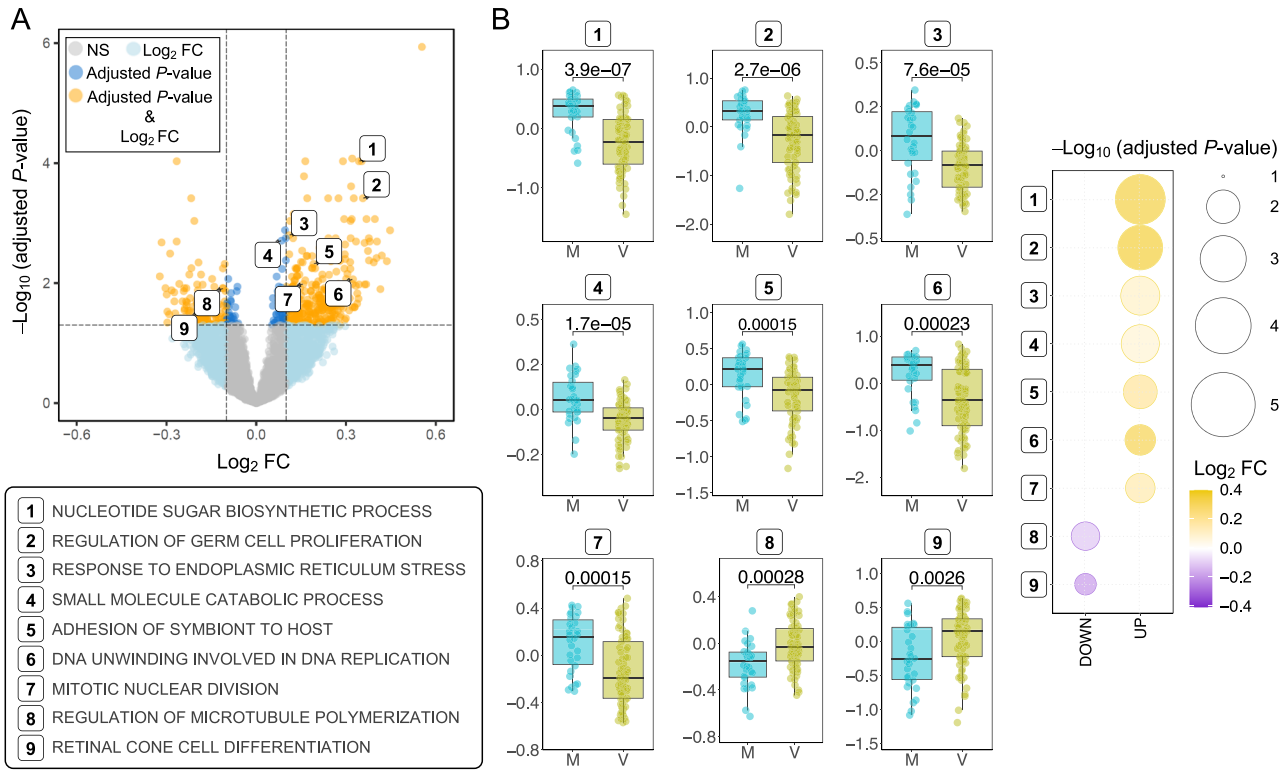


Fig. 5 | Differential pathways analysis. A Volcano plot showing the genes in DRPs between conditions: *M. pneumoniae* pneumonia ($n = 30$) vs. viral pneumonia ($n = 77$). The statistical test used to obtain *P*-values was a two-sided moderated t-test implemented in the *limma* package. Adjustments for multiple comparisons were performed using the Benjamini-Hochberg procedure to control the false discovery rate (FDR). The plot shows the $\log_2$ fold change (Log_2FC) on the x-axis and the $-\log_{10}$ adjusted *P*-value on the y-axis. Pathways meeting the criteria of an adjusted *P*-value < 0.05 and a $\log_2FC > |0.1|$ were colored in yellow, pathways with an adjusted *P*-value < 0.05 but a $\log_2FC < |0.1|$ were colored in dark blue, genes with an adjusted *P*-value > 0.05 but a $\log_2FC > |0.1|$ were colored in light blue and non-significant genes were colored in grey. **B** Boxplots representing the pathway activity of the DRPs for the *M. pneumoniae* ($n = 30$) and viral pneumonia ($n = 77$) groups in which genes from the transcriptomic signatures are involved (left panel). Two-sided Wilcoxon rank-sum test *P*-values are displayed. The boxes are defined by the upper and lower quartiles (Q1 and Q3), with the median displayed as a bold-colored horizontal line. The whiskers extend to the most extreme data points within 1.5 times the interquartile range (IQR) from the box. Bubble plot of the DRPs in which genes included in the transcriptomic signatures are involved (right panel). Numeric identifies indicate DRPs in which genes from the transcriptomic signatures are involved. The statistical test used to obtain *P*-values was a two-sided moderated t-test implemented in the *limma* package. Adjustments for multiple comparisons were performed using the Benjamini-Hochberg procedure to control the false discovery rate (FDR). Size of the points is proportional to the $-\log_{10}$ adjusted *P*-value and color bar indicates the $\log_2FC$ values.

glycosyltransferases. Mycoplasmas, with their small genome, cannot synthesize some essential biomolecules de novo, relying on obtaining them from the host with which they live in close association. UDP-N-acetylglucosamine (GlcNAc), a nucleotide sugar, is one of the major carbon sources utilized by several human pathogens, including *M. pneumoniae*, and its catabolism is also involved in the regulation of virulence and pathogenesis[24]. *M. pneumoniae* has been reported to produce biofilm on the epithelial surface, contributing to evasion of host immune system and providing resistance to antibiotics[25,26]. Biofilms, complex communities of microorganisms, typically bacteria, encased within a self-produced extracellular matrix, can contribute to persistent infections[27]. Polysaccharides containing GlcNAc contribute to the structure and virulence of these biofilms in different microorganisms[28] including *M. pneumoniae*.

Another significantly altered pathway is the "Interleukin-12 (IL-12) mediated signalling pathway", found to be over-regulated in *M. pneumoniae* compared with viral pneumonia infection. IL-12 plays an immune regulatory role in innate and adaptive responses during infection, triggering the production of interferon-γ and promoting the differentiation of type 1 T helper cells[29]. Elevated levels of IL-12 have been correlated with the severity of pneumonia in *M. pneumoniae* infections and slower microbiologic and histological recovery in mice, suspected to be related to a boost in the innate immune host response[30].

Remarkably, some of the pathways involving genes included in our signatures, are among or related to the top 20 most significant pathways from the global analysis of *M. pneumoniae* vs. viral pneumonias (Supplementary Fig. 3, Supplementary Data 7). This suggests an important role of these genes in the host-response to *M. pneumoniae* infection. Nucleotide sugar biosynthesis is one of the most significantly up-regulated pathways in the global analysis and, as described above, is closely related to pathogens' metabolism. The endoplasmic reticulum stress or the endoplasmic reticulum unfolded protein response is known to be involved in the innate immune signalling response to different pathogens[31]. Pathogens hijack endoplasmic reticulum functions for their own benefit leading to a stress state in host cells. The unfolded protein response mitigates the endoplasmic reticulum stress, but if it cannot counterbalance this stress, apoptotic and autophagic processes are activated, contributing to the disease pathogenesis. Other pathways related to the regulation of cell cycle, specifically around the mitotic phase, are largely describe to be altered in viral processes[32].

Although our results show evidence of biological processes differentially activated in *M. pneumoniae* pneumonia, the specific

contribution of these pathways to the *M. pneumoniae* pathogenesis needs further exploration.

There are a few limitations in the present study. First, while we analyze the largest cohorts of *M. pneumoniae* infections to date, the sample size remain limited, highlighting the need to explore additional cohorts and sub-phenotypes, such as patients with respiratory symptoms of non-infectious etiology. Second, our signatures were derived from microarray data, leaving room for the discovery of potentially superior signatures once RNAseq data become available. Third, technical validations of these signatures using gold standard methods such as dPCR or qPCR would be valuable, as well as exploring the possibility of translating these signatures into a POC diagnostic tool. Fourth, since gene expression patterns can vary across populations[33], it would be advantageous to examine the performance of the signatures in diverse populations from different geographic regions.

This study emphasizes the urgent need for expanded research on *M. pneumoniae*, providing compelling evidence of this microorganism's unique ability to elicit a distinct gene expression response in infected pneumonia patients. By leveraging this pathogen-specific reaction, we successfully derived flexible gene expression signatures, comprising 3 to 10 transcripts. Validation in an independent cohort of patients provided further evidence indicating their specificity and sensitivity in distinguishing *M. pneumoniae* from other viral, bacterial and undiagnosed pneumonias. This breakthrough enables accurate differentiation between pneumonia caused by *M. pneumoniae* and other pneumonias of different etiologies, providing a framework for appropriate treatment and offering promising avenues for further exploration and potential clinical applications.

Methods

Samples and study design

We selected microarray blood transcriptomic profiles from 151 children hospitalized with pneumonia and 20 healthy controls as reported by Wallihan et al. [17]. The pneumonia cases included infections caused by *M. pneumoniae* ($n = 30$), various viral infections ($n = 77$), pyogenic bacteria ($n = 5$), co-infections not involving *M. pneumoniae* ($n = 10$), co-infections involving *M. pneumoniae* ($n = 11$), and pneumonia infections without detected pathogen ($n = 18$); Fig. 1. The data were downloaded from the Gene Expression Onmibus (GEO) database with accession number GSE103119. Inclusion criteria required evidence of acute infection, signs or symptoms of respiratory illness, and radiologic confirmation of lower respiratory tract disease. For more detailed information on the inclusion and exclusion criteria, see Supplemental Table 1 in Wallihan et al. [17].

Additionally, we have used RNAseq data from blood of pediatric patients with viral and bacterial pneumonia not caused by *M. pneumoniae* recruited, recruited through the European Union Childhood Life-threatening Infectious Diseases Study (EUCLIDS- https://www.diamonds2020.eu/our-research-history/euclids/[34–36] consortium available in GEO and ArrayExpress under accession numbers GSE261482/E-MTAB-11671); Fig. 1. Specifically, the EUCLIDS discovery cohort contains RNAseq data for 39 definitive non-mycoplasma bacterial and 9 definitive viral pneumonias, all of them used and reported by Viz-Lasheras et al.[18](clinical data available in Table 1[18]); see definitions in ref. 14.

For validation, we used new RNAseq data from an additional pediatric cohort comprising blood samples obtained from children infected with *M. pneumoniae* ($n = 9$), samples from viral ($n = 10$) and bacterial ($n = 22$) pneumonias, samples from pneumonias with no causal pathogen detected ($n = 57$) along with healthy controls ($n = 35$) and controls with asthma ($n = 8$); Fig. 1. Clinical features and demographic information of this cohort are provided in Supplementary Data 3. These samples were recruited under the umbrella of *i*) the EUCLIDS Study (EUCLIDS validation), *ii*) the PErsonalised Risk assessment in Febrile illness to Optimise Real-life Management across the European Union (PERFORM - https://www.perform2020.org/) and *iii*) the Diagnosis and Management of Febrile Illness using RNA Personalised Molecular Signature Diagnosis (DIAMONDS - https://www.diamonds2020.eu) consortiums[16,37,38] and are available in ArrayExpress under accession numbers E-MTAB-12793, E-MTAB-14564 and E-MTAB-14588. Pneumonia phenotype in the non-mycoplasma pneumonia cohort (EUCLIDS discovery), and in the validation cohort (EUCLIDS validation-PERFORM-DIAMONDS) was defined as follows: *i*) Inflammation of one or both lungs (lobar, segmental or multilobar collapse/consolidation) on chest X-ray (excluding perihilar consolidation or patchy consolidation); *ii*) Clinical symptoms compatible with acute respiratory infection and radiological evidence of consolidation or pleural effusion: alveolar consolidation (dense or fluffy opacity occupying part or all of a lobe or lung, with or without air-bronchograms) or pleural effusion (fluid in the lateral pleural space, excluding minor or oblique fissure), spatially associated with a pulmonary parenchymal infiltrates (including other infiltrate) or obliterating enough of the hemithorax to obscure opacity. All samples were collected within 24 h of admission. *M. pneumoniae* pneumonia was diagnosed using PCR and/or serology in throat swab samples along with clinical symptoms compatible with atypical pneumonia (Supplementary Data 8).

Inclusion and ethics statement

Harmonized procedures for patient recruitment, classification, clinical data and sample collection, processing and storage were followed across the participating centers. Written, informed consent was obtained from parents or guardians at all sites using locally approved permissions. See study protocol EUCLIDS: see DOI: 10.1016/S2352-4642(18)30113-5; study protocol PERFORM: see DOI: 10.1093/cid/ciad615; study protocol DIAMONDS: see https://doi.org/10.1186/ISRCTN12394803. Each participating country's Ethics Committee (EC) approval is available upon request. The study conformed to the ethical principles for medical research involving human participants outlined in the Declaration of Helsinki. This study was designed and implemented by clinicians and researchers embedded within treating clinical teams and this is reflected in the opportunities to contribute meaningfully to the data analysis, interpretation, and writing. This is reflected in the authorship. Members of the wider EUCLIDS, PERFORM and DIAMONDS Search Study consortia are identified in the Supplementary Information.

RNAseq analysis

Whole blood was collected into PAXgene blood RNA tubes (BD Biosciences; catalog number: 762165) and stored at −80 °C. For the EUCLIDS cohort, total RNA was isolated using PAXgene blood isolation kit according to the manufacturer's instructions (Qiagen; catalog number: 762174). RNA sequencing was conducted on a HiSeq 4000 (Illumina) platform, with library preparation and sequencing of 30 million 75 or 100 bp paired-end reads. The Illumina's TruSeq RNA Sample Preparation Kit was used for library preparation, and ribosomal and globin RNA depletion was performed using the Illumina® Ribo-Zero Gold kit (Illumina; catalog number: MRZG12324). For the PERFORM/DIAMONDS cohort, RNA was isolated using PAXgene blood miRNA isolation kit according to the manufacturer's instructions (Qiagen; catalog number: 763134). An additional DNAse treatment was carried out with the RNA clean & concentrator kit (Zymo Research; catalog number: R1013) prior to sequencing. RNA was quantified using RiboGreen (Invitrogen; catalog number: R11490) on the FLUOstar OPTIMA plate reader (BMG Labtech) and the integrity analyzed on the TapeStation 2200 (Agilent, catalog number: 5067–5576; 5067–5578; 5067–5577). After a normalization step, a strand specific library preparation was completed using NEBNext® Ultra™ II mRNA kit (New England Biolabs; catalog number: E7770) and NEB rRNA/globin depletion probes following manufacturer's recommendations (New

England Biolabs; catalog number: E7750). Individual libraries were normalized using Qubit, pooled together and diluted. The sequencing was performed using a 150 paired-end configuration in a Novaseq6000 platform (Illumina). Quality control of raw data was carried out using *FastQC*[39], alignment and read counting were performed using *STAR*[40], alignment filtering was done with *SAMtools*[41] and read counting was carried out using *FeatureCounts*[42].

Statistical analysis

Microarray data pre-process and normalization was performed using the *illuminaHumanv4.db*[43] and *limma*[44] packages. RNAseq data was processed for batch correction using control samples and *COMBAT-Seq*[45] package. Data was subsequently normalized with *DESeq2*[46] package. Principal component analysis (PCA) was conducted to explore the different groups in the data and check for potential outliers. A differential expression (DE) analysis was carried out using *limma*[44] package and accounting for differences in age, in order to compare the blood transcriptome of pneumonia patients with *M. pneumoniae* infection ($n = 30$) vs. viral pneumonia infections ($n = 77$); step 1 in Fig. 1.

In order to identify subsets of genes that could serve as predictive transcriptomic signatures differentiating *M. pneumoniae* pneumonia from other pneumonias of viral etiology, we randomly split the dataset ($n = 107$[17]); into 1000 independent subsets comprising 70% of the samples (TA datasets) and other 1000 subsets containing the remaining 30% of the samples (TE datasets); step 2 in Fig. 1. A predictive transcriptomic signature was computed using the R package *glmnet*[47] for each of the 1000 TA. To do that, a logistic *LASSO* regression model was fitted with the alpha parameter set to 1 and a 10-fold cross validation (step 3 in Fig. 1); 276 differentially expressed genes (DEGs) were included as input for the logistic regression based on $|\mathrm{Log_2FC}| > 1$, adjusted *P*-value < 0.01, and a $\mathrm{Log_2}$ average expression > 2 (Supplementary Data 1).

