Methodological aspects of precision medicine with application to primary Sjögren's disease

Cheïma Boudjeniba

MATHÉMATIQUES ET INFORMATIQUE
Sciences
Université Paris Cité

UNIVERSITE PARIS CITE

École doctorale de Sciences Mathématiques de Paris Centre (ED386)

Laboratoire Mathématiques Appliquées à Paris 5 (MAP5)

Methodological aspects of precision medicine with application to primary Sjögren's disease

Par **Cheïma BOUDJENIBA**

Thèse de doctorat de **Mathématiques Appliquées**

Dirigée par **Etienne BIRMELE**
Présentée et soutenue publiquement le 07/12/2023

Devant un jury composé de :

Antoine CHAMBAZ, Prof. HDR	Université de Paris Cité	Président
Michael BARNES	Queen Mary Université de Londres	Rapporteur
Damien CHAUSSABEL, PhD	Jackson Laboratory	Rapporteur
Chiara BALDINI, MD-PhD	Université de Pise	Examinatrice
Pilar BRITO-ZERON, MD-PhD	Hopital Sanitas CIMA	Examinatrice
Etienne BIRMELE, Prof. HDR	Université de Strasbourg	Directeur
Etienne BECHT, PhD	Laboratoires SERVIER	Co-encadrant
Benno SCHWIKOWSKI, PhD	Institut Pasteur	Co-encadrant

Université Paris Cité
Laboratoire MAP5-Mathématiques Appliquées à Paris 5
Campus Saints-Pères
45 Rue des Saints-Pères,
75006 Paris
https://www.map5.mi.parisdescartes.fr/

Institut Pasteur
Computational Biology Department
Computational Systems Biomedicine Laboratory
28 Rue du Dr Roux,
75015 Paris
https://research.pasteur.fr/en/team/csb/

Les laboratoires Servier
Institut de Recherche et Développement Paris-Saclay
Translational Medicine - Quantitative Pharmacology
Rue Francis Perrin,
91190 Gif-sur-Yvette
https://www.servier.com/

Acknowledgments

I would like to express my sincere gratitude to all those who have contributed to the completion of this thesis.

First and foremost, I naturally wish to express my profound gratitude to all my supervisors. Your unwavering support, invaluable scientific insights, and supervision have been instrumental to my growth.
Mickaël GUEDJ, your belief in and proposal of this subject instilled in me the confidence I lacked. Our scientific exploration was immensely enjoyable, and you opened the door to a captivating realm of research.
To **Benno SCHWIKOWSKI**, your propensity for challenging the status quo has pushed my boundaries further.
Etienne BIRMELE, your exceptional academic guidance and availability ensured we adhered to deadlines and maintained rigorous statistical standards, even when delving far from your scientific comfort zone.
I extend my heartfelt gratitude to **Etienne BECHT** for being an extraordinary supervisor. Your commitment to the project, despite joining midway, was remarkable. The values I've imbibed from you extend beyond scientific excellence; you've imparted lessons in patience when confronting setbacks and unwavering support through my highs and lows. I earnestly hope for an opportunity to collaborate with you again in this field. Your brilliance is truly inspiring.

To Professor **Antoine CHAMBAZ**, President of the Jury. I extend my gratitude to you for agreeing to assess this work and preside over the jury.

To Professors **Michael BARNES** and **Damien CHAUSSABEL,**, Reviewers. I express my thanks to you for the insightful feedbacks and constructive criticism during the thesis review process. Your expertise added depth and rigor to my research, and I am grateful for the time and effort they dedicated to evaluating my work.

My thanks also go to all the colleagues who have helped me :
Laurence LAIGLE, thank you for guiding me throughout my thesis. By bringing your vast scientific knowledge, you managed to give meaning to my results.
My dear colleagues at Servier, **Nanna BARNKOB**, Antoine HAMON,

Florent LEFORT, Jack SWINDLE, Bastien CHASSAGNOL, Perrine SORET, Céline LEFEBVRE, Jeremy MANRY, Yufei LUO, Sahar ELOUEJ, Antoine BICHAT, Sophie COURTADE-GAIANI, Fabien MELCHIORE, Christelle RODRIGUES, François RIGLET. Many thanks to immunology team, Phillipe MOINGEON, Emiko DESVAUX, Sandra HUBERT and Audrey AUSSY, I greatly appreciate your expertise in immunology; your ability to share your passion has made collaborating with you an absolute delight.

To my colleagues and friends at the Pasteur Institute, **Mara, Diana, Federico** and **Océane**, you have been my sunshine during this thesis. We have shared so much, your presence has brightened my journey.

Je souhaite enfin exprimer ma reconnaissance envers tous les membres de ma famille proche. Maman, Papa, Amina, Amine, Abdelkader, Vanessa, Djaber, Chloé, Hamza, Rachida et sa famille, ainsi que mes tantes et oncles, je vous adresse mes sincères remerciements pour votre soutien inébranlable et votre présence tout au long de ces années. Merci de m'avoir accordé votre confiance et de m'avoir constamment encouragé sur le chemin qui mène à ce jour.

J'adresse également mes remerciements pour tous les moments légers et joyeux à mes petits neveux et nièces, Nevine, Kenzy, Eline, Aïden, Arïa et Paul. Je leur souhaite tout le succès et l'épanouissement dans tous vos projets futures.