The accuracy of the predictive transcriptomic signatures was measured by calculating the area under the receiver operating characteristic curve (AUC) with 95% confidence intervals (CI) using the *pROC* package[48]. The optimal cut-point value (cut-off) that maximize sensitivity and specificity, was calculated using the *OptimalCutPoints* R package[49]. AUC values were computed for the 1000 TA and the corresponding TE datasets (step 4 in Fig. 1), and also for the non-*M. pneumoniae* viral and bacterial pneumonias in the EUCLIDS cohort; step 5 in Fig. 1. Among the 1000 signatures of size *n* transcripts, we selected those *i* signatures of size $n = 3$ to 10 transcripts. Next, on these selected *i* signatures, we computed the score $n\text{-SS}_i = (0.4 \times \mathrm{AUC_{TA}}) + (0.4 \times \mathrm{AUC_{TE}}) - (0.2 \times \mathrm{AUC_{EU}})$ (step 6 in Fig. 1); this scoring method considers the AUC value for both training and test samples, while penalizing performance on non-*M. pneumoniae* pneumonia infections in the EUCLIDS cohort. We selected the best transcriptomic signature of size $n = 3$ to 10 ($n\text{-SS}_{\mathrm{MAX}}$) among the different signatures with the same number of transcripts. The AUC values for the signatures with the $n\text{-SS}_{\mathrm{MAX}}$ values, were then obtained in the 999 complete training resamples (1000 minus the one used to generate the $n\text{-SS}_{\mathrm{MAX}}$); step 7 in Fig. 1. In addition, the robustness of the best signatures to differentiate *M. pneumoniae* pneumonia from all sub-groups included in the original pneumonia dataset[17] was evaluated by calculating the AUC values in the following contrasts: *i) M. pneumoniae* pneumonia vs. viral pneumonia, *ii*) vs. bacterial pneumonia, *iii*) vs. viral pneumonia + non-mycoplasma bacterial pneumonia + non-mycoplasma co-infected pneumonias, *iv)* vs. *M. pneumoniae* pneumonia co-infections, *v*) vs. pneumonia without detected pathogen, and *vi) vs.* healthy controls; step 8 in Fig. 1.

The performance of other signatures available in the literature to differentiate *M. pneumoniae* from viral pneumonias were also investigated. In particular, we tested the 5-transcript signature recently developed to differentiate non-Mycoplasma pneumonia from viral pneumonia[18], and the 2-transcript signature developed to differentiate viral from bacterial infections[14]; step 9 in Fig. 1.

Validation of the best signatures

The performance and accuracy of the predictive transcriptomic signatures were validated using new blood RNAseq data generated from an additional pediatric cohort, comprising new samples from EUCLIDS and PERFORM-DIAMONDS cohorts. Coefficients and intercepts from each *LASSO* model were applied to the new dataset to perform ROC analysis and calculate the AUC, sensitivity, and specificity of the signatures. Six different comparisons were tested using the validation cohort: *i*) *M. pneumoniae* pneumonia vs. viral pneumonia, *ii*) vs. bacterial pneumonia, *iii*) vs. both viral and non-mycoplasma bacterial pneumonias, *iv*) vs. pneumonias without detected pathogen; *v*) vs. controls with asthma, and *vi*) vs. healthy controls, step 10 in Fig. 1.

GSVA pathway analysis

Biological pathways differentially involved in viral and *M. pneumoniae* pneumonia were inferred from gene expression data using the GSVA algorithm included in the Gene Set Variation Analysis (*GSVA*) R package[50]. Gene Ontology (GO) biological pathways gene set collection from the Molecular Signatures Database (MSigDB)[51] was used as reference database. Significantly differentially regulated pathways (DRPs) were determined using the *limma* package[44] with the viral groups as a reference and a threshold of adjusted *P*-value < 0.05.

All graphics were created using R software v.4.3.2 (www-r-project.org).

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

New RNAseq data generated in this study and used for validation are available in ArrayExpress repository (https://www.ebi.ac.uk/biostudies/arrayexpress) under accession number E-MTAB-14588. RNASeq data previously analyzed in other studies are available in ArrayExpress under accession numbers E-MTAB-12793 and E-MTAB-14564 and in Gene Expression Omnibus NCBI (GEO; http://www.ncbi.nlm.nih.gov/geo/) under accession number GSE261482. Details are provided in Supplementary Data 4.

Code availability

Original R code has been deposited at github (https://github.com/Sandraviz/Mycoplasma-pneumoniae_-Transcriptomic-Signatures.git), https://doi.org/10.5281/zenodo.14258748[52].

References

1. Larcher, R., Boudet, A., Roger, C., Villa, F. & Loubet, P. *Mycoplasma pneumoniae* is back! Is it the next pandemic? *Anaesth Crit Care. Pain. Med.* **43**, 101338 (2024).
2. Meyer Sauteur, P. M. et al. *Mycoplasma pneumoniae* detections before and during the COVID-19 pandemic: results of a global survey, 2017 to 2021. *Euro Surveill.* **27**, 2100746 (2022).
3. Meyer Sauteur, P. M. et al. *Mycoplasma pneumoniae*: delayed re-emergence after COVID-19 pandemic restrictions. *Lancet Microbe* **5**, e100–e101 (2024).
4. Urbieta, A. D. et al. *Mycoplasma pneumoniae* at the rise not only in China: rapid increase of Mycoplasma pneumoniae cases also in Spain. *Emerg. Microbes Infect.* **13**, 2332680 (2024).
5. Nordholm, A. C. et al. *Mycoplasma pneumoniae* epidemic in Denmark, October to December, 2023. *Euro Surveill.* **29**, 2300707 (2024).
6. Wreghitt, T. *Mycoplasma pneumoniae*: current outbreak. *Epidemiol. Infect.* **152**, e47 (2024).

7. Bolluyt, D. C. et al. Increased incidence of *Mycoplasma pneumoniae* infections and hospital admissions in the Netherlands, November to December 2023. *Euro Surveill.* **29**, 2300724 (2024).
8. Parums, D. V. Editorial: Outbreaks of post-pandemic childhood pneumonia and the re-emergence of endemic respiratory infections. *Med. Sci. Monit.* **29**, e943312 (2023).
9. Daxboeck, F., Krause, R. & Wenisch, C. Laboratory diagnosis of *Mycoplasma pneumoniae* infection. *Clin. Microbiol. Infect.: Off. Publ. Eur. Soc. Clin Microbiol Infect Dis.* **9**, 263–273 (2003).
10. Landry, M. L. Immunoglobulin M for acute infection: True or false? *Vaccin. Immunol.* **23**, 40–45 (2016).
11. Zhang, L., Zong, Z. Y., Liu, Y. B., Ye, H. & Lv, X. J. PCR versus serology for diagnosing *Mycoplasma pneumoniae* infection: a systematic review & meta-analysis. *Indian J. Med. Res.* **134**, 270–280 (2011).
12. Atkinson, T. P., Balish, M. F. & Waites, K. B. Epidemiology, clinical manifestations, pathogenesis and laboratory detection of *Mycoplasma pneumoniae* infections. *FEMS Microbiol. Rev.* **32**, 956–973 (2008).
13. Rueda, Z. V. et al. Etiology and the challenge of diagnostic testing of community-acquired pneumonia in children and adolescents. *BMC Pediatr.* **22**, 169 (2022).
14. Herberg, J. A. et al. Diagnostic test accuracy of a 2-transcript host RNA signature for discriminating bacterial vs viral infection in febrile children. *JAMA* **316**, 835–845 (2016).
15. Barral-Arca, R., Pardo-Seco, J., Martinón-Torres, F. & Salas, A. A 2-transcript host cell signature distinguishes viral from bacterial diarrhea and it is influenced by the severity of symptoms. *Sci. Rep.* **8**, 8043 (2018).
16. Habgood-Coote, D. et al. Diagnosis of childhood febrile illness using a multi-class blood RNA molecular signature. *Med* **4**, 635–654.e635 (2023).
17. Wallihan, R. G. et al. Molecular distance to health transcriptional score and disease severity in children hospitalized with community-acquired pneumonia. *Front. Cell Infect. Microbiol.* **8**, 382 (2018).
18. Viz-Lasheras, S. et al. A 5-transcript signature for discriminating viral and bacterial etiology in pediatric pneumonia. *iScience* https://doi.org/10.1016/j.isci.2025.111747 (2025).
19. Spuesens, E. B. et al. Carriage of *Mycoplasma pneumoniae* in the upper respiratory tract of symptomatic and asymptomatic children: an observational study. *PLoS Med.* **10**, e1001444 (2013).
20. Wainman, L. M., Sathyanarayana, S. H. & Lefferts, J. A. Applications of digital polymerase chain reaction (dPCR) in molecular and clinical testing. *J. Appl. Lab. Med.* **9**, 124–137 (2024).
21. Daubenspeck, J. M., Jordan, D. S., Simmons, W., Renfrow, M. B. & Dybvig, K. General N-and O-Linked glycosylation of lipoproteins in mycoplasmas and role of exogenous oligosaccharide. *PLoS One* **10**, e0143362 (2015).
22. Klement, M. L., Ojemyr, L., Tagscherer, K. E., Widmalm, G. & Wieslander, A. A processive lipid glycosyltransferase in the small human pathogen *Mycoplasma pneumoniae*: involvement in host immune response. *Mol. Microbiol.* **65**, 1444–1457 (2007).
23. Jordan, D. S., Daubenspeck, J. M., Laube, A. H., Renfrow, M. B. & Dybvig, K. O-linked protein glycosylation in *Mycoplasma*. *Mol. Microbiol.* **90**, 1046–1053 (2013).
24. Ansari, S., Kumar, V., Bhatt, D. N., Irfan, M. & Datta, A. N-Acetylglucosamine sensing and metabolic engineering for attenuating human and plant pathogens. *Bioengineering* **9**, 64 (2022).
25. Simmons, W. L. et al. Type 1 and type 2 strains of *Mycoplasma pneumoniae* form different biofilms. *Microbiol* **159**, 737–747 (2013).
26. Lluch-Senar, M. et al. Comparative "-omics" in *Mycoplasma pneumoniae* Clinical Isolates Reveals Key Virulence Factors. *PLoS One* **10**, e0137354 (2015).
27. Branda, S. S., Vik, S., Friedman, L. & Kolter R. Biofilms: the matrix revisited. *Trends Microbiol.* **13**, 20–26 (2005).
28. Chen, M., Yu, Q. & Sun, H. Novel strategies for the prevention and treatment of biofilm related infections. *Int. J. Mol. Sci.* **14**, 18488–18501 (2013).
29. Brombacher, F., Kastelein, R. A. & Alber, G. Novel IL-12 family members shed light on the orchestration of Th1 responses. *Trends Immunol.* **24**, 207–212 (2003).
30. Salvatore, C. M., Fonseca-Aten, M., Katz-Gaynor, K., Gomez, A. M. & Hardy, R. D. Intranasal interleukin–12 therapy inhibits *Mycoplasma pneumoniae* clearance and sustains airway obstruction in murine pneumonia. *Infect. Immun.* **76**, 732–738 (2008).
31. Choi, J. A. & Song, C. H. Insights into the role of endoplasmic reticulum stress in infectious diseases. *Front. Immunol.* **10**, 3147 (2019).
32. Bagga, S. & Bouchard, M. J. Cell cycle regulation during viral infection. *Methods Mol. Biol.* **1170**, 165–227 (2014).
33. Barral-Arca, R., Pardo-Seco, J., Bello, X., Martinón-Torres, F. & Salas, A. Ancestry patterns inferred from massive RNAseq data. *RNA* **25**, 857–868 (2019).
34. Borghini, L. et al. Identification of regulatory variants associated with genetic susceptibility to meningococcal disease. *Sci. Rep.* **9**, 6966 (2019).
35. Martinon-Torres, F. et al. Life-threatening infections in children in Europe (the EUCLIDS Project): a prospective cohort study. *Lancet Child. Adolesc. Health* **2**, 404–414 (2018).
36. Kumar, V. et al. Variation in CFHR3 determines susceptibility to meningococcal disease by controlling factor H concentrations. *Am. J. Hum. Genet.* **109**, 1680–1691 (2022).
37. Moradi Marjaneh, M. et al. Analysis of blood and nasal epithelial transcriptomes to identify mechanisms associated with control of SARS-CoV-2 viral load in the upper respiratory tract. *J. Infect.* **87**, 538–550 (2023).
38. Jackson, H. et al. Characterisation of the blood RNA host response underpinning severity in COVID-19 patients. *Sci. Rep.* **12**, 12216 (2022).
39. Andrews, S. FastQC: a quality control tool for high throughput sequence data. Available online. *Retrieved May* **17**, 2018 (2010).
40. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
41. Li, H. et al. The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
42. Liao, Y., Smyth, G. K. & Shi, W. *featureCounts*: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2014).
43. Dunning, M., Lynch, A. & Eldridge, M. *illuminaHumanv4*.db: Illumina HumanHT12v4 annotation data (chip illuminaHumanv4). *R package version 1.26.0*. (2015).
44. Ritchie, M. E. et al. *limma* powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47 (2015).
45. Zhang, Y., Parmigiani, G. & Johnson, W. E. ComBat-seq: batch effect adjustment for RNA-seq count data. *NAR. Genom. Bioinform* **2**, lqaa078 (2020).
46. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with. *DESeq2. Genome Biol.* **15**, 550 (2014).
47. Friedman, J., Hastie, T. & Tibshirani, R. Regularization paths for generalized linear models *via* coordinate descent. *J. Stat. Softw.* **33**, 1–22 (2010).
48. Robin, X. et al. *pROC*: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinform* **12**, 77 (2011).
49. López-Ratón, M., Rodríguez-Álvarez, M. X., Cadarso-Suárez, C. & Gude-Sampedro, F. *OptimalCutpoints*: An R package for selecting optimal cutpoints in diagnostic tests. *J. Stat. Softw.* **61**, 1–36 (2014).
50. Hänzelmann, S., Castelo, R. & Guinney, J. GSVA: gene set variation analysis for microarray and RNA-seq data. *BMC Bioinforma.* **14**, 7 (2013).

51. Liberzon, A. et al. The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Syst.* **1**, 417–425 (2015).
52. Viz-Lasheras, S., Gómez-Carballa, A., Bello, X., F., M.-T. & Salas, A. A diagnostic host-specific transcriptome response for *Mycoplasma pneumoniae* pneumonia to guide pediatric patient treatment. https://doi.org/10.5281/zenodo.14258748 (2024).

Acknowledgements

This research project was made possible through the access granted by the Galician Supercomputing Center (CESGA) to its supercomputing infrastructure. The supercomputer FinisTerrae III and its permanent data storage system have been funded by the Spanish Ministry of Science and Innovation, the Galician Government, and the European Regional Development Fund (ERDF). This study also received support by i) ISCIII: TRINEO: PI22/00162; DIAVIR: DTS19/00049; Resvi-Omics: PI19/01039 (to A.S.), ReSVinext: PI16/01569, Enterogen: PI19/01090, OMI-COVI-VAC: PI22/00406 (to F.M.-T.), cofinanciados FEDER, ii) GAIN: IN607B 2020/08 and IN607A 2023/02 (to A.S.), GEN-COVID (IN845D 2020/23 (to F.M.-T.), IIN607A2021/05 (to F.M.-T.) and IN677D 2024/06 (to A.G.-C.); iii) ACIS: BI-BACVIR (PRIS-3, to A.S.), CovidPhy (SA 304 C, to A.S.), PneumoTrack (PRIST-VAL, to A.S.), and Respisal (PRIST-VAL, to F.M.-T.); v) consorcio Centro de Investigación Biomédica en Red de Enfermedades Respiratorias (CB21/06/00103; to A.S. and F.M.-T.) and vi) Spanish Ministry of Science and Innovation (MCIN)/Spanish Research Agency (AEI) (KAWA-TesT: PID2022-142156OB-I00, to A.G-.C.). This study has been funded by Instituto de Salud Carlos III (ISCIII) through the project "CP23/00080" and co-funded by the European Union (A.G-.C.). L.J.S. was supported by the NOMIS foundation. This work was supported by the European Union's Horizon 2020 research and innovation programme under Grant Agreement Nos. 668303 (PERFORM) and 848196 (DIAMONDS). The funders were not involved in the study design, collection, analysis, interpretation of data, the writing of this article or the decision to submit it for publication.

Author contributions

A.S., F.M-.T., A.G-.C. and S.V-.L. conceptualized and designed the study. A.S., F.M-.T., M.L., M.vF., C.F., M.E., W.Z., A.J.P., T.T.K., S.A., M.P., H.A.M. and R.dG. provided key resources. I.R-.C., A.I.D., J.A.H., A.J.C., E.D.C., were responsible for clinical classification of the patients and collection of clinical metadata. M.T., D.Z., C.V., S.C.P., L.J.S., D.S.K., U.V.B. and P.A. contributed to patients recruitment. VJW and NS coordinated sample processing and biobanking. S.V-.L., X.B., A.G-.C., M.K. and D.H-.C. conducted the bioinformatics and computational biology analysis. Computational models and machine learning approaches used to analyze the data were developed by A.S., S.V-.L. and A.G-.C. A.S., S.V-.L. and A.G-.C. wrote the first draft of the manuscript. All authors contributed to discussions regarding the interpretation of results and reviewed and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-025-55932-9.

Correspondence and requests for materials should be addressed to Antonio Salas.

Peer review information *Nature Communications* thanks Charles Langelier and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. A peer review file is available.

Reprints and permissions information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

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/.

© The Author(s) 2025

Sandra Viz-Lasheras[1,2,3,4,26], Alberto Gómez-Carballa[1,2,3,4,26], Xabier Bello[1,2,3,4], Irene Rivero-Calle[3,4,5], Ana Isabel Dacosta[3,4,5], Myrsini Kaforou[6], Dominic Habgood-Coote[6], Aubrey J. Cunnington[6], Marieke Emonts[7,8,9], Jethro A. Herberg[6], Victoria J. Wright[6], Enitan D. Carrol[10,11], Stephane C. Paulus[12], Werner Zenz[13], Daniela S. Kohlfürst[13], Nina Schweintzger[13], Michiel Van der Flier[14,15], Ronald de Groot[15], Luregn J. Schlapbach[16], Philipp Agyeman[17], Andrew J. Pollard[12], Colin Fink[18], Taco T. Kuijpers[19], Suzanne Anderson[20], Ulrich Von Both[21], Marko Pokorn[22], Dace Zavadska[23], María Tsolia[24], Henriëtte A. Moll[25], Clementien Vermont[25], Michael Levin[6], Federico Martinón-Torres[3,4,5], Antonio Salas[1,2,3,4] ✉, On behalf of EUCLIDS, PERFORM, and DIAMONDS consortia

[1]Unidade de Xenética, Instituto de Ciencias Forenses, Facultade de Medicina, Universidade de Santiago de Compostela, 15782, Calle San Francisco sn, Galicia, Spain. [2]Genética de Poblaciones en Biomedicina (GenPoB) Research Group, Instituto de Investigación Sanitaria (IDIS), 15706 Hospital Clínico Universitario de Santiago (SERGAS), Galicia, Spain. [3]Genetics, Vaccines and Infections Research Group (GenViP), Instituto de Investigación Sanitaria de Santiago (IDIS), 15706 Hospital Clínico Universitario de Santiago (SERGAS), Galicia, Spain. [4]Centro de Investigación Biomédica en Red de Enfermedades Respiratorias (CIBER-ES), Madrid, Spain. [5]Translational Pediatrics and Infectious Diseases, Department of Pediatrics, 15706 Hospital Clínico Universitario de Santiago de Compostela, Santiago de Compostela, Galicia, Spain. [6]Department of Infectious Disease, Imperial College London, London W2 1PG, UK. [7]Paediatric Immunology, Infectious Diseases & Allergy, Great North Children's Hospital, Newcastle upon Tyne Hospitals NHS Foundation Trust, Newcastle

upon Tyne NE1 4LP, UK. [8]Translational and Clinical Research Institute, Newcastle University, Newcastle upon Tyne NE2 4HH, UK. [9]NIHR Newcastle Biomedical Research Centre based at Newcastle upon Tyne Hospitals NHS Trust and Newcastle University, Newcastle upon Tyne NE4 5PL, UK. [10]Department of Infectious Diseases, Alder Hey Children's NHS Foundation Trust, Liverpool L12 2AP, UK. [11]Department of Clinical Infection, Microbiology and Immunology, Institute of Infection, Veterinary and Ecological Sciences, University of Liverpool, Liverpool L69 7BE, UK. [12]Department of Paediatrics, University of Oxford and the NIHR Oxford Biomedical Research Centre, Oxford OX3 9DU, UK. [13]Department of General Paediatrics, Medical University of Graz, Graz, Auenbruggerplatz 34/2, 8036 Graz, Austria. [14]Pediatric Infectious Diseases and Immunology, Wilhelmina Children's Hospital, University Medical Center Utrecht, Utrecht 3508 AB, The Netherlands. [15]Pediatric Infectious Diseases and Immunology, Amalia Children's Hospital, and Section Pediatric Infectious Diseases, Laboratory of Medical Immunology, Department of Laboratory Medicine, Radboud Institute for Molecular Life Sciences, Radboud University Medical Center, Nijmegen 6500 HB, The Netherlands. [16]Department of Intensive Care and Neonatology, and Children's Research Center, University Children's Hospital Zürich, University of Zürich, Zürich, Switzerland. [17]Department of Pediatrics, Inselspital, Bern University Hospital, University of Bern, Bern, Switzerland. [18]Micropathology Ltd, University of Warwick, Warwick CV4 7EZ, UK. [19]Division of Pediatric Immunology, Rheumatology and Infectious diseases, Emma Children's Hospital, Amsterdam Univiersyt Medical Center (Amsterdam UMC), Amsterdam 1105 AZ, The Netherlands. [20]Medical Research Council Unit at the London School of Hygene & Tropical Medicine, Banjul, The Gambia. [21]Infectious Diseases, Department of Pediatrics, Dr von Hauner Children's Hospital, University Hospital, LMU Munich, Munich, Germany. [22]Division of Paediatrics, University Medical Centre Ljubljana and Medical Faculty, University of Ljubljana, Ljubljana, Slovenia. [23]Children's Clinical University Hospital, Rīga Stradins University, Rīga, Latvia. [24]Second Department of Paediatrics, National and Kapodistrian University of Athens (NKUA), School of Medicine, Panagiotis & Aglaia, Kyriakou Children's Hospital, Athens, Greece. [25]Department of Paediatric Infectious Diseases and Immunology, Erasmus MC Sophia Children's Hospital, Rotterdam, Netherlands. [26]These authors contributed equally: Sandra Viz-Lasheras, Alberto Gómez-Carballa. ✉e-mail: antonio.salas@usc.es

EUCLIDS, PERFORM, and DIAMONDS consortia

Sandra Viz-Lasheras[1,2,3,4,26], Alberto Gómez-Carballa[1,2,3,4,26], Xabier Bello[1,2,3,4], Irene Rivero-Calle[3,4,5], Ana Isabel Dacosta[3,4,5], Myrsini Kaforou[6], Dominic Habgood-Coote[6], Aubrey J. Cunnington[6], Marieke Emonts[7,8,9], Jethro A. Herberg[6], Victoria J. Wright[6], Enitan D. Carrol[10,11], Stephane C. Paulus[12], Werner Zenz[13], Daniela S. Kohlfürst[13], Nina Schweintzger[13], Michiel Van der Flier[14,15], Ronald de Groot[15], Luregn J. Schlapbach[16], Philipp Agyeman[17], Andrew J. Pollard[12], Colin Fink[18], Taco T. Kuijpers[19], Suzanne Anderson[20], Ulrich Von Both[21], Marko Pokorn[22], Dace Zavadska[23], María Tsolia[24], Henriëtte A. Moll[25], Clementien Vermont[25], Michael Levin[6], Federico Martinón-Torres[3,4,5] & Antonio Salas[1,2,3,4] ✉

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

6. DISCUSSION

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

6. DISCUSSION

Lower respiratory tract infections, and specifically pneumonia, remain a significant global health challenge being a leading cause of death worldwide. Their study requires continued efforts from different disciplines and perspectives that include the investigation of the complex relationship between the causal pathogens and their hosts. Traditional studies have primarily focused on the analysis of clinical symptoms or the detection of specific pathogens, often overlooking the intricate role of the host's biological response and genetic factors in disease progression and susceptibility.

A multi-omic approach applied to the analysis of the molecular factors, specific response and genetic background of the host to the LRTIs can significantly improve the understanding of disease mechanisms, outcomes and allow us to identify host biomarkers and develop new diagnostic tools. This host-centered research has revealed that there are multiple variables influencing immune responses to infections. Among these, risk factors, age, and sex are pivotal features. Understanding these variations is crucial for tailoring treatments and preventive strategies to specific populations, which could significantly improve health outcomes.

Transcriptomic studies and new available technologies have particularly emerged as powerful tools for advancing the understanding of lower tract respiratory diseases. By analyzing gene expression patterns of different demographic groups and different tissue samples, researchers can identify molecular pathways involved in disease progression and uncover potential diagnostic and therapeutic targets in a more specific manner. Besides, other advancements in genetic research have highlighted the role of genetic variants in disease susceptibility and response to infections and treatments. Identifying these variants enables stratification of individuals based on risk, fostering personalized medicine approaches. For instance, genetic predispositions can help predict the likelihood of developing severe forms of respiratory infections, facilitating targeted interventions and more effective management. This genetic dimension complements host-centered studies by providing a foundation for risk-based diagnostics and personalized therapeutic strategies (185).

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Current diagnostic methods for respiratory infections often focus only on clinical presentation or the identification of the specific pathogen, leading to ambiguous or delayed diagnoses using current guidelines and diagnostic techniques. There are some groups of biomarkers and imaging tools which are routinely used to give an indication of the underlying etiology. However, it is often challenging draw accurate conclusions from the integration of such diverse information and sources (189). By changing the focus toward the host's transcriptomic response researchers can develop novel diagnostic tools that speed up the process of diagnosis or prognosis with high specificity and sensitivity instead of relying solely on clinical presentation or pathogen detection. Using ML algorithms researchers can build predictive models and find transcriptomic signatures or biomarkers offering insights into disease state and potential outcomes. This host-centered approach not only enhances diagnostic accuracy but also opens new avenues for therapeutic interventions revolutionizing management of respiratory infections. It facilitates more precise therapies, more prudent use of antibiotics, and, consequently, reduces the incidence of antimicrobial resistance while mitigating the adverse effects of incorrect treatment on patients. This novel approach represents a paradigm shift in diagnosis of lower tract respiratory infectious diseases.

6.1. STUDY OF THE HOST RESPONSE TO COVID-19: IMPLICATIONS FOR DISEASE SEVERITY

COVID-19, caused by the SARS-CoV-2 virus, has rapidly evolved into a global pandemic, revealing unique characteristics that distinguish it from other respiratory infections. The virus spread through the body leading to a spectrum of symptoms, from mild cough to severe pneumonia and acute respiratory distress syndrome. Viral colonization typically starts in a localized tissue environment, with the upper respiratory tract serving as the primary entry point; however, the infection does not always result in a significant systemic impact. A notable feature of COVID-19 is the heterogeneous immune response observed among infected individuals. While some patients exhibit robust innate immune activation, particularly through the IFN response, others experience inadequate immune activation, resulting in viral replication and severe inflammatory responses. This imbalance can lead to significant tissue damage and fatal clinical outcomes. These differences in patient’s immune responses could explain the different disease severities observed between patients and also between males and females. It has been shown that males experience more severe outcomes than females, including higher rates of hospitalization and mortality.

One of the key approaches to understanding the severity of respiratory infections is the study of the host's immune response as both innate and adaptive immune mechanisms are crucial for controlling viral and bacterial replication.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

To explore local and systemic differential host response to SARS-CoV-2 infection, the expression of 594 immune response genes was analyzed in three different tissues (blood, buccal and nasal) from COVID-19 patients affected by different disease severities (mild, moderate and severe) using nCounter technology.

Our analysis pointed to the existence of a set of immune-related genes and pathways that are tissue-specific and correlated to the COVID-19 severity and strongly suggest that disease severity at systemic level might be dependent on local immune response. Key severity signals involved genes related to innate immune response and cytokine signaling. In severe cases, higher expression was observed in blood samples (compared to mild cases), while the opposite trend, with lower expression, was seen in nasal samples. In saliva, hospitalized patients showed higher expression of certain genes related to immune activation. This imbalance between local and systemic immune activation suggests a pathogenic mechanism in which the failure to control the infection in its early stages leads to an uncontrolled systemic inflammatory response.

In the analysis of blood samples from COVID-19 patients most of the DEGs identified were over-expressed compared to controls. Some of these DEGs showed similar expression in other viral and bacterial infections compared to controls, indicating a shared immune response to severe infections. DEGs in severe cases indicate a stronger immune activation compared to mild or moderate cases, with a marked increase in neutrophil-related genes and inflammatory pathways, along with a significant suppression of adaptive immune responses in severe patients. However, the IFN-I pathway, with a main role in immune regulation and critical for early viral defense (190), was strongly activated in mild and moderate cases but suppressed in severe ones. Consistently, in severe cases *vs.* all other categories we observed DEGs related with neutrophil activation (such as *S100A8*/*S100A9* and *CAMP)* that encode proteins usually identified in Neutrophil Extracellular Traps (NETs) and other genes involved in blood coagulation processes that together may contribute to microthrombus formation (191). T-cell receptor signaling through ZAP-70, critical for T-cell proliferation and antiviral response, was suppressed, consistent with severe COVID-19 pathology. Globally, we found that among the main factors contributing to disease progression in severe cases are an impairment of IFN response, suppression of T and B responses (lymphopenia), strong neutrophil-driven immune response with abnormal increase of low density neutrophils (with high capacity to release NETs) and also enrichment in blood coagulation–related genes (85).

In the nasal epithelium analysis, a strong activation of the antiviral immune response was observed in mild cases (and to a lesser extent in moderate cases), driven by a robust IFN-II response, which led to the activation of IFN-induced antiviral genes (*IFITM1*/*IFIT2*), and over-expression of genes encoding for CXCR3 receptor ligands chemokines *CXCL9, CXCL10* and

USC DE COMPOSTELA

CXCL11. These chemokines are highly inducible by IFN gamma, and once joined to the receptor IFN is activated and triggers calcium release to the cytosol, acting as a second messenger to activate downstream signaling transduction within cells. This process triggered downstream cascades linked to the Th1, NK, and cytotoxic T-cell adaptive immune responses, suggesting efficient local control of the infection during the early stages. In contrast, severe cases displayed a more limited local antiviral response and a heightened innate immune response, marked by an overexpression of genes associated with monocyte-macrophage recruitment (*CD163*, *CLEC4E* and *MARCO*) and CC pro-inflammatory chemokines genes (*CCL3*, *CCL4*). These chemokines have been already described as severity-related biomarkers of COVID-19 in airway epithelial cells (192, 193). In addition, pathway analysis revealed disruptions in NAD^+ metabolism in severe cases compared to mild disease, suggesting potential NAD and ATP depletion in host cells. Some intracellular pathogens are known to manipulate host cell NAD^+ levels, compromising several innate immune pathways (194).

Contrary to findings in nasal epithelium, mild patients did not shown signs of immune response against the virus in the buccal mucosa, suggesting that the infection was effectively controlled locally preventing its spread to lower respiratory tract (LRT) and systemic colonization. On the other hand, in severe cases (and to a lesser degree moderate cases) the buccal mucosa showed clear signs of viral activity, cellular arrest, and viral spread to the LRT, with over expression of genes linked to viral dynamics. During the infection, viruses hijack cellular machinery to their own benefit, altering cellular pathways and the levels of several metabolites. For instance, *LGALS3* (galectin-3) was over-expressed in moderate and severe saliva samples, contributing to the pro-inflammatory response, enhancing macrophage survival and recruitment, and potentially promoting severe infection outcomes (195, 196). Moreover, it has been reported as a potential therapeutic target for COVID-19 (197). Other key up-regulated gene in saliva from severe cases was *PLAU* (urokinase-type plasminogen activator), which, along with *CD59* and *CD9*, may facilitate virus-host cell interactions and immune evasion, promoting viral spread. *PLAU* activates plasmin, a protease essential for viral entry by cleaving the SARS-CoV-2 spike protein and enhancing cell infection potential (198). We also observed higher *PLAU* expression in the nasal epithelium of severe cases, indicating a potential tissue crosslink. Tetraspanins like *CD9*, known to mediate viral entry in other coronaviruses, were also implicated, possibly aiding viral dissemination through extracellular vesicles (199). *CD59*, an inhibitor of complement-mediated lysis, may help the virus evade the immune system by integrating into the viral envelope or modulating host expression, as seen in other enveloped viruses (200). Additionally, severe cases showed up-regulation of genes involved in cell junctions, adhesion, and exosome production, processes that could facilitate viral spread. The immune response in the buccal cavity, due to its anatomical position and its connection to the upper and lower respiratory tract, may play a crucial role in viral transmission both between

individuals and within individuals, facilitating viral spread to the LRT and contributing to more severe disease outcomes. All these findings suggest that saliva might not only serve as a reliable indicator of disease severity but could also offer therapeutic opportunities to block early viral spread.

Co-expression modules analysis of blood, nasal epithelium and saliva samples yielded equivalent results to the individual genes approach. Most significantly correlated modules with COVID-19 severity were found un nasal epithelium and saliva. In nasal epithelium samples, two different modules were significantly correlated with mild and severe disease respectively. Hub genes from the mild-related module were mostly implicated in CXCR chemokine receptor binding processes, interferon gamma signaling and adaptative T-cell immune response. By contrast, hub genes within the severe-related module were linked to IL-1 response, CCR chemokine receptor binding, NAD metabolism, and an innate response involving neutrophils and TLR cascades. The co-expression module significantly correlated to the severe phenotype in saliva samples encompasses all up-regulated genes from the differential expression analysis. Notably, many of these genes also emerged as hub genes within the module, indicating their central role in the severe disease signature. These hub genes were associated with processes such as lymphocyte proliferation, activation, apoptosis, cell-cell adhesion, and cell cycle arrest. The most connected gene within this module was *SRC*, a member of the Src family kinases (SFKs), known for their pivotal role in regulating viral life cycles and mediating the interplay between host cell functions and viral demands (201). Inhibitors of SFKs have been proposed as potential treatments for COVID-19 due to their capacity to block viral dissemination (202).