Summary

Title: Advancing Precision Medicine in Primary Sjögren's Syndrome: Integrative Approaches and Gene Modules

In the dynamic field of medical research, the last two decades have seen significant strides in molecular biology, driven by technologies like next-generation sequencing. This shift has steered medicine away from a one-size-fits-all model toward precision medicine, acknowledging the unique nature of each individual. The interplay of genes and the environment shapes health, symptoms, and treatment responses, necessitating personalized therapies. Notably, molecular subtypes in cancer, tied to specific treatments, exemplify this progress, reinforced by FDA-approved prognostic signatures.

This doctoral project focuses on Primary Sjögren's Syndrome (pSS), an autoimmune disorder affecting moisture-producing glands. The IMI2 NECESSITY project, a collaborative effort between academia and industry, aims to deepen the understanding of pSS and identify new clinical and molecular markers for potential trials.

Utilizing diverse omics and clinical data, the project navigates the complexity of pSS, seeking consensus in patient stratification. Employing systems immunology, it decodes disrupted molecular networks, unveiling hidden drivers of immunity. Figure 1 outlines the constructed pipeline.

To interpret the transcriptome in pSS, we jointly analyzed four independent datasets profiling blood samples from 265 pSS patients. Gene clustering across datasets showcased reproducible gene modules, representing primary biological features in the blood transcriptomic profile of pSS patients.

Biological significance of consensus modules (CM) was enhanced by interpreting them using public pathway and blood cell transcriptome databases. Correlations with cellular frequencies validated these interpretations, allowing identification of gene modules indicative of rare cell populations and non-immune cell type signatures.

The identified gene modules provide a foundation for translational research in pSS, offering potential biomarkers from the blood transcriptomic profile. Complementing recent studies on disease stratification, our approach highlights functional and cellular composition differences between patient subgroups.

The research also aims to identify treatment response factors through historical clinical trial data within the IMI2 NECESSITY consortium, contributing to future trial sensitivity. Direct collaboration underscores the in-

terdisciplinary and collaborative nature of the project.

This doctoral work not only addresses pSS complexity but aligns with the broader mission of advancing precision medicine. As interdisciplinary efforts converge, innovative therapeutic strategies in autoimmune disorders become conceivable. The future of pSS research looks promising, with opportunities for exploration using advanced tools like single RNA sequencing and spatial transcriptomics. Concurrently, the development of computational tools tailored for RNA sequencing data analysis addresses inherent challenges.

Keywords: Precision Medicine, Sjögren's Syndrome, Gene Modules.

Résumé

Titre : Avancées de la médecine de précision dans le syndrome de Sjögren primaire : Approches intégratives et Modules de gènes

Dans le domaine dynamique de la recherche médicale, les deux dernières décennies ont vu des avancées significatives en biologie moléculaire, impulsées par des technologies telles que le séquençage de nouvelle génération. Ce changement a orienté la médecine loin d'un modèle unique pour tous, vers la médecine de précision, reconnaissant la nature unique de chaque individu. L'interaction entre les gènes et l'environnement façonne la santé, les symptômes et les réponses au traitement, nécessitant des thérapies personnalisées. Notamment, les sous-types moléculaires dans le cancer, liés à des traitements spécifiques, illustrent cette progression, renforcée par des signatures pronostiques approuvées par la FDA.

Ce projet de doctorat se concentre sur le syndrome de Sjögren primaire (pSS), un trouble auto-immun affectant les glandes productrices d'humidité. Le projet IMI2 NECESSITY, une collaboration entre le milieu académique et l'industrie, vise à approfondir la compréhension du pSS et à identifier de nouveaux marqueurs cliniques et moléculaires pour des essais potentiels.

En utilisant des données omiques et cliniques diverses, le projet navigue dans la complexité du pSS en recherchant un consensus dans la stratification des patients. En employant l'immunologie des systèmes, il décrypte les réseaux moléculaires perturbés, révélant les moteurs cachés de l'immunité. La Figure 1 décrit le pipeline construit.

Pour interpréter le transcriptome dans le pSS, nous avons analysé conjointement quatre ensembles de données indépendants profilant des échantillons sanguins de 265 patients atteints de pSS. Le regroupement génique entre les ensembles de données a mis en évidence la reproductibilité des modules géniques, représentant les principales caractéristiques biologiques dans le profil transcriptomique sanguin des patients atteints de pSS.

La signification biologique des modules consensuels (MC) a été renforcée en les interprétant à l'aide de bases de données publiques de voies et de transcriptomes de cellules sanguines. Les corrélations avec les fréquences cellulaires ont validé ces interprétations, permettant l'identification de modules géniques indicatifs de populations cellulaires rares et de signatures de types cellulaires non immunitaires.

Les modules géniques identifiés servent de base à la recherche translationnelle dans le pSS, offrant des biomarqueurs potentiels issus du profil transcrip-

tomique sanguin. En complément des récentes études sur la stratification de la maladie, notre approche met en lumière les différences fonctionnelles et cellulaires entre les sous-groupes de patients.

La recherche vise également à identifier les facteurs de réponse au traitement à travers les données historiques d'essais cliniques au sein du consortium IMI2 NECESSITY, contribuant à la sensibilité des futurs essais. La collaboration directe souligne la nature interdisciplinaire et collaborative du projet.