These findings underscore the importance of tissue-specific expression in the early phases of infection, suggesting that a localized immune response in the nasal epithelium plays a crucial role in preventing viral spread to more distant tissues and that buccal cavity might play a key role in SARS-CoV-2 infection and dissemination in patients with worse prognosis. Thus, disease severity at systemic level might be strongly dependent on local immune response. Expanding our knowledge on these localized immune dynamics is crucial for improving COVID-19 management as it can provide insights into disease progression and also help identify potential specific biomarkers for COVID-19 disease severity, guiding clinical decisions thought developing effective diagnostic, prognostic and therapeutic strategies.

There is substantial evidence suggesting that biological differences between males and females contribute to variations in the immune response to respiratory infections. These immunological particularities might be behind the sex biased incidence observed in COVID-19. One factor that was associated with the severity of viral infections is the expression of Toll-like receptor 7 (*TLR7*), whose protein plays a critical role in the detection of viral RNA and the

USC DE COMPOSTELA

subsequent induction of IFN-I production and other proinflammatory cytokines release. This type of interferon is an essential component of the innate immune response, providing an early defense mechanism against pathogens (101). Deficiencies in *TLR7* expression or functionality can lead to diminished IFN-I production, resulting in an impaired immune response and an increased susceptibility to severe manifestations of the disease.

On the other hand, it was described that *TLR7*, which is linked to the X chromosome, as well as other genes related to the immune response, escape from XCI, especially if they are located in specific regions of the chromosome with a high probability to scape (100). This phenomenon can produce an imbalanced dosage of the gene product between females and males. As *TLR7* gene is located in one of those highly probable XCI escape region, it is hypothesized that could have a biallelic dose in females and monoallelic in males. This differences in gene expression of *TLR7* were proposed as a driven factor for the differences in COVID-19 outcomes and disease progression between females and males. The understanding of these sex-specific biological mechanisms can help to decipher key factors of disease severity, guiding management, prognosis and treatment.

Employing a transcriptomic and epigenomic approach, multiple cohorts of COVID-19 patients (including samples from blood, saliva, and nasal tissues), and an accurate and homogeneous classification of patients by severity (ICU non-ICU and WHO score), we found that gene expression and epigenetic profiles of *TLR7* gene differ between sexes and relate to disease outcomes or severity. The analyses provide various evidence supporting the hypothesis that *TLR7* is differentially expressed in males and females with severe COVID-19 and may be involved in the more severe outcomes of man compared to women. The results show that males tend to exhibit lower *TLR7* expression levels compared to females, particularly in severe cases, which aligns with existing epidemiological data indicating a higher incidence and severity of COVID-19 in males. In females, as expected, a higher dose of the product of *TLR7* expression was found, which may be related to the immunological advantage and more effective antiviral response already observed against other viral infections (96, 97). Thus, regardless of the tissue type or severity classification applied, females with severe disease consistently exhibited higher *TLR7* expression compared to males, with blood showing the most pronounced differences in expression levels.

Furthermore, some of the epigenetic modifications (DNA methylation patterns) involving *TLR7* gene that can be also contributing to this sex-related differences were also investigated. Present data suggest that males with severe COVID-19 exhibited alterations in DNA methylation at key sites within the *TLR7* gene, which could contribute to its lower expression. The identification of differentially methylated positions (DMPs) in the *TLR7* gene highlights

USC DE COMPOSTELA

the complexity of gene regulation and its potential implications for immune responses in male patients. The presence of hypermethylated sites in severe male patients compared to mild cases suggests that epigenetic alterations could play a critical role in modulating *TLR7* expression and, consequently, immune responses, providing a deeper insight into the understanding of this biological phenomenon.

However, while *TLR7* expression and methylation patterns provide some insights into the sex bias observed in COVID-19 outcomes, it cannot fully explain the severity discrepancies. Other genetic, hormonal, and environmental factors likely contribute to this phenomenon and should be further investigates.

The value and novelty of this study relies in different aspects that include the employment of a robust multi-omic approach. The integration of transcriptomic and epigenomic data provides a deeper understanding not only the gene expression patterns of *TLR7* but also about how epigenetic modifications, such as DNA methylation, may influence its function in different patient cohorts. Moreover, the study highlights the importance of analyzing gene expression across different tissues (blood, saliva, and nasal samples). Such multi-tissue analyses can provide insights into localized immune responses and help identify biomarkers for disease progression and severity. In addition, the study is approached from a multi-cohort and multi-platform perspective, which offers a more realistic scenario and avoids biased results. Using two different technologies for the analysis and using some overlapping samples provided an opportunity to compare the ability of RNA-seq and nCounter transcriptomic technologies to accurately capture *TLR7* gene expression. The results from both analyses showed a significant correlation, thereby offering technical validation for the expression data. These findings contribute to the growing body of literature advocating for sex-specific considerations in the clinical practice and research of respiratory infections.

6.2. CHALLENGES IN PNEUMONIA DIAGNOSIS, TREATMENT, AND SUSCEPTIBILITY. THE RESPONSE LIES IN THE HOST

Pneumonia is a complex infectious disease influenced by various factors, including the host's immune response during the acute phase of infection, and genetic susceptibility factors that influence an individual's risk of developing pneumonia or experiencing severe outcomes. Identifying host genetic variants linked to immune response, inflammation, and pathogen recognition can uncover at-risk populations and inform preventive strategies such as vaccines or personalized treatments.

USC UNIVERSIDADE

There were previous attempts to identify host mutations associated with immunological processes related to pneumonia but never in such a large cohort. Several mutations associated

with antigen recognition (CD14, toll-like receptors, etc.), proinflammatory and anti-inflammatory responses (including TNF-α, interleukin-6, interleukin-10, etc.), and effector mechanisms (such as angiotensin-converting enzyme, coagulation system genes, etc.) have been proposed to be linked to the host's response to pneumonia (203).

The present study identified novel genetic variants linked to CAP susceptibility. The single allele association test identifies the variant rs1056322 as the most statistically significant SNP under a recessive model, and it is located in the *FAM118A* gene. Additionally, two closely related SNPs, rs6007010 and rs6007594, which are in high linkage disequilibrium, show suggestive statistical significance. To date, there is no solid information about the role of this gen in pneumonia infection or other infectious diseases and this association should be further validated. Interestingly, *FAM118A* has been recently associated to human bronchial epithelial cells infected with respiratory syncytial virus (204).

The most important findings of the present study come from exploring the pathogenicity accumulated in single genes. The most significant genes in SKAT-O collapsing association tests were *COL13A1*, *C20orf96*, *HEXIM1*, *NPAS4*, and *STRC*. Remarkably, none of these genes had been previously associated with pneumonia, highlighting the value of WES in uncovering previously unrecognized genetic factors. The *NPAS4* gene, has been recently found to be significantly downregulated during Zika virus infection and associated with neuronal dysfunction (205). *COL13A1*, that has been previously associated with the respiratory tract in the context of hyperbaric oxygen (206) and *HEXIM1* gene is involved in the regulation of DNA virus-mediated innate immune response (207) and has been also associated with childhood onset asthma (208). However, there is no functional information about the *C20orf96* gene as encode for an unknown protein and its role should be further explored.

The main strength of the study lies in the variety of statistical approaches used to support the findings, as well as reducing cofounding effects and biases such as population stratification or genotyping errors. Besides, using SKAT-O test it was possible to evaluate the combined effect of rare variants within selected genes. which may have a cumulative effect on disease risk that cannot be detected by single-variant tests. While the findings from SKAT-O provided additional insights into the genetic basis of CAP, further replication studies are needed to validate these results in independent cohorts.

In addition, the evaluation of gene expression patterns for the five most significant genes emerging from the collapsing association tests (*COL13A1*, *C20orf96*, *HEXIM1*, *NPAS4*, and *STRC*) and the *FAM18A* detected in the genotyping test, revealed a significant differential expression in pneumonia patients compared to controls. However, these differences were slightly different when evaluating COVID-19 pneumonia and other non-COVID-19 pneumonia

and also when comparing adult and pediatric populations. For instance, *FAM118A* gene showed downregulated expression in both adult and pediatric pneumonia patients compared to controls. *COL13A1* (that has been previously associated with respiratory tract diseases) (206) was downregulated both in non-COVID-19 and in severe COVID-19 pneumonia patients. *HEXIM1* gene, that is a role player in several immunological processes including regulation of DNA virus-mediated innate immune response (207) and it is also associated with childhood onset asthma (208), was found to be overexpressed in COVID-19 pneumonia, more clearly in the most severe cases.

Using gene expression data, predictive performances were also assessed and the results showed predictive power for some combinations of the best identified candidates. Genes that better predict non-COVID-19 pneumonia in adults are *C20orf96* and *NPAS4* with AUC value of 0.64 while in the pediatric cohort are *FAM118* and *STRC*, with AUC value of 0.79 and *COL13A1, FAM118A*, and *HEXIM1* to predict COVID-19 pneumonia. Additionally, genome-wide co-expression correlations shed light into gene-gene interactions at the pathway levels, highlighting an involvement of *FAM118A* and *HEXIM1* genes in methylation regulation processes and a differential implication in pathways related to "DNA geometric changes" and "DNA dependent DNA replication".

Pneumonia is characterized by diverse etiologies, symptomatology, and treatment options, complicating clinical decisions regarding antibiotic prescriptions and hospitalization, especially for vulnerable population such as children. Current molecular methods for bacterial detection in pneumonia patients remain limited, primarily due to the low specificity and sensibility but also difficulties in obtaining samples directly from the infection site. This complexity results in a significant medical cost and unnecessary antibiotic prescriptions being made preventively in cases where bacterial infection cannot be definitively ruled out. This scenario not only raises costs but also contributes to the growing issue of antimicrobial resistance. Study of the host response to pneumonia during the acute phase of infection provides real-time insight into how the immune system responds to the pathogen and can reveal molecular mechanisms driving disease progression, offering targets for diagnosis, prognosis and therapeutic intervention.

Previous studies on transcriptomic biomarkers for pneumonia have also attempted to address pneumonia disease complexity through RNA profiling. For instance, Wallihan et al. (209) generated blood transcriptomic profiles from pediatric CAP cases to assess disease severity, reporting genes linked to inflammation and immune regulation. However, this study was limited by its specificity for cases of *M. pneumoniae* infection, which may not represent the full spectrum of bacterial and viral pathogens in pediatric pneumonia. To date, only one

USC DE COMPOSTELA

study has explored blood transcriptomics in the context of pediatric pneumonia etiology but it focuses on a highly specific population from a malaria-endemic region and result in a transcriptomic signature with a high number of genes, which limits its applicability in clinical practice for diagnostic purposes (210).

In the present study the whole blood transcriptomic patterns of pediatric pneumonia and healthy controls were analyzed, and a pneumonia-specific, minimalistic gene expression signature was developed to differentiate bacterial from viral pneumonia with high accuracy. The findings revealed that pneumonia in children induces extensive modifications in the host's transcriptome compared to healthy controls, impacting the expression of over 5,000 genes in the blood of affected patients. These changes were validated through a different technology and an independent cohort of pediatric pneumonia cases. The top DEGs detected between pneumonia patients and controls in the RNA-seq cohort, involved genes like *OLFM4*, *MMP8*, and *CD177*, which aligned with results in the microarray dataset (independent cohort). These genes encode proteins released from neutrophil granules and have previously been proposed as candidate biomarkers for diagnosing community-acquired pneumonia (CAP) and predicting sepsis. For example, *MMP8* is a collagenase produced by neutrophils that facilitates their migration and has been investigated as a diagnostic biomarker for CAP (211, 212). Its expression has been found to be elevated in both adult and pediatric pneumonia patients, although its limited specificity constrains its use as a standalone diagnostic biomarker. *OLFM4*, a glycoprotein within neutrophil granules, has been associated with severity in infectious diseases and poor outcomes in sepsis (213). Meanwhile, *CD177*, a protein critical for regulating neutrophil migration and activation, has been linked with higher mortality in severe infections such as SARS-CoV-2 (214). In the same line, enrichment analyses highlighted altered pathways within the immune system, particularly those involving the activation and migration of neutrophils, that play a crucial role in immune responses to pulmonary infections. During pneumonia infection, neutrophils migrate to the affected lung tissue, releasing enzymes and cytokines that, while effective at eliminating pathogens, can also contribute to tissue damage and exacerbate inflammation. In addition to the neutrophil-mediated innate response, it was also observed an enrichment of humoral immunity processes and other innate immune responses commonly activated in infections. While these processes assist in pathogen clearance, their excessive activation can worsen clinical outcomes, increasing the risk of severe sepsis and mortality (215). The observed overactivation of immune responses in pediatric pneumonia cases underscores the importance of a precise diagnosis to guide appropriate treatment.

Given the above-mentioned need to enhance diagnostic approaches for pediatric pneumonia, developing new tools to assess infection etiology is essential. This would promote

more responsible antibiotic use and reduce the risks associated with inappropriate treatments. A comparative analysis of viral versus bacterial pneumonia cases revealed 282 differentially expressed genes, most of them showing under-expression in bacterial pneumonia. From these candidate genes, a five-transcript blood signature (*FAM20A*, *BAG3*, *TDRD9*, *MXRA7*, and *KLF14*) was identified demonstrating high accuracy in distinguish between bacterial and viral pneumonia (AUC 0.95, sensitivity 0.8, specificity 1.0 in de discovery cohort). When comparing this five-gene pneumonia-specific signature with a previously published transcriptomic signature for broadly differentiating viral from bacterial infections (124) it showed enhanced specificity and applicability in this respiratory disease context with respect to the existing signatures. Despite both signatures demonstrating substantial diagnostic accuracy, the pneumonia-targeted five-gene approach may offer an extra value, particularly in scenarios where pneumonia is the suspected syndrome. This signature also performed well in the initial-validation cohort, where probable bacterial and probable viral cases were included probable bacterial and probable viral cases, showing robustness across clinical scenarios that mirror real-world challenges. In addition, the proposed transcriptomic signature maintains the high sensitivity and specificity in the final validation analysis using an independent pediatric pneumonia cohort (AUC of 0.92), underscoring its applicability and potential clinical utility.

Among the selected genes of the five-transcript blood signature, some are associated with known antimicrobial activities. For instance, *BAG3* has been recently linked to bacterial proliferation, while *FAM20A* and *TDRD9* have shown increased expression in respiratory distress and sepsis patients (216, 217). These findings suggest that the five-transcript signature can serve as a reliable diagnostic tool, potentially differentiating bacterial from viral pneumonia even before full clinical symptoms emerge. This capability could significantly reduce diagnostic times compared to traditional culture-based methods, potentially shortening hospital stays and avoiding unnecessary tests and antibiotic prescriptions. Finally, the validation of this transcriptomic signature in samples collected across multiple hospitals in different countries reduces the potential for population or regional bias. Although the signature was discovered in a quite small subset of samples, its integration in a machine learning algorithm allows continuous refinement and adaptation as new data emerges from various populations and pathogens.

Co-expression analysis of viral and bacterial pneumonias highlighted clusters of co-expressed genes indicating distinct immunological pathways activated in each type of pneumonia. Specifically, two gene modules were correlated with viral pneumonia, while one was correlated with bacterial pneumonia. The “bacterial” module is primarily involved in innate responses, including cytokine production, acute inflammation, and neutrophil degranulation, along with TLR signaling deficiencies. TLRs play a crucial role in limiting the spread of

USC UNIVERSIDADE

invading bacteria and its activation of TLRs triggers innate immune pathways and upregulates proinflammatory mediators, which recruit neutrophils to the lungs during bacterial infections. A balanced inflammatory response is key to pathogen control; however, prolonged and excessive cytokine release, coupled with heightened neutrophil activity, has been linked to increased risk of severe sepsis and adverse outcomes in CAP cases (215). In contrast, the modules associated with viral pneumonia revealed processes linked to adaptive immunity mediated by T-cells, including T-cell receptor signaling, activation, differentiation, and proliferation. These T-cell responses are essential for combating viral infections, as they facilitate pathogen clearance through cytokine release. CD8+ T-cells, for example, have been shown to be vital in clearing various viral infections, while CD4+ T-cells serve as helper cells, supporting B-cells and CD8+ T-cells and providing independent antiviral functions (218, 219).

In summary, this study has revealed distinct immune mechanisms and gene signatures activated in viral and bacterial pneumonia, representing a substantial advance in the field. The five-transcript signature, in particular, has shown high discriminatory power, offering early diagnostic capability that could easily improve pediatric pneumonia management and treatment decisions. This advancement may reduce hospital stays, alleviate unnecessary antibiotic prescriptions, and improve outcomes for pediatric patients.

During this study, it became evident that existing transcriptomic signatures, which have previously proven effective in distinguishing between viral and bacterial infections (124), including the novel pneumonia signature described above, were not effective in differentiating *M. pneumoniae* pneumonia from viral pneumonia in children. The rise in cases of *M. pneumoniae* since May 2023 has aroused significant concerns in the clinical community. Infections caused by this pathogen necessitate specific treatment approaches, as they do not respond to standard antibiotic therapies typically used for suspected bacterial pneumonia. Importantly, atypical pneumonia, and specifically *M. pneumoniae* pneumonia, triggers an unique and very specific host response, differing from other bacteria or viruses in terms of immune activation pathways and gene expression patterns, limiting the broader application of previous findings in clinical scenarios.

In order to fill this gap in pneumonia diagnosis, the study focused on the development of specific transcriptomic signatures that accurately differentiate with a high sensitivity and specificity *M. pneumoniae* pneumonia against all other pneumonias caused by viral and bacterial pathogens. We found that *M. pneumoniae* pneumonia leads to the alteration of over 3,000 genes in the blood of affected children compared to samples from children with viral pneumonia, enabling the development of eight novel signatures specifically differentiating pneumonia caused by *M. pneumoniae* from viral pneumonia.

USC DE COMPOSTELA

Utilizing an innovative simulation-drive framework that divides the data in 1,000 different possible subsets, we identified the eight distinct transcript combinations ranging from three to ten transcripts, most of which demonstrated an area under the curve value greater than 0.70 in an independent cohort of pneumonia samples. The performance of these signatures increased with the number of predictors. However, while the ten-transcript signature performed best (with an AUC value of 0.95 in the discovery cohort), the three-transcript signature also displayed unexpectedly high accuracy (showing an AUC value of 0.84). Furthermore, all signatures were validated in an independent cohort of pediatric samples from children with other viral and bacterial pneumonia, demonstrating high accuracy (with AUC values from 0.70 to 1.00) and showcasing both reproducibility and generalizability of the model which add significant robustness to the results. Importantly, these signatures maintain their high specificity when distinguishing *M. pneumoniae* infections from other bacterial pneumonias and non-mycoplasma co-infections (showing AUC values that ranged from 0.89 to 0.99 in the discovery cohort and from 0.74 to 0.85 in the validation cohort). Identifying these unique transcriptomic signatures could significantly enhance diagnosis and treatment for affected children, guiding the choice of macrolide antibiotics over beta-lactam antibiotics, that are usually the first choice. Since these signatures demonstrate high sensitivity, they can ensure effective detection of *M. pneumoniae* cases (true positives), even in the first stages of the disease, ultimately improving patient management and outcomes.

Differential pathway analysis further reveals distinct biological processes that are over-activated in *M. pneumoniae* infections compared to viral pneumonias. Many of these altered pathways are linked to nucleotide sugar metabolism, which is vital for pathogen survival and infection due to its role in carbohydrate biosynthesis, protein glycosylation, and glycosylated lipid production. Mycoplasmas, with their limited genome, rely on the host for some essential biomolecules. For example, UDP-N-acetylglucosamine (GlcNAc) is a crucial carbon source for various human pathogens, including *M. pneumoniae,* and its catabolism regulates virulence and pathogenesis. *M. pneumoniae* is known to form biofilms on epithelial surfaces, aiding in immune evasion and antibiotic resistance (220). Biofilms, which are communities of microorganisms within a self-produced matrix, can contribute to persistent infections, with polysaccharides containing GlcNAc playing a significant role in their structure and virulence (221). Another important pathway found to be over-regulated in *M. pneumoniae* infections was the Interleukin-12 (IL-12) signaling pathway, which plays a crucial role in immune regulation during infection, promoting the production of interferon-γ and differentiation of type 1 T helper cells. Elevated IL-12 levels have been associated with increased severity of *M. pneumoniae* pneumonia and slower recovery rates in murine models (222), suggesting a link to increased innate immune responses.

USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Notably, some pathways related to the genes included in the signatures overlap with the most significant pathways from the broader analysis of *M. pneumoniae* versus viral pneumonias, indicating these genes have critical roles in the host's response to *M. pneumoniae* infection. These findings highlight unique biological processes in *M. pneumoniae* pneumonia, but further research is required to clarify the specific contributions of these pathways to the pathogen's virulence.

The selection and utility of these signatures depend on various clinical factors, such as available equipment, budget, sample size, and urgency for results. Few technologies can simultaneously measure such a number of transcripts at a reasonable cost and speed. For instance, RNA sequencing is not the best choice for studying a limited number of target genes due to its time and cost demands. In contrast, NanoString technology offers high precision and convenience for targeting a limited number of genes, though it is often limited to research settings rather than routine medical laboratories. Conventional RT-qPCR, while the gold standard, can be labor-intensive and less cost-effective as the number of target genes increases. Novel PCR techniques, like one-step digital PCR (dPCR), enable the combination of up to five genes with high sensitivity and fewer procedural steps, making them a viable option for investigating biomarkers in some proposed signatures. Besides, future efforts should focus on developing feasible point-of-care devices capable of performing rapid and reliable RNA quantification in devices that can be easily integrated in diverse clinical settings. Thus, research into rapid, patient-friendly, inexpensive, and accessible diagnostic modalities is essential to appropriately characterize and manage LRTIs such as pneumonia, potentially reducing diagnostic turnaround times, guiding timely therapeutic decisions and becoming accessible in a variety of clinical settings and not only in big hospital with specialized equipment. Reducing the number of predictors can increase the likelihood of adapting these signatures for point-of-care (POC) testing, with the three-transcript signature being a promising candidate due to its high accuracy even when compared to signatures with more predictors.

Despite the promise of all the presented signatures (pneumonia five-gene signature and the eight *M. pneumoniae* pneumonia signatures) further validation (in larger and heterogeneous cohorts) is needed to ensure the model can be generalized across epidemiological differences in pathogen prevalence, in areas with high rates of viral co-infections or antimicrobial resistance and in other respiratory non-infectious conditions that may be confusing at the presentation of symptoms.

The multi-omic integration of gene expression, epigenetic, and genetic studies provides a comprehensive understanding of the molecular mechanisms underlying the host response to respiratory infections, such as COVID-19 and CAP. The identification of diagnostic signatures

USC DE COMPOSTELA

specific to *M. pneumoniae* and the differentiation between viral and bacterial pneumonia represents a significant move towards personalized and precision medicine approach, enabling faster and more accurate diagnoses as well as more targeted and effective treatments. This not only reduces treatment delays but also allows for the initiation of therapy prior to obtaining an accurate diagnosis. Furthermore, as the interaction between a pathogen and host can vary widely depending on genetic background, immune responses, and environmental factors, these studies hold particular relevance for improving the management of respiratory diseases.

As transcriptomic technologies and next-generation sequencing continue to advance and decrease in cost, it is anticipated that we will observe a transformation in the diagnosis and treatment of respiratory diseases, leading to a substantial improvement in clinical outcomes for patients worldwide. Although substantial effort is required to address the gap between in vitro findings and their successful translation into clinical practice, significant progress is already being made in the development of technologies that are increasingly viable for point-of-care applications, such as FilmArray (223) , MinIon (224), and other rapid nucleic acid detection methods like loop-mediated isothermal amplification (LAMP) (225, 226). Consequently, the scientific findings presented in this thesis hold considerable potential for clinical application in the near future.

USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

7. CONCLUSIONS

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

7. CONCLUSIONS

The results obtained from the analysis carried out during this doctoral thesis have allowed to reach the following conclusions:

i) Disease severity at systemic level might be strongly dependent on local immune response to COVID-19. Evidence highlights that immune response within the nasal cavities, particularly the balance between antiviral interferon responses and inflammatory pathways, may dictate whether the virus is contained locally or disseminates systemically, thereby influencing disease progression and severity.

ii) The buccal cavity may play a pivotal role in the systemic dissemination of SARS-CoV-2 infection. Transcriptomic analyses reveal that severe COVID-19 cases show active viral replication and disrupted cellular functions in oral tissues. This suggests that the buccal cavity functions not only serves as an entry point for the virus but also contributes significantly to its spread throughtout the body.

iii) Sex-specific differences in *TLR7* expression and methylation patterns appear to influence COVID-19 severity. Males, particularly those with severe disease, consistently exhibit lower *TLR7* expression than females. This reduction, driven by genetic and epigenetic mechanisms linked to the X chromosome, contributes to impaired antiviral responses and a heightened risk of severe outcomes, underscoring the key role of *TLR7* in modulating disease severity.

iv) We have identified genetic susceptibility variants associated with community-acquired pneumonia (CAP). Novel associations have been observed in genes such as *FAM118A*, *NPAS4*, *C20orf96*, *COL13A1*, and *STRC*, providing insights intro genetic predispositions to CAP.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

v) Gene expression analysis reveal significant dysregulation of CAP-associated candidate genes in both non-COVID-19 pneumonia and severe COVID-19 pneumonia. This findings suggests the involvement of these genes in critical host-pathogen interactions and their potential contribution to disease severity, supporting further investigation into their use asl biomarkers.

vi) Specific gene expression responses to viral and bacterial pneumonia have led to the identification of a robust 5-gene transcriptomic signature capable of accurately differentiating between the two conditions. This signature has been successfully validated in an independent cohort, demonstrating its reliability. This approach shows promise for use in point-of-care diagnostic devices, offering a rapid and precise tool to improve the clinical management of pediatric pneumonia and reduce unnecessary antibiotic prescriptions for viral infections.

vii) Viral and bacterial pneumonia activate distinct immunological pathways. Bacterial pneumonia is characterized by innate immune responses, including cytokine production, acute inflammation, neutrophil degranulation, and deficiencies in TLR signaling. In contrast, viral pneumonia predominantly engages adaptive immunity mediated by T-cells, reflecting a fundamentally different host defense strategy.

viii) *Mycoplasma pneumoniae* pneumonia induces highly specific changes in host gene expression, distinguishable through transcriptomic signatures. A combination of 8 transcripts from 17 genes enable accurate differentiation of *M. pneumoniae* pneumonia from other bacterial and viral pneumonias. Given that *M. pneumoniae* is resistant to beta-lactam antibiotics, implementing these signatures in clinical practice could enhance treatment decisions and minimize inappropriate antibiotic use.

ix) *M pneumoniae* uniquely modulates host biological processes. Notably, it over-activates nucleotide sugar metabolism, a pathway critical for pathogen survival and infection, and the interleukin-12 signaling pathway, essential for immune regulation during infection. Genes associated with these pathways highlight their pivotal role in the host's specific response to *M. pneumoniae*.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

.

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

8. REFERENCES

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

8. REFERENCES

1. World Health Organization. Respiratory infections 2024 [Available from: https://platform.who.int/mortality/themes/theme-details/topics/topic-details/MDB/respiratory-infections.
2. Vos T, Lim SS, Abbafati C, Abbas KM, Abbasi M, Abbasifard M, et al. Global burden of 369 diseases and injuries in 204 countries and territories, 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019. The lancet. 2020;396(10258):1204-22.
3. Li Z-J, Zhang H-Y, Ren L-L, Lu Q-B, Ren X, Zhang C-H, et al. Etiological and epidemiological features of acute respiratory infections in China. Nature communications. 2021;12(1):5026.
4. Mandell LA, Wunderink RG, Anzueto A, Bartlett JG, Campbell GD, Dean NC, et al. Infectious Diseases Society of America/American Thoracic Society consensus guidelines on the management of community-acquired pneumonia in adults. Clinical infectious diseases. 2007;44(Supplement_2):S27-S72.
5. Hammond A, Halliday A, Thornton HV, Hay AD. Predisposing factors to acquisition of acute respiratory tract infections in the community: a systematic review and meta-analysis. BMC infectious diseases. 2021;21:1-11.
6. Troeger C, Forouzanfar M, Rao PC, Khalil I, Brown A, Swartz S, et al. Estimates of the global, regional, and national morbidity, mortality, and aetiologies of lower respiratory tract infections in 195 countries: a systematic analysis for the Global Burden of Disease Study 2015. The Lancet Infectious Diseases. 2017;17(11):1133-61.
7. Turner HC, Thwaites GE, Clapham HE. Vaccine-preventable diseases in lower-middle-income countries. The Lancet Infectious Diseases. 2018;18(9):937-9.
8. Bender RG, Sirota SB, Swetschinski LR, Dominguez R-MV, Novotney A, Wool EE, et al. Global, regional, and national incidence and mortality burden of non-COVID-19 lower respiratory infections and aetiologies, 1990–2021: a systematic analysis from the Global Burden of Disease Study 2021. The Lancet Infectious Diseases. 2024.
9. Abbadi A, Gentili S, Tsoumani E, Brandtmüller A, Hendel MK, Salomonsson S, et al. Impact of lower-respiratory tract infections on healthcare utilization and mortality in older adults: a Swedish population-based cohort study. Aging Clinical and Experimental Research. 2024;36(1):146.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

10. Safiri S, Mahmoodpoor A, Kolahi A-A, Nejadghaderi SA, Sullman MJ, Mansournia MA, et al. Global burden of lower respiratory infections during the last three decades. Frontiers in Public Health. 2023;10:1028525.
11. Martinón-Torres F, Carmo M, Platero L, Drago G, López-Belmonte J, Bangert M, et al. Clinical and economic burden of respiratory syncytial virus in Spanish children: the BARI study. BMC Infectious Diseases. 2022;22(1):759.
12. Zhang S, Akmar LZ, Bailey F, Rath BA, Alchikh M, Schweiger B, et al. Cost of respiratory syncytial virus-associated acute lower respiratory infection management in young children at the regional and global level: a systematic review and meta-analysis. The Journal of infectious diseases. 2020;222(Supplement_7):S680-S7.
13. Talic S, Shah S, Wild H, Gasevic D, Maharaj A, Ademi Z, et al. Effectiveness of public health measures in reducing the incidence of covid-19, SARS-CoV-2 transmission, and covid-19 mortality: systematic review and meta-analysis. bmj. 2021;375.
14. Besnier E, Thomson K, Stonkute D, Mohammad T, Akhter N, Todd A, et al. Which public health interventions are effective in reducing morbidity, mortality and health inequalities from infectious diseases amongst children in low-and middle-income countries (LMICs): an umbrella review. PLoS One. 2021;16(6):e0251905.
15. World Health Organization. Infection prevention and control 2024 [Available from: https://www.who.int/health-topics/infection-prevention-and-control#tab=tab_1.
16. Kleynhans J, Tempia S, Shioda K, Von Gottberg A, Weinberger DM, Cohen C. Estimated impact of the pneumococcal conjugate vaccine on pneumonia mortality in South Africa, 1999 through 2016: An ecological modelling study. PLoS Medicine. 2021;18(2):e1003537.
17. Rivero-Calle I, Pardo Seco J, Raguindin P, Alvez F, Gómez-Rial J, Salas A, et al. Routine infant vaccination of pneumococcal conjugate vaccines has decreased pneumonia across all age groups in Northern Spain. Human Vaccines & Immunotherapeutics. 2020;16(6):1446-53.
18. Watt JP, Wolfson LJ, O'Brien KL, Henkle E, Deloria-Knoll M, McCall N, et al. Burden of disease caused by Haemophilus influenzae type b in children younger than 5 years: global estimates. The Lancet. 2009;374(9693):903-11.
19. Gessner BD, Adegbola RA. The impact of vaccines on pneumonia: key lessons from Haemophilus influenzae type b conjugate vaccines. Vaccine. 2008;26:B3-B8.
20. World Health Organization. Global Influenza Programme [Available from: https://www.who.int/teams/global-influenza-programme/vaccines.
21. Abbasi J. Studies suggest COVID-19 vaccine boosters save lives. JAMA. 2022;327(2):115-.
22. World Health Organization. Global COVID-19 Vaccination Strategy in a Changing World: July 2022 update 2022 [Available from: https://www.who.int/publications/m/item/global-covid-19-vaccination-strategy-in-a-changing-world--july-2022-update.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