Ce travail doctoral aborde non seulement la complexité du pSS, mais s'inscrit dans la mission plus large de faire progresser la médecine de précision. À mesure que les efforts interdisciplinaires convergent, des stratégies thérapeutiques innovantes dans les troubles auto-immuns deviennent envisageables. L'avenir de la recherche sur le pSS semble prometteur, avec des opportunités d'exploration à l'aide d'outils avancés tels que le séquençage unique de l'ARN et la transcriptomique spatiale. Parallèlement, le développement d'outils informatiques adaptés à l'analyse des données de séquençage de l'ARN répond aux défis inhérents.

Mots-clés : Médecine de précision, Syndrome de Sjögren, Modules de gènes.

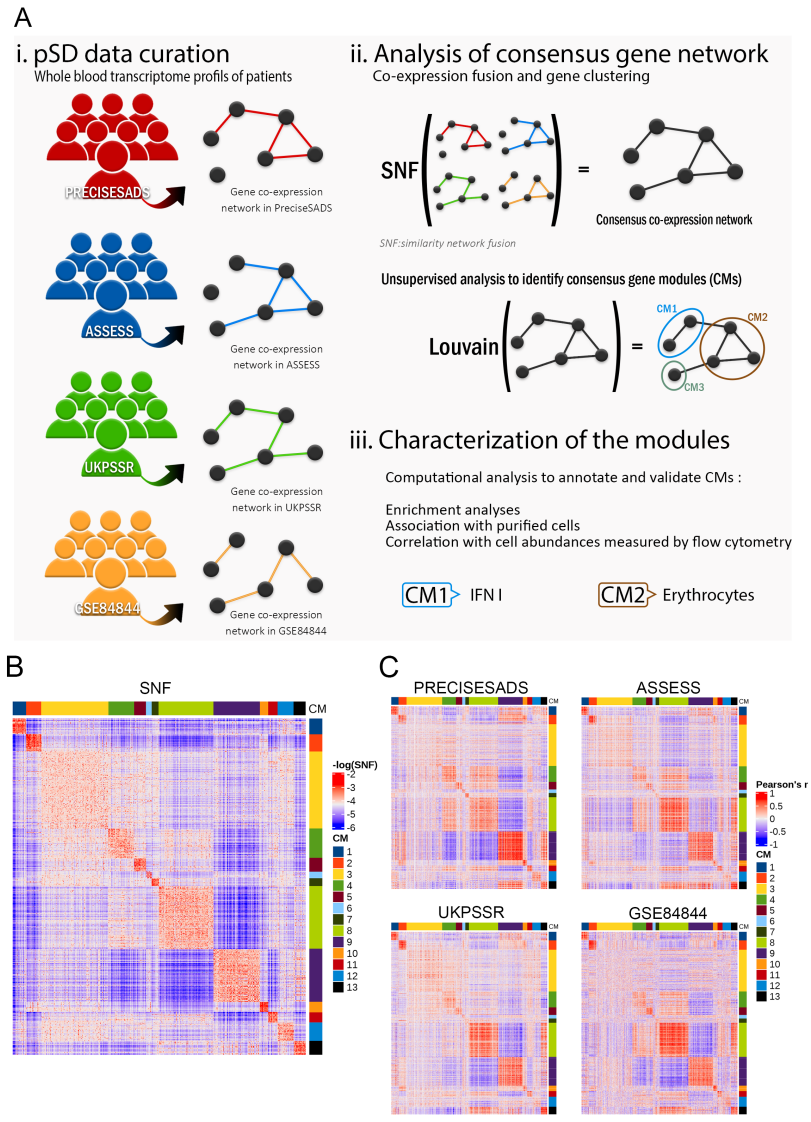
Résumé. Aspects méthodologiques de la médecine de précision avec application à la maladie de Sjögren

Dans le paysage en constante évolution de la recherche médicale, les deux dernières décennies ont été le témoin de progrès remarquables en biologie moléculaire, catalysés par des techniques innovantes telles que le séquençage de nouvelle génération. Un changement de paradigme s'est produit, éloignant la médecine de l'approche taille unique. La médecine de précision, con-cept basé sur la compréhension que l'état de chaque individu est distinct, a émergé. Les gènes et l'environnement s'entremêlent pour influencer la santé, les symptômes et les résultats du traitement, signifiant que des thérapies sur mesure sont essentielles. Un exemple est les sous-types moléculaires du cancer, associés à des traitements dédiés. Ce progrès a été souligné par la réalisation d'une signature pronostique approuvée par la FDA, mettant en lumière le potentiel transformateur de la médecine de précision. Cependant, avec l'expansion exponentielle des données médicales, il devient essentiel de relever les défis posés par leur volume, leur hétérogénéité et leur richesse.

Dans le domaine de la médecine de précision, deux méthodologies clés se démarquent : la stratification des patients et la prédiction des répondeurs au traitement. Mon projet de doctorat s'inscrit dans ce contexte dynamique, en mettant l'accent sur la maladie de Sjögren primaire (pSD). La pSD, un trouble auto-immun, affecte les glandes productrices d'humidité, entraînant une sécheresse de la bouche et des yeux, ainsi que d'autres symptômes débilitants. La variabilité tant dans les aspects biologiques que cliniques chez les patients a posé d'importants obstacles, laissant les patients atteints de pSD sans cure définitive. Face à ce défi, le projet IMI2 NECESSITY présente une plateforme collaborative où le monde académique et l'industrie convergent pour approfondir notre compréhension de la maladie et dévoiler de nouveaux marqueurs cliniques et moléculaires pour des essais potentiels. Guidé par cet effort collectif, le projet entreprend des trajectoires doubles.

Premièrement, l'étude capitalise sur diverses tentatives de stratification des patients atteints de pSD (cf. Tarn et al, Soret et al., et Trutschel et al.), provenant de cohortes indépendantes et englobant des données omiques et cliniques variées. En se concentrant sur l'obtention d'un consensus similaire à des efforts réussis dans le domaine de l'oncologie, cette approche navigue dans la complexité de la pSD, cherchant à établir des groupes de patients robustes. En utilisant la boîte à outils complexe de l'immunologie des systèmes, le projet se plonge dans le décodage des réseaux moléculaires perturbés, révélant les moteurs cachés de l'immunité et dévoilant leurs manifestations cliniques

en aval. La **figure 1** résume le pipeline construit lors de la thèse.



.

Pour surmonter les difficultés dans l'interprétation du transcriptome dans le contexte de la pSD, nous avons analysé conjointement quatre ensembles de données transcriptomiques indépendants profilant des échantillons de sang

total de 265 patients atteints de pSD. Nous avons utilisé des méthodes de regroupement pour identifier les principaux axes de variation à travers ces quatre ensembles de données. Comme les algorithmes de regroupement sont sensibles au bruit, nous avons mis en place une méthode pour effectuer une analyse de regroupement de gènes sur une représentation conjointe de la matrice de corrélations géniques par paire à travers les quatre ensembles de données, plutôt que sur chaque ensemble de données séparément. Pour ce faire, nous avons remodelé les quatre matrices observées de corrélations géniques par paire sous forme de graphes et utilisé l'algorithme Similarity Network Fusion pour obtenir une représentation graphique consensuelle du réseau de corrélations géniques à travers les quatre cohortes, sur laquelle nous avons appliqué l'algorithme de regroupement de graphes Louvain. Nous avons montré de manière significative que les modules géniques que nous avons identifiés sont reproductibles à travers les quatre cohortes sur lesquelles ils ont été découverts, ainsi que sur une cohorte indépendante. Ces modules de gènes représentent donc les principales caractéristiques biologiques contenues dans le profil transcriptomique du sang total chez les patients atteints de pSD, facilitant ainsi son interprétation pour la recherche translationnelle.

Afin de rendre les CM (consensus modules) plus biologiquement significatifs, nous les avons interprétés en utilisant différentes bases de données publiques de voies et de transcriptomes de cellules sanguines. Cela nous a permis d'identifier à la fois des modules fonctionnels (signalisation de l'interféron ou prolifération cellulaire) et des modules reflétant la composition cellulaire du sang des patients. De manière significative, nous avons observé des corrélations très significatives entre l'expression des modules géniques et les fréquences cellulaires correspondantes ou les niveaux de cytokines, validant ainsi ces interprétations biologiques dérivées de manière computationnelle. Ces dernières années, des méthodes de déconvolution transcriptomique ont été proposées pour estimer les proportions cellulaires à partir de mesures transcriptomiques. La plupart de ces méthodes reposent sur des profils transcriptomiques moyens de référence de types cellulaires, généralement dérivés de cellules purifiées du sang de donneurs sains, et utilisent des gènes discriminants entre les populations cellulaires dans un contexte donné, tel que le cancer. En revanche, notre approche est guidée par les variations observées dans le sang des patients atteints de pSD à travers plusieurs cohortes, garantissant que les signatures géniques des types cellulaires identifiés sont valides dans ce contexte. De plus, cette approche axée sur les données nous a permis de définir des modules géniques indicatifs de populations cellulaires rares

telles que les éosinophiles ou des signatures de types cellulaires non immunitaires tels que les érythrocytes ou les plaquettes, qui ne sont généralement pas quantifiées par les algorithmes de déconvolution. De plus, nous avons trouvé des modules fonctionnels (CM1 type 1 IFN et CM7 Cycle cellulaire) qui ne correspondent pas aux variations dans les fréquences des types cellulaires sanguins. Les modules géniques consensuels décrits ici pourraient donc aider à comprendre la physiopathologie complexe de la pSD, car ils représentent des sources de l'hétérogénéité dans le transcriptome sanguin des patients atteints de pSD, biologiquement significatives, reproductibles et sensibles.

Les modules géniques que nous avons identifiés peuvent servir de base pour la recherche translationnelle en pSD, en fournissant une liste concise de biomarqueurs potentiels fournis par le profil transcriptomique du sang total. Plusieurs études indépendantes se sont récemment concentrées sur la stratification de la maladie en sous-groupes distincts de patients, basée sur les profils transcriptomiques du sang total ou les caractéristiques cliniques. Ces systèmes de classification peuvent devenir pertinents dans les futurs essais cliniques, car de nouveaux traitements peuvent bénéficier uniquement à un sous-ensemble restreint de patients. Notre approche complémente ces classifications en mettant en évidence les différences de composition fonctionnelle et cellulaire entre les sous-groupes de patients, ainsi qu'en soulignant les consensus et les différences entre les systèmes de classification. Nos analyses suggèrent notamment que les sous-groupes de patients dans les systèmes de stratification publiés basés sur le transcriptome peuvent être distingués en fonction de la mesure de trois variables : la fréquence des neutrophiles dans le sang périphérique, la concentration d'interféron de type 1, ainsi que la fréquence des érythrocytes ou des plaquettes dans le sang.