23. National Institutes of Health. How boosters produce broad protection against COVID-19 2023 [Available from: https://www.nih.gov/news-events/nih-research-matters/varied-boosters-produce-broad-protection-against-covid-19.
24. Moline HL. Early Estimate of Nirsevimab Effectiveness for Prevention of Respiratory Syncytial Virus–Associated Hospitalization Among Infants Entering Their First Respiratory Syncytial Virus Season—New Vaccine Surveillance Network, October 2023–February 2024. MMWR Morbidity and mortality weekly report. 2024;73.
25. Razzaghi H. Maternal Respiratory Syncytial Virus Vaccination and Receipt of Respiratory Syncytial Virus Antibody (Nirsevimab) by Infants Aged 8 Months—United States, April 2024. MMWR Morbidity and Mortality Weekly Report. 2024;73.
26. Surie D, Self WH, Zhu Y, Yuengling KA, Johnson CA, Grijalva CG, et al. RSV vaccine effectiveness against hospitalization among US adults 60 years and older. JAMA. 2024.
27. Ares-Gómez S, Mallah N, Santiago-Pérez M-I, Pardo-Seco J, Pérez-Martínez O, Otero-Barrós M-T, et al. Effectiveness and impact of universal prophylaxis with nirsevimab in infants against hospitalisation for respiratory syncytial virus in Galicia, Spain: initial results of a population-based longitudinal study. The Lancet Infectious Diseases. 2024.
28. Hammitt LL, Dagan R, Yuan Y, Baca Cots M, Bosheva M, Madhi SA, et al. Nirsevimab for prevention of RSV in healthy late-preterm and term infants. New England Journal of Medicine. 2022;386(9):837-46.
29. Terstappen J, Hak SF, Bhan A, Bogaert D, Bont LJ, Buchholz UJ, et al. The respiratory syncytial virus vaccine and monoclonal antibody landscape: the road to global access. The Lancet Infectious Diseases. 2024.
30. Hsiao A, Hansen J, Timbol J, Lewis N, Isturiz R, Alexander-Parrish R, et al. Incidence and estimated vaccine effectiveness against hospitalizations for all-cause pneumonia among older US adults who were vaccinated and not vaccinated with 13-valent pneumococcal conjugate vaccine. JAMA network open. 2022;5(3):e221111-e.
31. Li Z, Lu G, Meng G. Pathogenic fungal infection in the lung. Frontiers in immunology. 2019;10:1524.
32. Graffelman AW, Neven AK, le Cessie S, Kroes AC, Springer MP, van den Broek PJ. Pathogens involved in lower respiratory tract infections in general practice. British Journal of general practice. 2004;54(498):15-9.
33. van Doorn HR, Yu H. Viral respiratory infections. Hunter's tropical medicine and emerging infectious diseases: Elsevier; 2020. p. 284-8.
34. Ieven M, Coenen S, Loens K, Lammens C, Coenjaerts F, Vanderstraeten A, et al. Aetiology of lower respiratory tract infection in adults in primary care: a prospective study in 11 European countries. Clinical Microbiology and Infection. 2018;24(11):1158-63.
35. Mitchell A, Mitchell T. Streptococcus pneumoniae: virulence factors and variation. Clinical Microbiology and Infection. 2010;16(5):411-8.
36. Pickles RJ, DeVincenzo JP. Respiratory syncytial virus (RSV) and its propensity for causing bronchiolitis. The Journal of pathology. 2015;235(2):266-76.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

37. McGonagle D, Sharif K, O'Regan A, Bridgewood C. The role of cytokines including interleukin-6 in COVID-19 induced pneumonia and macrophage activation syndrome-like disease. Autoimmunity reviews. 2020;19(6):102537.
38. Rytter H, Jamet A, Coureuil M, Charbit A, Ramond E. Which current and novel diagnostic avenues for bacterial respiratory diseases? Frontiers in Microbiology. 2020;11:616971.
39. Zheng Y, Qiu X, Wang T, Zhang J. The diagnostic value of metagenomic next–generation sequencing in lower respiratory tract infection. Frontiers in cellular and infection microbiology. 2021;11:694756.
40. Chen X, Ding S, Lei C, Qin J, Guo T, Yang D, et al. Blood and bronchoalveolar lavage fluid metagenomic next-generation sequencing in pneumonia. Canadian Journal of Infectious Diseases and Medical Microbiology. 2020;2020(1):6839103.
41. Zhang D, Yang D, Makam AN. Utility of blood cultures in pneumonia. The American journal of medicine. 2019;132(10):1233-8.
42. Miao Q, Ma Y, Wang Q, Pan J, Zhang Y, Jin W, et al. Microbiological diagnostic performance of metagenomic next-generation sequencing when applied to clinical practice. Clinical Infectious Diseases. 2018;67(suppl_2):S231-S40.
43. Mahony JB. Detection of respiratory viruses by molecular methods. Clinical microbiology reviews. 2008;21(4):716-47.
44. Chang AB, Ooi MH, Perera D, Grimwood K. Improving the diagnosis, management, and outcomes of children with pneumonia: where are the gaps? Frontiers in pediatrics. 2013;1:29.
45. Noviello S, Huang DB. The basics and the advancements in diagnosis of bacterial lower respiratory tract infections. Diagnostics. 2019;9(2):37.
46. Goossens H, Ferech M, Vander Stichele R, Elseviers M. Outpatient antibiotic use in Europe and association with resistance: a cross-national database study. The Lancet. 2005;365(9459):579-87.
47. Trust W. Review on antimicrobial resistance. Tackling drug-resistant infections globally: Final report and recommendations.; 2016.
48. Prina E, Ranzani OT, Torres A. Community-acquired pneumonia. The Lancet. 2015;386(9998):1097-108.
49. Torres A, Catia C, Niederman MS, Rosario M, Chalmers JD, Wunderink RG. Pneumonia (Primer). Nature Reviews: Disease Primers. 2021;7(1).
50. Cillóniz C, Ewig S, Polverino E, Marcos MA, Esquinas C, Gabarrús A, et al. Microbial aetiology of community-acquired pneumonia and its relation to severity. Thorax. 2011;66(4):340-6.
51. Metlay JP, Waterer GW, Long AC, Anzueto A, Brozek J, Crothers K, et al. Diagnosis and treatment of adults with community-acquired pneumonia. An official clinical practice guideline of the American Thoracic Society and Infectious Diseases Society of America. American journal of respiratory and critical care medicine. 2019;200(7):e45-e67.
52. Wootton DG, Aston SJ, Felton T. The challenge of antimicrobial prescribing for hospital-acquired pneumonia. Journal of Hospital Infection. 2020;104(2):198-9.

53. Saraya T. The history of Mycoplasma pneumoniae pneumonia. Frontiers in microbiology. 2016;7:364.
54. Daubenspeck JM, Jordan DS, Simmons W, Renfrow MB, Dybvig K. General N-and O-linked glycosylation of lipoproteins in mycoplasmas and role of exogenous oligosaccharide. PLoS One. 2015;10(11):e0143362.
55. Atkinson TP, Balish MF, Waites KB. Epidemiology, clinical manifestations, pathogenesis and laboratory detection of Mycoplasma pneumoniae infections. FEMS microbiology reviews. 2008;32(6):956-73.
56. Schot MJ, Dekker AR, Giorgi WG, Hopstaken RM, de Wit NJ, Verheij TJ, et al. Diagnostic value of signs, symptoms and diagnostic tests for diagnosing pneumonia in ambulant children in developed countries: a systematic review. NPJ primary care respiratory medicine. 2018;28(1):1-11.
57. Rambaud-Althaus C, Althaus F, Genton B, D'Acremont V. Clinical features for diagnosis of pneumonia in children younger than 5 years: a systematic review and meta-analysis. The Lancet infectious diseases. 2015;15(4):439-50.
58. Scott JAG, Wonodi C, Moïsi JC, Deloria-Knoll M, DeLuca AN, Karron RA, et al. The definition of pneumonia, the assessment of severity, and clinical standardization in the Pneumonia Etiology Research for Child Health study. Clinical infectious diseases. 2012;54(suppl_2):S109-S16.
59. Eslamy HK, Newman B. Pneumonia in normal and immunocompromised children: an overview and update. Radiologic Clinics. 2011;49(5):895-920.
60. Prina E, Ceccato A, Torres A. New aspects in the management of pneumonia. Critical care. 2016;20(1):1-9.
61. Cillóniz C, Torres A, Niederman MS. Management of pneumonia in critically ill patients. bmj. 2021;375.
62. Yun KW. Community-acquired pneumonia in children: updated perspectives on its etiology, diagnosis, and treatment. Clinical and Experimental Pediatrics. 2023;67(2):80.
63. Berti E, Galli L, de Martino M, Chiappini E. International guidelines on tackling community-acquired pneumonia show major discrepancies between developed and developing countries. Acta Pædiatrica. 2013;102:4-16.
64. Izadnegahdar R, Cohen AL, Klugman KP, Qazi SA. Childhood pneumonia in developing countries. The Lancet Respiratory Medicine. 2013;1(7):574-84.
65. Lynch T, Platt R, Gouin S, Larson C, Patenaude Y. Can we predict which children with clinically suspected pneumonia will have the presence of focal infiltrates on chest radiographs? Pediatrics. 2004;113(3):e186-e9.
66. Grief SN, Loza JK. Guidelines for the Evaluation and Treatment of Pneumonia. Primary Care: Clinics in Office Practice. 2018;45(3):485-503.
67. Torres A, Niederman MS, Chastre J, Ewig S, Fernandez-Vandellos P, Hanberger H, et al. Summary of the international clinical guidelines for the management of hospital-acquired and ventilator-acquired pneumonia. ERJ open research. 2018;4(2).
68. Cilloniz C, Liapikou A, Torres A. Advances in molecular diagnostic tests for pneumonia. Current Opinion in Pulmonary Medicine. 2020;26(3):241-8.

69. Christ-Crain M, Opal SM. Clinical review: the role of biomarkers in the diagnosis and management of community-acquired pneumonia. Critical care. 2010;14(1):1-11.
70. Principi N, Esposito S. Biomarkers in pediatric community-acquired pneumonia. International journal of molecular sciences. 2017;18(2):447.
71. Virkki R, Juven T, Rikalainen H, Svedström E, Mertsola J, Ruuskanen O. Differentiation of bacterial and viral pneumonia in children. Thorax. 2002;57(5):438-41.
72. Daxboeck F, Krause R, Wenisch C. Laboratory diagnosis of Mycoplasma pneumoniae infection. Clinical microbiology and infection : the official publication of the European Society of Clinical Microbiology and Infectious Diseases. 2003;9(4):263-73.
73. Landry ML. Immunoglobulin M for Acute Infection: True or False? Vaccine Immunol. 2016;23(7):40-5.
74. Zhang L, Zong ZY, Liu YB, Ye H, Lv XJ. PCR versus serology for diagnosing Mycoplasma pneumoniae infection: a systematic review & meta-analysis. Indian J Med Res. 2011;134(3):270-80.
75. Rueda ZV, Aguilar Y, Maya MA, Lopez L, Restrepo A, Garces C, et al. Etiology and the challenge of diagnostic testing of community-acquired pneumonia in children and adolescents. BMC Pediatr. 2022;22(1):169.
76. Gómez-Carballa A, Bello X, Pardo-Seco J, Martinón-Torres F, Salas A. Mapping genome variation of SARS-CoV-2 worldwide highlights the impact of COVID-19 super-spreaders. Genome Research. 2020;30(10):1434-48.
77. Abbasi J. COVID-19 and mRNA vaccines—first large test for a new approach. Jama. 2020;324(12):1125-7.
78. Wiersinga WJ, Rhodes A, Cheng AC, Peacock SJ, Prescott HC. Pathophysiology, transmission, diagnosis, and treatment of coronavirus disease 2019 (COVID-19): a review. Jama. 2020;324(8):782-93.
79. Jin X, Lian J-S, Hu J-H, Gao J, Zheng L, Zhang Y-M, et al. Epidemiological, clinical and virological characteristics of 74 cases of coronavirus-infected disease 2019 (COVID-19) with gastrointestinal symptoms. Gut. 2020;69(6):1002-9.
80. Huang C, Wang Y, Li X, Ren L, Zhao J, Hu Y, et al. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. The lancet. 2020;395(10223):497-506.
81. Louie JK, Acosta M, Winter K, Jean C, Gavali S, Schechter R, et al. Factors associated with death or hospitalization due to pandemic 2009 influenza A (H1N1) infection in California. Jama. 2009;302(17):1896-902.
82. Wu C, Chen X, Cai Y, Zhou X, Xu S, Huang H, et al. Risk factors associated with acute respiratory distress syndrome and death in patients with coronavirus disease 2019 pneumonia in Wuhan, China. JAMA internal medicine. 2020;180(7):934-43.
83. Scialo F, Daniele A, Amato F, Pastore L, Matera MG, Cazzola M, et al. ACE2: the major cell entry receptor for SARS-CoV-2. Lung. 2020;198:867-77.
84. Hadjadj J, Yatim N, Barnabei L, Corneau A, Boussier J, Smith N, et al. Impaired type I interferon activity and inflammatory responses in severe COVID-19 patients. Science. 2020;369(6504):718-24.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

85. Zheng H-Y, Xu M, Yang C-X, Tian R-R, Zhang M, Li J-J, et al. Longitudinal transcriptome analyses show robust T cell immunity during recovery from COVID-19. Signal transduction and targeted therapy. 2020;5(1):294.
86. Aschenbrenner AC, Mouktaroudi M, Krämer B, Oestreich M, Antonakos N, Nuesch-Germano M, et al. Disease severity-specific neutrophil signatures in blood transcriptomes stratify COVID-19 patients. Genome medicine. 2021;13:1-25.
87. Jackson H, Rivero Calle I, Broderick C, Habgood-Coote D, d'Souza G, Nichols S, et al. Characterisation of the blood RNA host response underpinning severity in COVID-19 patients. Scientific reports. 2022;12(1):12216.
88. Jackson HR, Miglietta L, Habgood-Coote D, D'Souza G, Shah P, Nichols S, et al. Diagnosis of multisystem inflammatory syndrome in children by a whole-blood transcriptional signature. Journal of the Pediatric Infectious Diseases Society. 2023;12(6):322-31.
89. Woodhouse EW, McClain MT, Woods CW. Harnessing the host response for precision infectious disease diagnosis. Clinical Microbiology Reviews. 2024:e00078-24.
90. Eckhardt M, Hultquist JF, Kaake RM, Hüttenhain R, Krogan NJ. A systems approach to infectious disease. Nature Reviews Genetics. 2020;21(6):339-54.
91. Netea MG, Joosten LA, Latz E, Mills KH, Natoli G, Stunnenberg HG, et al. Trained immunity: a program of innate immune memory in health and disease. Science. 2016;352(6284):aaf1098.
92. Parkin J, Cohen B. An overview of the immune system. The Lancet. 2001;357(9270):1777-89.
93. Brinkmann V, Reichard U, Goosmann C, Fauler B, Uhlemann Y, Weiss DS, et al. Neutrophil extracellular traps kill bacteria. science. 2004;303(5663):1532-5.
94. Akashi K, Traver D, Miyamoto T, Weissman IL. A clonogenic common myeloid progenitor that gives rise to all myeloid lineages. Nature. 2000;404(6774):193-7.
95. Iwasaki A, Medzhitov R. Regulation of adaptive immunity by the innate immune system. science. 2010;327(5963):291-5.
96. Kyu HH, Vongpradith A, Sirota SB, Novotney A, Troeger CE, Doxey MC, et al. Age–sex differences in the global burden of lower respiratory infections and risk factors, 1990–2019: results from the Global Burden of Disease Study 2019. The Lancet Infectious Diseases. 2022;22(11):1626-47.
97. Klein SL, Flanagan KL. Sex differences in immune responses. Nature Reviews Immunology. 2016;16(10):626-38.
98. Fink AL, Klein SL. The evolution of greater humoral immunity in females than males: implications for vaccine efficacy. Current opinion in physiology. 2018;6:16-20.
99. Youness A, Miquel C-H, Guéry J-C. Escape from X chromosome inactivation and the female predominance in autoimmune diseases. International journal of molecular sciences. 2021;22(3):1114.
100. Souyris M, Cenac C, Azar P, Daviaud D, Canivet A, Grunenwald S, et al. TLR7 escapes X chromosome inactivation in immune cells. Science immunology. 2018;3(19):eaap8855.
101. Petes C, Odoardi N, Gee K. The toll for trafficking: toll-like receptor 7 delivery to the endosome. Frontiers in immunology. 2017;8:1075.

102. Asano T, Boisson B, Onodi F, Matuozzo D, Moncada-Velez M, Maglorius Renkilaraj MRL, et al. X-linked recessive TLR7 deficiency in~ 1% of men under 60 years old with life-threatening COVID-19. Science immunology. 2021;6(62):eabl4348.
103. Van Der Made CI, Simons A, Schuurs-Hoeijmakers J, van den Heuvel G, Mantere T, Kersten S, et al. Presence of genetic variants among young men with severe COVID-19. Jama. 2020;324(7):663-73.
104. Butler-Laporte G, Povysil G, Kosmicki JA, Cirulli ET, Drivas T, Furini S, et al. Exome-wide association study to identify rare variants influencing COVID-19 outcomes: Results from the Host Genetics Initiative. PLoS genetics. 2022;18(11):e1010367.
105. Lyon MF. Sex chromatin and gene action in the mammalian X-chromosome. American journal of human genetics. 1962;14(2):135.
106. Spiering AE, de Vries TJ. Why females do better: the X chromosomal TLR7 gene-dose effect in COVID-19. Frontiers in Immunology. 2021;12:756262.
107. Tukiainen T, Villani A-C, Yen A, Rivas MA, Marshall JL, Satija R, et al. Landscape of X chromosome inactivation across human tissues. Nature. 2017;550(7675):244-8.
108. Scully EP, Haverfield J, Ursin RL, Tannenbaum C, Klein SL. Considering how biological sex impacts immune responses and COVID-19 outcomes. Nature Reviews Immunology. 2020;20(7):442-7.
109. Zhang Q, Bastard P, Cobat A, Casanova J-L. Human genetic and immunological determinants of critical COVID-19 pneumonia. Nature. 2022;603(7902):587-98.
110. Takahashi T, Ellingson MK, Wong P, Israelow B, Lucas C, Klein J, et al. Sex differences in immune responses that underlie COVID-19 disease outcomes. Nature. 2020;588(7837):315-20.
111. Karczewski KJ, Snyder MP. Integrative omics for health and disease. Nature Reviews Genetics. 2018;19(5):299-310.
112. Tounta V, Liu Y, Cheyne A, Larrouy-Maumus G. Metabolomics in infectious diseases and drug discovery. Molecular omics. 2021;17(3):376-93.
113. Lowe R, Shirley N, Bleackley M, Dolan S, Shafee T. Transcriptomics technologies. PLoS computational biology. 2017;13(5):e1005457.
114. Westermann AJ, Gorski SA, Vogel J. Dual RNA-seq of pathogen and host. Nature Reviews Microbiology. 2012;10(9):618-30.
115. Rajagopala SV, Bakhoum NG, Pakala SB, Shilts MH, Rosas-Salazar C, Mai A, et al. Metatranscriptomics to characterize respiratory virome, microbiome, and host response directly from clinical samples. Cell reports methods. 2021;1(6).
116. Gliddon HD, Herberg JA, Levin M, Kaforou M. Genome-wide host RNA signatures of infectious diseases: discovery and clinical translation. Immunology. 2018;153(2):171-8.
117. Warsinske H, Vashisht R, Khatri P. Host-response-based gene signatures for tuberculosis diagnosis: A systematic comparison of 16 signatures. PLoS medicine. 2019;16(4):e1002786.
118. Anderson ST, Kaforou M, Brent AJ, Wright VJ, Banwell CM, Chagaluka G, et al. Diagnosis of childhood tuberculosis and host RNA expression in Africa. New England Journal of Medicine. 2014;370(18):1712-23.