Deuxièmement, la recherche vise à identifier les facteurs de réponse au traitement à travers des données historiques d'essais cliniques au sein du consortium IMI2 NECESSITY. En ciblant les patients répondeurs et en affinant les critères cliniques, cette initiative contribuerait à renforcer la sensibilité dans les essais futurs. L'accès à des ensembles de données uniques et l'interaction directe avec les principaux intervenants du consortium IMI 2 NECESSITY et les initiatives internes de Servier soulignent la nature interdisciplinaire et collaborative du projet.

Ce doctorat aborde non seulement la complexité de la pSD, mais résonne également avec la mission plus large de faire progresser la médecine de précision, où des informations basées sur les données ouvrent la voie à des paradigmes de soins individualisés. À mesure que les efforts interdis-

ciplinaires convergent, le potentiel de stratégies thérapeutiques innovantes dans le domaine des troubles auto-immuns devient palpable. L'avenir de la recherche sur la maladie de Sjögren primaire semble prometteur, avec de nombreuses possibilités en attente d'exploration à travers des outils de pointe tels que la séquence unique de l'ARN et/ou la transcriptomique spatiale. De plus, l'exploration de données plus spécifiques aux tissus pourrait significativement améliorer notre compréhension des mécanismes d'action à différents niveaux cellulaires, ouvrant potentiellement la voie à de nouvelles stratégies thérapeutiques dans la gestion de la Sjögren.

Simultanément à ces découvertes, je tiens à souligner le développement de plusieurs outils informatiques. Bien que je n'aie pas rédigé d'articles méthodologiques autonomes, il est à noter que les données de séquençage de l'ARN, comme discuté précédemment dans cette thèse, présentent des défis de complexité, de taille et d'interprétation. Par conséquent, en collaboration, j'ai conçu et validé des outils et méthodologies informatiques spécifiques adaptés à l'analyse des données générées dans le cadre de mes différents projets.

Mots clés : Médecine de précision, Maladie du Sjögren, Modules de Genes.

Contents

1 Introduction to immunology

Immunology is a captivating branch of life science that I discovered when I started my PhD. It delves into the remarkable defense mechanisms orchestrating our body's protection against pathogens and diseases. This field explores the intricate interactions between the immune system and various agents, aiming to understand how immunity is acquired, maintained, and regulated. From the fundamental states of immune cells, such as T cells, B cells, and macrophages, to the production of antibodies and cytokines, immunology unravels the fascinating complexities that underpin our ability to fend off infections. By comprehending the mechanisms of studying the physiopathology of diseases, we gain invaluable insights into vaccine development, immunotherapy, and overall health maintenance.

This section presents a comprehensive overview of the fundamental principles and essential knowledge in immunology. It covers the basics and core concepts that are crucial for understanding the intricacies of the immune system.

1.1 Key components of the immune system

Although not intended to be an exhaustive subsection of the human immune system, this subsection highlights pertinent aspects relevant to my PhD. It covers both the innate and adaptive immune responses, which depend on a vast network of specialized immune cells grouped into seven major categories. For further in-depth understanding, I recommend the comprehensive textbook "Janeway's Immunology" [1].

1.1.1 Cellular components

1. Granulocytes (Neutrophils, Eosinophils and Basophils), the most abundant immune cells in blood, are phagocytes with the ability to engulf and eliminate harmful foreign agents and dying cells. As crucial components of the innate immune system, characterized by cytoplasmic granules housing cytotoxic molecules, they swiftly migrate to infection sites to combat and destroy pathogens, playing a pivotal role in the initial defense against infections [2, 3].

2. Monocytes, comprising approximately 10% of peripheral leukocytes in humans, migrate to the site of infection during infections. There, they

transform into effector cells, specifically macrophages and dendritic cells, while simultaneously releasing substantial quantities of inflammatory molecules, thereby activating and sustaining the immune response [4].

3. Macrophages, distinct from circulating cells, reside in various organs as large phagocytes involved in both innate immunity and diverse biological processes. Apart from pathogen elimination like neutrophils, they perform tasks such as dead cell removal, iron and fatty acid metabolism. Depending on their location, they take on different names and unique functions, including Kupffer's cells in the liver, Langerhans' cells in the skin, and microglia in the brain. Derived from monocytes as they enter tissues from the peripheral blood, macrophages serve as effector cells. Their primary function centers around phagocytosis, internalizing and destroying cells, bacteria, or small bodies through acidification, proteases, and reactive oxygen species. Additionally, macrophages can present antigens to CD4+ T lymphocytes at low levels [5].

4. Dendritic cells (DCs) are specialized antigen-presenting cells (APC) that bridge the innate and adaptive immune systems. Through phagocytosis, they process pathogens, presenting antigens to T lymphocytes. This activation of antigen-specific T cells facilitates adaptive immunity. With their long dendrites sensing molecular patterns, DCs internalize extracellular antigens, presenting them to CD4+ and CD8+ T cells via antigen cross-presentation. Additionally, DCs employ macro- pinocytosis to process antigenic content from small volumes of extracellular fluids. Their maturation state and received stimuli influence the expression of co-stimulatory or co-inhibitory ligands and various cytokines, shaping the functionality of the immune cells they stimulate [6].

5. T cells, distinct from previously described myeloid cells, are lymphocytes with a relatively small size (5-10 μm diameter), and a large nucleus. Their T-Cell Receptor (TCR) specifically identifies antigens presented by DCs or macrophages. Each T cell possesses a unique TCR generated through controlled DNA recombination and mutation, enabling recognition of specific antigens. Upon activation by a cognate antigen, T cells proliferate and express effector molecules, leading to functions like cytotoxicity (CD8+ T cells) or regulating other immune cells (CD4+ T cells). T lymphocytes differentiate from myeloid cells in

the bone marrow and migrate to the thymus. Thymocytes within the thymus undergo random DNA rearrangement, resulting in diverse TCR sequences. TCRs can sense peptides displayed by Major Histocompatibility Complex (MHC) molecules. Thymocytes are selected based on positive interactions with MHC molecules and are negatively selected against self-antigens. T cells that meet these criteria leave the thymus and circulate throughout the body [7].

6. Similar to T cells, each B cell expresses a randomly generated receptor by V(D)J recombination, the B-cell receptor (BCR). Unlike the TCR, the BCR does not require the presentation of peptides by MHC molecules to exert its function. The BCR resembles a membrane-bound antibody. Upon recognizing a soluble antigen, the B cell undergoes active division and generates a soluble form of the BCR called an antibody. Antibodies play a vital role in neutralizing viruses by preventing them from infecting host cells and also enhance the phagocytosis of bacteria and larger pathogens by macrophages. In the next chapter, a more comprehensive explanation of antibodies will be provided. B cells, along with T cells, constitute the two arms of the adaptive immune response [8].

7. Innate lymphoid cells (ILCs) have been recently identified due to their limited presence in the blood and resemblance to T lymphocytes. These lymphocytes lack the expression of encoded receptors like TCR or BCR. The predominant ILCs are Natural Killer (NK) cells, functioning as cytotoxic agents, capable of eliminating infected, stressed, and cancerous cells in a non-specific manner, independent of antigen recognition [9].

Each of the previously described cell type in addition to being located in several places, they possesses distinct properties and functions, making them individually insufficient to provide complete protection against pathogens (**Figure 1**). Therefore, immune cells must engage in close communication to orchestrate an effective immune response.

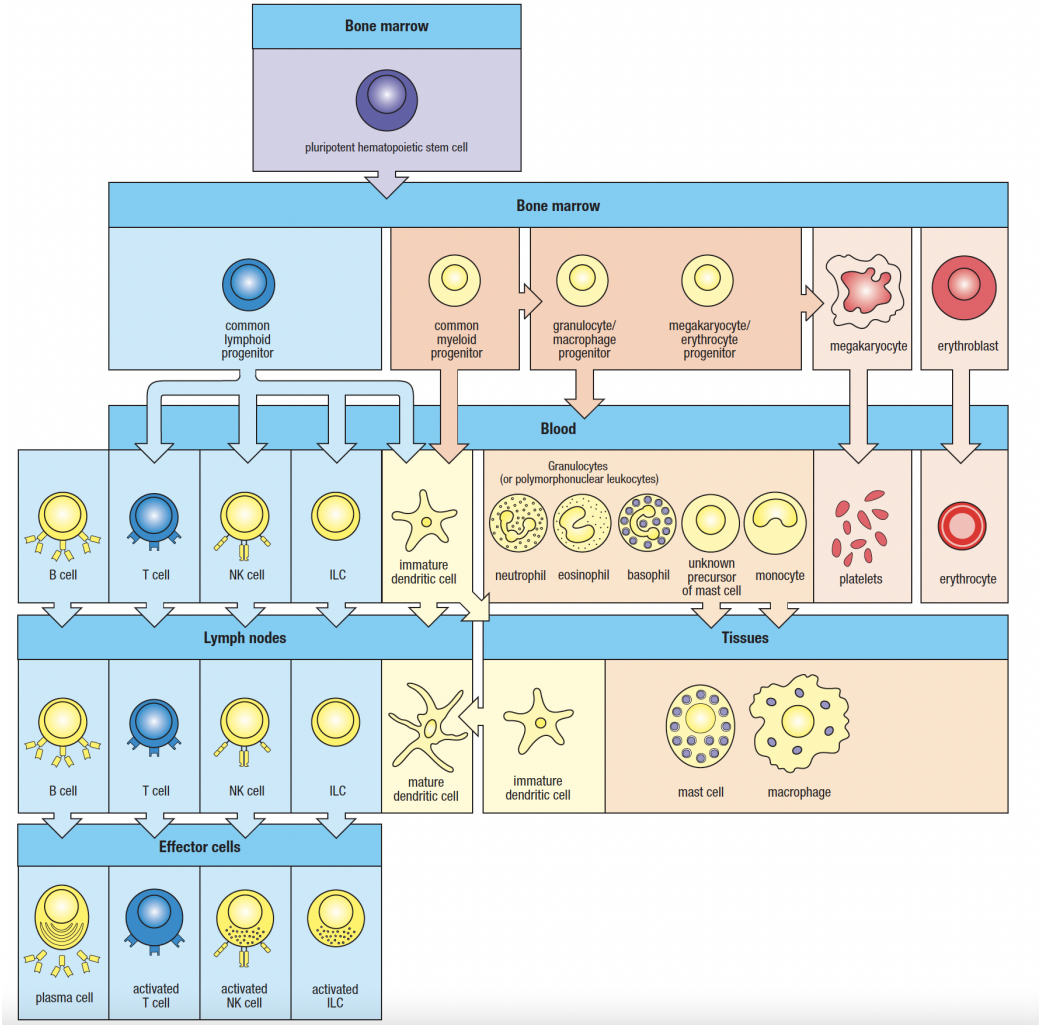


Figure 1: Cellular elements of the blood, including immune system cells, originate from pluripotent hematopoietic stem cells within the bone marrow. From (Murphy, 2017)