119. Kaforou M, Wright VJ, Oni T, French N, Anderson ST, Bangani N, et al. Detection of tuberculosis in HIV-infected and-uninfected African adults using whole blood RNA expression signatures: a case-control study. PLoS medicine. 2013;10(10):e1001538.
120. Herberg JA, Kaforou M, Gormley S, Sumner ER, Patel S, Jones KD, et al. Transcriptomic profiling in childhood H1N1/09 influenza reveals reduced expression of protein synthesis genes. The Journal of infectious diseases. 2013;208(10):1664-8.
121. Barral-Arca R, Gómez-Carballa A, Cebey-López M, Bello X, Martinón-Torres F, Salas A. A meta-analysis of multiple whole blood gene expression data unveils a diagnostic host-response transcript signature for respiratory syncytial virus. International journal of molecular sciences. 2020;21(5):1831.
122. Gómez-Carballa A, Barral-Arca R, Cebey-López M, Currás-Tuala MJ, Pischedda S, Gómez-Rial J, et al. Host transcriptomic response following administration of rotavirus vaccine in infants' mimics wild type infection. Frontiers in immunology. 2021;11:580219.
123. Lee HJ, Georgiadou A, Walther M, Nwakanma D, Stewart LB, Levin M, et al. Integrated pathogen load and dual transcriptome analysis of systemic host-pathogen interactions in severe malaria. Science translational medicine. 2018;10(447):eaar3619.
124. Herberg JA, Kaforou M, Wright VJ, Shailes H, Eleftherohorinou H, Hoggart CJ, et al. Diagnostic test accuracy of a 2-transcript host RNA signature for discriminating bacterial vs viral infection in febrile children. Jama. 2016;316(8):835-45.
125. Channon-Wells S, Habgood-Coote D, Vito O, Galassini R, Wright VJ, Brent AJ, et al. Integration and validation of host transcript signatures, including a novel 3-transcript tuberculosis signature, to enable one-step multiclass diagnosis of childhood febrile disease. Journal of Translational Medicine. 2024;22(1):802.
126. Habgood-Coote D, Wilson C, Shimizu C, Barendregt AM, Philipsen R, Galassini R, et al. Diagnosis of childhood febrile illness using a multi-class blood RNA molecular signature. Med. 2023;4(9):635-54. e5.
127. Gómez-Carballa A, Barral-Arca R, Cebey-López M, Bello X, Pardo-Seco J, Martinón-Torres F, et al. Identification of a minimal 3-transcript signature to differentiate viral from bacterial infection from best genome-wide host RNA biomarkers: a multi-cohort analysis. International Journal of Molecular Sciences. 2021;22(6):3148.
128. Gómez-Carballa A, Cebey-López M, Pardo-Seco J, Barral-Arca R, Rivero-Calle I, Pischedda S, et al. A qPCR expression assay of IFI44L gene differentiates viral from bacterial infections in febrile children. Scientific reports. 2019;9(1):1-12.
129. Chaussabel D, Pascual V, Banchereau J. Assessing the human immune system through blood transcriptomics. BMC biology. 2010;8:1-14.
130. He S, Wang L-H, Liu Y, Li Y-Q, Chen H-T, Xu J-H, et al. Single-cell transcriptome profiling of an adult human cell atlas of 15 major organs. Genome biology. 2020;21:1-34.
131. Li C-X, Zhang L, Yan Y-R, Ding Y-J, Lin Y-N, Zhou J-P, et al. A narrative review of exploring potential salivary biomarkers in respiratory diseases: Still on its way. Journal of Thoracic Disease. 2021;13(7):4541.
132. Landry ML, Foxman EF. Antiviral response in the nasopharynx identifies patients with respiratory virus infection. The Journal of infectious diseases. 2018;217(6):897-905.

133. Choi YH, Kim JK. Dissecting cellular heterogeneity using single-cell RNA sequencing. Molecules and cells. 2019;42(3):189-99.
134. van IJzendoorn DG, Szuhai K, Briaire-de Bruijn IH, Kostine M, Kuijjer ML, Bovée JV. Machine learning analysis of gene expression data reveals novel diagnostic and prognostic biomarkers and identifies therapeutic targets for soft tissue sarcomas. PLoS computational biology. 2019;15(2):e1006826.
135. Huang Z, Johnson TS, Han Z, Helm B, Cao S, Zhang C, et al. Deep learning-based cancer survival prognosis from RNA-seq data: approaches and evaluations. BMC medical genomics. 2020;13:1-12.
136. Jiang Y, Zhang Z, Wang W, Huang W, Chen C, Xi S, et al. Biology-guided deep learning predicts prognosis and cancer immunotherapy response. Nature Communications. 2023;14(1):5135.
137. Ryan FJ, Hope CM, Masavuli MG, Lynn MA, Mekonnen ZA, Yeow AEL, et al. Long-term perturbation of the peripheral immune system months after SARS-CoV-2 infection. BMC medicine. 2022;20:1-23.
138. Barturen G, Carnero-Montoro E, Martínez-Bueno M, Rojo-Rello S, Sobrino B, Porras-Perales Ó, et al. Whole blood DNA methylation analysis reveals respiratory environmental traits involved in COVID-19 severity following SARS-CoV-2 infection. Nature Communications. 2022;13(1):4597.
139. Molina J, González-Gamarra A, Ginel L, Peláez ME, Juez JL, Artuñedo A, et al. Cappric study—characterization of community-acquired pneumonia in spanish adults managed in primary care settings. Microorganisms. 2021;9(3):508.
140. Bumgarner R. Overview of DNA microarrays: types, applications, and their future. Current protocols in molecular biology. 2013;101(1):22.1. 1-.1. 11.
141. Dunning M, Lynch A, Eldridge M. illuminaHumanv4. db: Illumina HumanHT12v4 annotation data (chip illuminaHumanv4). R package version. 2015;1(0):10.18129.
142. Dunning M, Eldridge M. illuminaHumanv3. db: Illumina HumanHT12v3 annotation data (chip illuminaHumanv3). R package version 1.26. 0. 2015.
143. Du P, Kibbe WA, Lin SM. lumi: a pipeline for processing Illumina microarray. Bioinformatics. 2008;24(13):1547-8.
144. Carvalho BS, Irizarry RA. A framework for oligonucleotide microarray preprocessing. Bioinformatics. 2010;26(19):2363-7.
145. Sweeney TE, Wong HR, Khatri P. Robust classification of bacterial and viral infections via integrated host gene expression diagnostics. Science translational medicine. 2016;8(346):346ra91-ra91.
146. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic acids research. 2015;43(7):e47-e.
147. Illumina. Introduction to SBS technology 2024 [Available from: https://www.illumina.com/science/technology/next-generation-sequencing/sequencing-technology.html.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

148. Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, Cervera A, McPherson A, et al. A survey of best practices for RNA-seq data analysis. Genome biology. 2016;17:1-19.
149. Risso D, Schwartz K, Sherlock G, Dudoit S. GC-content normalization for RNA-Seq data. BMC bioinformatics. 2011;12:1-17.
150. Andrews S. FastQC: a quality control tool for high throughput sequence data. Available online. Retrieved May. 2010;17:2018.
151. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29(1):15-21.
152. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence alignment/map format and SAMtools. bioinformatics. 2009;25(16):2078-9.
153. Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics. 2014;30(7):923-30.
154. Risso D, Ngai J, Speed TP, Dudoit S. Normalization of RNA-seq data using factor analysis of control genes or samples. Nature biotechnology. 2014;32(9):896-902.
155. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome biology. 2014;15(12):1-21.
156. Vandesompele J, De Preter K, Pattyn F, Poppe B, Van Roy N, De Paepe A, et al. Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes. Genome biology. 2002;3:1-12.
157. Zhong S. ctrlGene: assess the stability of candidate housekeeping genes. R package version 10 1, edn2019.
158. Bhattacharya A, Hamilton AM, Furberg H, Pietzak E, Purdue MP, Troester MA, et al. An approach for normalization and quality control for NanoString RNA expression data. Briefings in bioinformatics. 2021;22(3):bbaa163.
159. Newman AM, Steen CB, Liu CL, Gentles AJ, Chaudhuri AA, Scherer F, et al. Determining cell type abundance and expression from bulk tissues with digital cytometry. Nature biotechnology. 2019;37(7):773-82.
160. Newman AM, Liu CL, Green MR, Gentles AJ, Feng W, Xu Y, et al. Robust enumeration of cell subsets from tissue expression profiles. Nature methods. 2015;12(5):453-7.
161. Nwanganga F, Chapple M. Practical machine learning in R: John Wiley & Sons; 2020.
162. Gu Z, Eils R, Schlesner M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. Bioinformatics. 2016;32(18):2847-9.
163. Kassambara A. Machine learning essentials: Practical guide in R: Sthda; 2018.
164. Lazar C, Taminau J, Meganck S, Steenhoff D, Coletta A, Molter C, et al. A survey on filter techniques for feature selection in gene expression microarray analysis. IEEE/ACM transactions on computational biology and bioinformatics. 2012;9(4):1106-19.
165. Kaur A, Guleria K, Trivedi NK, editors. Feature selection in machine learning: Methods and comparison. 2021 International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE); 2021: IEEE.
166. Emmert-Streib F, Dehmer M. High-dimensional LASSO-based computational regression models: regularization, shrinkage, and selection. Machine Learning and Knowledge Extraction. 2019;1(1):359-83.

167. Wong T-T, Yeh P-Y. Reliable accuracy estimates from k-fold cross validation. IEEE Transactions on Knowledge and Data Engineering. 2019;32(8):1586-94.
168. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. Journal of statistical software. 2010;33(1):1.
169. Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez J-C, et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC bioinformatics. 2011;12(1):1-8.
170. Sing T, Sander O, Beerenwinkel N, Lengauer T. ROCR: visualizing classifier performance in R. Bioinformatics. 2005;21(20):3940-1.
171. López-Ratón M, Rodríguez-Álvarez MX, Cadarso-Suárez C, Gude-Sampedro F. OptimalCutpoints: an R package for selecting optimal cutpoints in diagnostic tests. Journal of statistical software. 2014;61:1-36.
172. Blighe K, Rana S, Lewis M. EnhancedVolcano: Publication-ready volcano plots with enhanced colouring and labeling. R package version. 2019;1(0):10.18129.
173. Wickham H. ggplot2. Wiley interdisciplinary reviews: computational statistics. 2011;3(2):180-5.
174. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC bioinformatics. 2008;9(1):1-13.
175. Miller HE, Bishop AJ. Correlation AnalyzeR: functional predictions from gene co-expression correlations. BMC bioinformatics. 2021;22:1-19.
176. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. Nature genetics. 2000;25(1):25-9.
177. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic acids research. 2000;28(1):27-30.
178. Fabregat A, Jupe S, Matthews L, Sidiropoulos K, Gillespie M, Garapati P, et al. The reactome pathway knowledgebase. Nucleic acids research. 2018;46(D1):D649-D55.
179. Chicco D, Agapito G. Nine quick tips for pathway enrichment analysis. PLoS computational biology. 2022;18(8):e1010348.
180. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences. 2005;102(43):15545-50.
181. Hänzelmann S, Castelo R, Guinney J. GSVA: gene set variation analysis for microarray and RNA-seq data. BMC bioinformatics. 2013;14:1-15.
182. Yu G, Wang L-G, Han Y, He Q-Y. clusterProfiler: an R package for comparing biological themes among gene clusters. Omics: a journal of integrative biology. 2012;16(5):284-7.
183. Egger G, Liang G, Aparicio A, Jones PA. Epigenetics in human disease and prospects for epigenetic therapy. Nature. 2004;429(6990):457-63.
184. Aryee MJ, Jaffe AE, Corrada-Bravo H, Ladd-Acosta C, Feinberg AP, Hansen KD, et al. Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. Bioinformatics. 2014;30(10):1363-9.

185. MacArthur D, Manolio T, Dimmock D, Rehm H, Shendure J, Abecasis G, et al. Guidelines for investigating causality of sequence variants in human disease. Nature. 2014;508(7497):469-76.
186. Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. Genome research. 2009;19(9):1655-64.
187. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience. 2015;4(1):s13742-015-0047-8.
188. Lee S, Miropolsky L, Wu M. SKAT: SNP-set (Sequence) kernel association test. R package version. 2013;82.
189. Craig JC, Williams GJ, Jones M, Codarini M, Macaskill P, Hayen A, et al. The accuracy of clinical symptoms and signs for the diagnosis of serious bacterial infection in young febrile children: prospective cohort study of 15 781 febrile illnesses. Bmj. 2010;340.
190. Crouse J, Kalinke U, Oxenius A. Regulation of antiviral T cell responses by type I interferons. Nature Reviews Immunology. 2015;15(4):231-42.
191. Leppkes M, Knopf J, Naschberger E, Lindemann A, Singh J, Herrmann I, et al. Vascular occlusion by neutrophil extracellular traps in COVID-19. EBioMedicine. 2020;58.
192. Chua RL, Lukassen S, Trump S, Hennig BP, Wendisch D, Pott F, et al. COVID-19 severity correlates with airway epithelium–immune cell interactions identified by single-cell analysis. Nature biotechnology. 2020;38(8):970-9.
193. Xu Z-S, Shu T, Kang L, Wu D, Zhou X, Liao B-W, et al. Temporal profiling of plasma cytokines, chemokines and growth factors from mild, severe and fatal COVID-19 patients. Signal transduction and targeted therapy. 2020;5(1):100.
194. Mesquita I, Varela P, Belinha A, Gaifem J, Laforge M, Vergnes B, et al. Exploring NAD+ metabolism in host–pathogen interactions. Cellular and molecular life sciences. 2016;73:1225-36.
195. Wang H, Song X, Huang Q, Xu T, Yun D, Wang Y, et al. LGALS3 promotes treatment resistance in glioblastoma and is associated with tumor risk and prognosis. Cancer Epidemiology, Biomarkers & Prevention. 2019;28(4):760-9.
196. Liu F-T, Hsu DK. The role of galectin-3 in promotion of the inflammatory response. Drug news & perspectives. 2007;20(7):455-60.
197. Garcia-Revilla J, Deierborg T, Venero JL, Boza-Serrano A. Hyperinflammation and fibrosis in severe COVID-19 patients: galectin-3, a target molecule to consider. Frontiers in Immunology. 2020;11:2069.
198. Kam Y-W, Okumura Y, Kido H, Ng LF, Bruzzone R, Altmeyer R. Cleavage of the SARS coronavirus spike glycoprotein by airway proteases enhances virus entry into human bronchial epithelial cells in vitro. PloS one. 2009;4(11):e7870.
199. Hassanpour M, Rezaie J, Nouri M, Panahi Y. The role of extracellular vesicles in COVID-19 virus infection. Infection, Genetics and Evolution. 2020;85:104422.
200. Maloney BE, Perera KD, Saunders DR, Shadipeni N, Fleming SD. Interactions of viruses and the humoral innate immune response. Clinical immunology. 2020;212:108351.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