1.1.2 Cell communication

All cells possess the capacity to adapt to their ever-changing surroundings. This adaptation involves three key stages: sensing environmental signals/changes, processing signals, and modifying cellular processes. Unicellu-

lar organisms primarily respond to changes in local nutrients or temperature while cells in multicellular organisms must integrate multiple signals from neighboring cells. These cellular communications regulate various aspects of cell behavior—survival, division, metabolism, and movement—crucial to multicellular biology. Communication between cells is predominantly facilitated by soluble factors like proteins, lipids, carbohydrates, and sometimes gases. Some of these molecules function across great distances, while others impact nearby cells. Cells detect these factors through receptors, often situated at the cell membrane, capable of binding to the soluble factor. The binding of factors (alias, ligands) triggers physical or chemical changes in the receptor, a process commonly referred to as receptor activation. Hereafter, I present a few examples of the vast and intricate network of communication factors that contribute to the complex regulation of the immune system.

1. Interleukins (IL) primarily mediate communication between leukocytes and regulate various immune responses. One of the well-known interleukins is interleukin-6 (IL-6), which is involved in both pro-inflammatory and anti-inflammatory responses. It has been shown to play a critical role in the acute-phase response during infections and tissue injury [10]. Additionally, interleukin-7 (IL-7) is another important interleukin as it acts as a growth factor for lymphocytes, particularly T cells and B cells, promoting their survival, proliferation, and differentiation [11].

2. Chemokines (proteins of the CCL or CXCL families) regulate the migration and recruitment of immune cells to sites of infection or inflammation. Studies have explored their role in diseases such as cancer, autoimmune disorders, and infectious diseases, providing potential targets for therapeutic interventions [12, 13].

3. Tumor necrosis factor (TNF), also known as cachexin or cachectin, was previously referred to as tumor necrosis factor alpha (TNF-α). It functions as both an adipokine and a cytokine. Belonging to the TNF superfamily, TNF shares a common TNF domain with various transmembrane proteins. In its role as an adipokine, TNF contributes to insulin resistance and is implicated in obesity-related type 2 diabetes. As a cytokine, TNF plays a vital role in immune system communication. When macrophages, a type of white blood cell, detect an infection, they release TNF to signal other immune cells, initiating an inflammatory response.

4. The group of Type I interferons forms the most extensive Interferon category (IFN). Among humans, this group encompasses IFN-α, IFN-β, IFN-ϵ, IFN-κ, and IFN-ω, all of which are clustered on chromosome 9 (in human) and transmit signals through the Type I IFN heterodimeric receptor complex consisting of IFN-α receptor 1 (IFNAR1) and IFNAR2 subunits [14]. Each Type I IFN is encoded by a sole gene except for IFN-α, which contains 13 subtypes in humans [15]. Type I IFNs, as cytokines, impact the expression of numerous genes, resulting in profound changes within cells. IFN-α activates the cell by dimerizing its two-receptor chains, IFNAR1 and IFNAR2, both of which are present on all nucleated cells. Virtually every cell has the capability to produce IFN-α/β; however, during the course of an infection, dendritic cells produce the vast majority of IFN-α [16, 17]. The diverse Type I IFNs exhibit varying tissue expression patterns and binding affinities for the IFNAR1/2 receptor complex, thereby leading to distinct subtypes giving rise to varied outcomes in terms of antiviral, antiproliferative, and immunomodulatory activities. The study of IFNs in autoimmune diseases is a significant area of research aimed at understanding the role of these signaling proteins in immune dysregulation and disease pathogenesis. In the subsequent subsection, I will present an example of interferon signaling pathway.

This non-exhaustive list of communication factors comprises essential signaling molecules produced by various immune cells that play a crucial role in mediating communication and coordination within the immune system. These small proteins act as messengers, relaying information between cells to regulate immune responses, inflammation, and immune cell activation. Cytokines can have diverse effects, stimulating or suppressing immune activities, and their intricate interactions orchestrate the finely-tuned defense mechanisms of the body, contributing significantly to maintaining overall health and combating infections and diseases.

1.1.3 Description of common signaling pathways

A biological pathway is a series of interconnected molecular events and interactions within a cell or organism that collectively drive a specific biological process or function. These pathways involve the coordinated activity of various molecules, such as proteins, enzymes, and metabolites, which work

together to achieve a particular cellular response or outcome. Over the years, multiple signaling pathways have been identified and studied using cellular and molecular biology as well as genetic tools. Describing the multitude of signaling pathways identified in mammals would require an entire book. Thus, only the crucial pathways to be later discussed in this manuscript will be detailed in this subsection. As previously indicated in the preceding subsection, cell signaling hinges on the interaction between proteins and ligands. A significant proportion of cytokine receptors belong to the Receptor Tyrosine Kinase (RTK) family, particularly within the context of the JAK-STAT pathway. These receptors are directly associated with cytoplasmic tyrosine kinases named Janus Kinase (so called because they have two tandem kinase-like domains and thus resemble the two-headed mythical Roman god Janus), which phosphorylate and stimulate the transcriptional activity of effectors known as STAT (Signal Transducer and Activator of Transcription).