201. Pagano MA, Tibaldi E, Palù G, Brunati AM. Viral proteins and Src family kinases: Mechanisms of pathogenicity from a “liaison dangereuse”. World journal of virology. 2013;2(2):71.
202. Pillaiyar T, Laufer S. Kinases as potential therapeutic targets for anti-coronaviral therapy. Journal of Medicinal Chemistry. 2021;65(2):955-82.
203. Waterer GW, Wunderink RG. Genetic susceptibility to pneumonia. Clinics in chest medicine. 2005;26(1):29-38.
204. Gardinassi LG. A Cross-Study Biomarker Signature of Human Bronchial Epithelial Cells Infected with Respiratory Syncytial Virus. Advances in virology. 2016;2016(1):3605302.
205. Alpuche-Lazcano SP, Saliba J, Costa VV, Campolina-Silva GH, Marim FM, Ribeiro LS, et al. Profound downregulation of neural transcription factor Npas4 and Nr4a family in fetal mice neurons infected with Zika virus. PLoS Neglected Tropical Diseases. 2021;15(5):e0009425.
206. Yuan Y, Zhou Y, Li Y, Hill C, Ewing RM, Jones MG, et al. Deconvolution of RNA-seq analysis of hyperbaric oxygen-treated mice lungs reveals mesenchymal cell subtype changes. International journal of molecular sciences. 2020;21(4):1371.
207. Morchikh M, Cribier A, Raffel R, Amraoui S, Cau J, Severac D, et al. HEXIM1 and NEAT1 long non-coding RNA form a multi-subunit complex that regulates DNA-mediated innate immune response. Molecular cell. 2017;67(3):387-99. e5.
208. Zhu Z, Guo Y, Shi H, Liu C-L, Panganiban RA, Chung W, et al. Shared genetic and experimental links between obesity-related traits and asthma subtypes in UK Biobank. Journal of Allergy and Clinical Immunology. 2020;145(2):537-49.
209. Wallihan RG, Suárez NM, Cohen DM, Marcon M, Moore-Clingenpeel M, Mejias A, et al. Molecular distance to health transcriptional score and disease severity in children hospitalized with community-acquired pneumonia. Frontiers in cellular and infection microbiology. 2018;8:382.
210. Silterra J, Gillette MA, Lanaspa M, Pellé KG, Valim C, Ahmad R, et al. Transcriptional categorization of the etiology of pneumonia syndrome in pediatric patients in malaria-endemic areas. The Journal of infectious diseases. 2017;215(2):312-20.
211. Fang X, Duan S-F, Hu Z-Y, Wang J-J, Qiu L, Wang F, et al. Inhibition of Matrix Metalloproteinase-8 Protects Against Sepsis Serum Mediated Leukocyte Adhesion. Frontiers in Medicine. 2022;9.
212. Lauhio A, Hästbacka J, Pettilä V, Tervahartiala T, Karlsson S, Varpula T, et al. Serum MMP-8,-9 and TIMP-1 in sepsis: high serum levels of MMP-8 and TIMP-1 are associated with fatal outcome in a multicentre, prospective cohort study. Hypothetical impact of tetracyclines. Pharmacological research. 2011;64(6):590-4.
213. Liu W, Rodgers GP, editors. Olfactomedin 4 is a biomarker for the severity of infectious diseases. Open Forum Infectious Diseases; 2022: Oxford University Press US.
214. Levy Y, Wiedemann A, Hejblum BP, Durand M, Lefebvre C, Surenaud M, et al. CD177, a specific marker of neutrophil activation, is associated with coronavirus disease 2019 severity and death. Iscience. 2021;24(7).

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

215. Kellum JA, Kong L, Fink MP, Weissfeld LA, Yealy DM, Pinsky MR, et al. Understanding the inflammatory cytokine response in pneumonia and sepsis: results of the Genetic and Inflammatory Markers of Sepsis (GenIMS) Study. Archives of internal medicine. 2007;167(15):1655-63.
216. Juss JK, House D, Amour A, Begg M, Herre J, Storisteanu DM, et al. Acute respiratory distress syndrome neutrophils have a distinct phenotype and are resistant to phosphoinositide 3-kinase inhibition. American journal of respiratory and critical care medicine. 2016;194(8):961-73.
217. Davenport EE, Burnham KL, Radhakrishnan J, Humburg P, Hutton P, Mills TC, et al. Genomic landscape of the individual host response and outcomes in sepsis: a prospective cohort study. The Lancet Respiratory Medicine. 2016;4(4):259-71.
218. Kervevan J, Chakrabarti LA. Role of CD4+ T cells in the control of viral infections: Recent advances and open questions. International Journal of Molecular Sciences. 2021;22(2):523.
219. Swain SL, McKinstry KK, Strutt TM. Expanding roles for CD4+ T cells in immunity to viruses. Nature Reviews Immunology. 2012;12(2):136-48.
220. Ansari S, Kumar V, Bhatt DN, Irfan M, Datta A. N-acetylglucosamine sensing and metabolic engineering for attenuating human and plant pathogens. Bioengineering. 2022;9(2):64.
221. Simmons WL, Daubenspeck JM, Osborne JD, Balish MF, Waites KB, Dybvig K. Type 1 and type 2 strains of Mycoplasma pneumoniae form different biofilms. Microbiology. 2013;159(Pt_4):737-47.
222. Salvatore C, Fonseca-Aten M, Katz-Gaynor K, Gomez A, Hardy R. Intranasal interleukin-12 therapy inhibits Mycoplasma pneumoniae clearance and sustains airway obstruction in murine pneumonia. Infection and immunity. 2008;76(2):732-8.
223. Poritz MA, Blaschke AJ, Byington CL, Allen L, Nilsson K, Jones DE, et al. FilmArray, an automated nested multiplex PCR system for multi-pathogen detection: development and application to respiratory tract infection. PloS one. 2011;6(10):e26047.
224. Jain M, Olsen HE, Paten B, Akeson M. The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. Genome biology. 2016;17:1-11.
225. Parida M, Sannarangaiah S, Dash PK, Rao P, Morita K. Loop mediated isothermal amplification (LAMP): a new generation of innovative gene amplification technique; perspectives in clinical diagnosis of infectious diseases. Reviews in medical virology. 2008;18(6):407-21.
226. Garg N, Ahmad FJ, Kar S. Recent advances in loop-mediated isothermal amplification (LAMP) for rapid and efficient detection of pathogens. Current Research in Microbial Sciences. 2022;3:100120.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

ANNEX I: PUBLISHED OR ACCEPTED WORKS

USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

ANNEX I: Published or Accepted Works

ARTICLE 1: A multi-tissue study of immune gene expression profiling highlights the key role of the nasal epithelium in COVID-19 severity

DOI: 10.1016/j.envres.2022.112890

EDITORIAL PERMISSION TO REPRODUCE THE CONTENT: This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/bync-nd/4.0/). All articles published gold open access will be immediately and permanently free for everyone to read and download, copy and distribute.

ARTICLE 2: Sex-biased expression of the *TLR7* gene in severe covid-patients: insights from transcriptomics and epigenomics

DOI: 10.1016/j.envres.2022.114288

EDITORIAL PERMISSION TO REPRODUCE THE CONTENT: This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/bync-nd/4.0/). All articles published gold open access will be immediately and permanently free for everyone to read and download, copy and distribute.

ARTICLE 3: Whole exome sequencing identifies new susceptibility candidates underlying community-acquired pneumonia

DOI: doi.org/10.1016/j.gendis.2023.101170

EDITORIAL PERMISSION TO REPRODUCE THE CONTENT: For open access publishing, this journal uses an exclusive licensing agreement. Authors will retain copyright alongside scholarly usage rights, and KeAi Communications Co., Ltd. will be granted exclusive publishing and distribution rights.

ARTICLE 4: A 5-transcript signature for discriminating viral and bacterial etiology in pediatric pneumonia

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

PATENT APPLICATTION: There is a European Patent application derived from this work under application number No. 24 382 240.0.

EDITORIAL PERMISSION TO REPRODUCE THE CONTENT: CC BY-NC-ND 4.0 license allows users to distribute and copy the article; and include in a collective work (such as an anthology). These uses are permitted only for non-commercial purposes, and provided the user gives appropriate credit to the author(s) (with a link to the formal publication through the relevant DOI); provides a link to the license; and does not alter or modify the article.

ARTICLE 5: A diagnostic host-specific transcriptome response for *Mycoplasma pneumoniae* pneumonia to guide pediatric patient treatment

PATENT APPLICATTION: There is a European Patent application derived from this work under application number No. 24 382 675.7.

EDITORIAL PERMISSION TO REPRODUCE THE CONTENT: Authors have the right to reuse their article's Version of Record, in whole or in part, in their own thesis. Additionally, they may reproduce and make available their thesis, including Springer Nature content, as required by their awarding academic institution. Authors must properly cite the published article in their thesis according to current citation standards.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

ANNEX II: ETICAL APPROVAL

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

ANNEX II: Ethical Approval

XUNTA DE GALICIA
CONSELLERÍA DE SANIDADE

SERVIZO GALEGO de SAÚDE | Xerencia do Servizo Galego de Saúde

galicia

DICTAMEN DEL Comité de ética de la investigación con medicamentos de Galicia (CEIm-G)

Paula M. López Vázquez, Secretaria del Comité de ética de la investigación con medicamentos de Galicia (CEIm-G)

CERTIFICA:

Que este Comité evaluó en su reunión del día 24/04/2020:

Título: Diagnosis and Management of Febrile Illness using RNA Personalised Molecular Signature Diagnosis
Promotor: *Imperial College London*
Tipo de estudio: Otros
Version: version 1.0, 09 de marzo 2020 y Hoja de información para Padres/ Tutor legal Versión Versión 2.0. 22 de abril de 2020 y Hoja e información al menor maduro Versión 2.0. 22 de abril de 2020 y hoja de información para el/ la participante Versión 2.0. 22 de abril de 2020
Código del Promotor: DIAMONDS
Código de Registro: 2020/219

Y, tomando en consideración las siguientes cuestiones:

- La pertinencia del estudio, teniendo en cuenta el conocimiento disponible, así como los requisitos legales aplicables.
- La idoneidad del protocolo en relación con los objetivos del estudio, justificación de los riesgos y molestias previsibles para el sujeto, así como los beneficios esperados.
- Los principios éticos de la Declaración de Helsinki vigente.
- Los Procedimientos Normalizados de Trabajo del CEIm-G de Galicia

Emite un **INFORME FAVORABLE** para la realización del estudio por el/la investigador/a del centro:

Centros	Investigadores Principales
C.H. Universitario de Santiago	Federico Martinón Torres

Documento asinado dixitalmente por:
Paula López Vázquez (26/04/2020 17:45)
https://sede.xunta.gal/cve?idcve=5A0C-C4C4-B0AH-0F4O-TTJ5-Z158-7915-9592-12

Edificio Administrativo San Lázaro
15703 SANTIAGO DE COMPOSTELA
Teléfono: 881546425
ceic@sergas.es

XUNTA DE GALICIA
CONSELLERÍA DE SANIDADE

SERVIZO GALEGO de SAÚDE | Xerencia do Servizo Galego de Saúde

galicia

Y HACE CONSTAR QUE:

1. El comité cumple los requisitos legales vigentes aplicables a los Comités de ética de investigación.
2. El comité tanto en su composición como en sus PNTs cumple las Normas de Buena Práctica Clínica (CPMP/ICH/135/95).
3. La composición actual del comité es:

Susana María Romero Yuste (Presidenta). Médico Especialista en Reumatología.
Irene Zarra Ferro. (Vicepresidenta). Farmacéutica de Atención Especializada.
Paula Mª López Vázquez, (Secretaria). Médico Especialista en Farmacología Clínica.
Belen Pérez Marcos (Vicesecretaria). Farmacéutica de Atención Especializada
Rosendo Bugarín González. Médico Especialista en Medicina Familiar y Comunitaria.
Juan Casariego Rosón. Médico Especialista en Cardiología.
Nuria Carballeda Feijóo. Miembro lego. Representante de los intereses de los pacientes.
Juana Mª Cruz del Río. Trabajadora Social.
Rafael Álvaro Millán Calenti. Licenciado en Derecho. Asesor jurídico.
Juan Fernando Cueva Bañuelos. Médico Especialista en Oncología Médica.
José Álvaro Fernández Rial. Médico Especialista en Medicina Interna.
José Luis Fernández Trisac. Médico Especialista en Pediatría.
Mª José Ferreira Díaz. Diplomada Universitaria de Enfermería.
Agustín Pía Morandeira. Farmacéutico de Atención Primaria
Jorge Prado Casal. Licenciado en CC. Físicas. Experto en Protección de Datos.
Carmen Rodríguez-Tenreiro Sánchez. Licenciada en Farmacia.
Diego Santos García. Médico Especialista en Neurología.
Juan Vázquez Lago. Médico Especialista en Medicina Preventiva y Salud Pública.
Mª Asunción Verdejo González. Médico Especialista en Farmacología Clínica..

En Santiago de Compostela,

Documento asinado dixitalmente por:
Paula López Vázquez (26/04/2020 17:45)
https://sede.xunta.gal/cve?dcve=SAOC-G4G4-BQAH-OF4Q-TTJ5-Z158-7915-9592-12

Xacobeo 2021

Edificio Administrativo San Lázaro
15703 SANTIAGO DE COMPOSTELA
Teléfono: 881546425
ceic@sergas.es

SERVIZO GALEGO DE SAÚDE

DICTAMEN DEL Comité de ética de la investigación con medicamentos de Galicia (CEIm-G)

2020/219

Laura Romero Sánchez, Secretaria del Comité de ética de la investigación con medicamentos de Galicia (CEIm-G)

CERTIFICA:

Que este Comité evaluó en su reunión del día 31/07/24 la modificación del estudio:

Título: Diagnosis and Management of Febrile Illness using RNA Personalised Molecular Signature Diagnosis
Versión modificación: Protocolo v2.0 de 13 de mayo de 2024.
Tipo de estudio: Outros
Promotor: Imperial College London
Código del Promotor: DIAMONDS
Código de Registro: 2020/219

Y que este Comité acepta de conformidad con sus procedimientos normalizados de trabajo y tomando en cuenta los requisitos éticos, metodológicos y legales exigibles, que dicha enmienda sea incorporada al estudio de investigación.

Documento asinado dixitalmente por:
Laura Romero Sanchez (02/08/2024 09:03)
https://sede.xunta.gal/cve?idcve=SAOC-Q4G4-BOAH-OF4O-TGRE-2RWX-W1DH-9172-2582-2048-81

CEIm de Galicia
XERENCIA DO SERVIZO GALEGO DE SAÚDE
Complexo Administrativo de San Lázaro
15781 Santiago de Compostela
T. 881 546425
ceic@sergas.gal
https://acis.sergas.es/cartafol/Redes-de-Comites-de-Etica-da-Investigacion

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

SERVIZO GALEGO DE SAÚDE

Y HACE CONSTAR QUE:

1 El comité cumple los requisitos legales vigentes aplicables a los Comités de ética de investigación.
2 El comité tanto en su composición como en sus PNTs cumple las Normas de Buena Práctica Clínica (CPMP/ICH/135/95).
3 La composición actual del comité es:

Susana María Romero Yuste (Presidenta). Médico Especialista en Reumatología.
Diego Santos García (Vicepresidente). Médico Especialista en Neurología.
Laura Romero Sánchez (Secretaria). Médico Especialista en Alergología.
Martina Lema Oreiro (Vicesecretaria). Farmacéutica de Hospital.
Rosendo Bugarín González. Médico Especialista en Medicina Familiar y Comunitaria.
Nuria Carballeda Feijóo. Miembro lego. Representante de los intereses de los pacientes.
Juana Mª Cruz del Río. Trabajadora Social.
Rafael Álvaro Millán Calenti. Asesor jurídico.
José Álvaro Fernández Rial. Médico Especialista en Medicina Interna.
José Luis Fernández Trisac. Médico Especialista en Pediatría.
Mª José Ferreira Díaz. Diplomada Universitaria de Enfermería.
Agustín Pía Morandeira. Farmacéutico de Atención Primaria
Jorge Prado Casal. Licenciado en CC. Físicas. Experto en Protección de Datos.
Carmen Rodríguez-Tenreiro Sánchez. Licenciada en Farmacia.
Rafael Carlos Vidal Pérez. Médico Especialista en Cardiología.
Mª Asunción Verdejo González. Médico Especialista en Farmacología Clínica.
Paula Mª López Vázquez. Médico Especialista en Farmacología Clínica.

Firmado digitalmente en Santiago de Compostela

Documento asinado dixitalmente por:
Laura Romero Sanchez (02/08/2024 09:03)
https://sede.xunta.gal/cve?idcve=SAOC-Q4G4-BOAH-OF4O-TGRE-2RWX-W1DH-9172-2582-2048-81

CEIm de Galicia
XERENCIA DO SERVIZO GALEGO DE SAÚDE
Complexo Administrativo de San Lázaro
15781 Santiago de Compostela
T. 881 546425
ceic@sergas.gal
https://acis.sergas.es/cartafol/Redes-de-Comites-de-Etica-da-Investigacion





Lower respiratory tract infections represent a significant global health challenge, causing millions of deaths annually, particularly in children under five. This thesis adopts host-focused approach to unravel the complexities of respiratory infections, with a particular emphasis on COVID-19 and pneumonia. By integrating a multi-omic strategy (including genomics, transcriptomics, and epigenomics) with machine learning techniques, the research identifies specific immune response patterns, genetic and epigenetic factors, and novel transcriptomic biomarkers for the diagnosis and prognosis of respiratory infections. These insights deepen our understanding of host-pathogen interactions and have the potential to transform the management, diagnosis and treatment of respiratory infections.