Type I and type II interferon-mediated signalling pathway

The initial step in both type I and type II interferon mediated signalling is the activation of these receptor-associated JAKs, which occurs in response to a ligand-dependent rearrangement and dimerization of the receptor subunits, followed by autophosphorylation and activation of the associated JAKs (see **Figure 2**). As well as activation of classical JAK–STAT (signal transducer and activator of transcription)-signalling pathways (discussed later), activation of IFN-receptor-associated JAKs seems to regulate, either directly or indirectly, several other downstream cascades. Such diversity of signalling is consistent with the pleiotropic biological effects of IFNs on target cells and tissues.

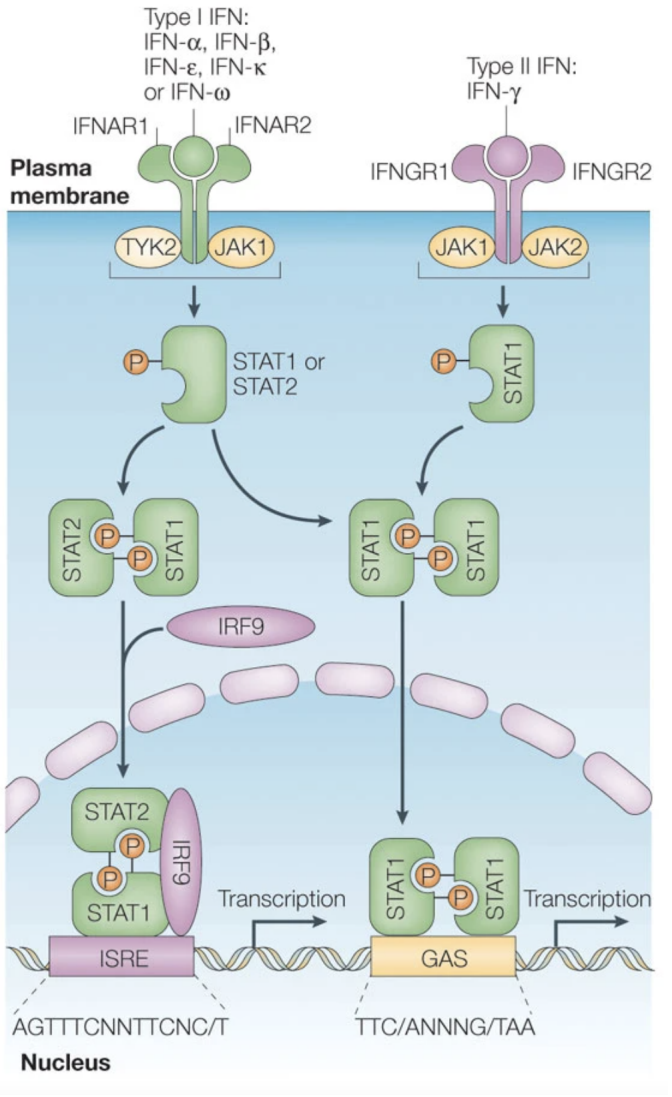


Figure 2: The JAK-STAT pathway and its activation by cytokines (Type I IFN or Type II IFN) on the plasma membrane. The binding of the cytokines to two receptor monomers allows the associated JAK to be close enough to phosphorylate each other on tyrosines and make them fully active. Once activated, they phosphorylate the receptor itself allowing the recruitment and subsequent phosphorylation of STAT proteins. The phosphorylated STAT proteins form a dimer that can enter the nucleus and activates the transcription of specific genes. From (Platanias, 2005)[18]

The NF-κB pathway While a considerable portion of cytokines exert their effects on cells through the JAK-STAT signaling pathways, several cytokines rely on an entirely distinct pathway known as the NF-κB pathway (for Nuclear Factor kappa-light-chain enhancer of activated B cells) (**Figure 3**). This pathway can be activated by various cytokines like Tumor Necrosis Factor alpha (TNF-α) or Interleukin 1 beta (IL1β) through their respective receptors. Additionally, it can also be triggered by the binding of pathogen components to specific proteins called Toll-like receptors (TLRs). Upon the binding of ligands to receptor monomers, these monomers assemble into dimers or trimers, leading to significant conformational changes that enable the phosphorylation of the IKK complex. Following phosphorylation, the IKK complex gains the capability to phosphorylate another protein complex composed of two NF-κB proteins and the inhibitory IκB protein. The phosphorylation of this complex prompts the degradation of IκB and the release of the NF-κB dimer, which then migrates to the nucleus to activate the transcription of inflammatory genes. Analogous to STAT proteins, five distinct NF-κB proteins have been identified in mammals, contributing to a diverse range of regulatory possibilities.

1.2 Systems Immunology, when computational science and life science meet

Microscopy serves as the foundational pillar of cellular biology. Originating and evolving in the 17th century, it has illuminated the concept of cells as the fundamental building blocks of life. Notably, it has also exposed the remarkable array and divergence of cells within an organism. This revelation underscores the essential imperative to explore such diversity in order to fathom intricate biological phenomena. In contrast to microscopy, molecular biology and genomic methodologies have the capacity to comprehensively analyze thousands of attributes. For instance, RNA expression levels can be gauged via RNA sequencing (RNA-seq), chromatin accessibility can be explored using Assay for Transposase-Accessible Chromatin with sequencing (ATAC-seq), and DNA mutations can be detected through DNA sequencing (DNA-seq). However, this breadth of analysis comes at the expense of cellular resolution: these approaches usually demand hundreds of cells, thereby limiting their utility in probing cellular heterogeneity. This paradigm underwent a profound transformation with the emergence of the genomic field,

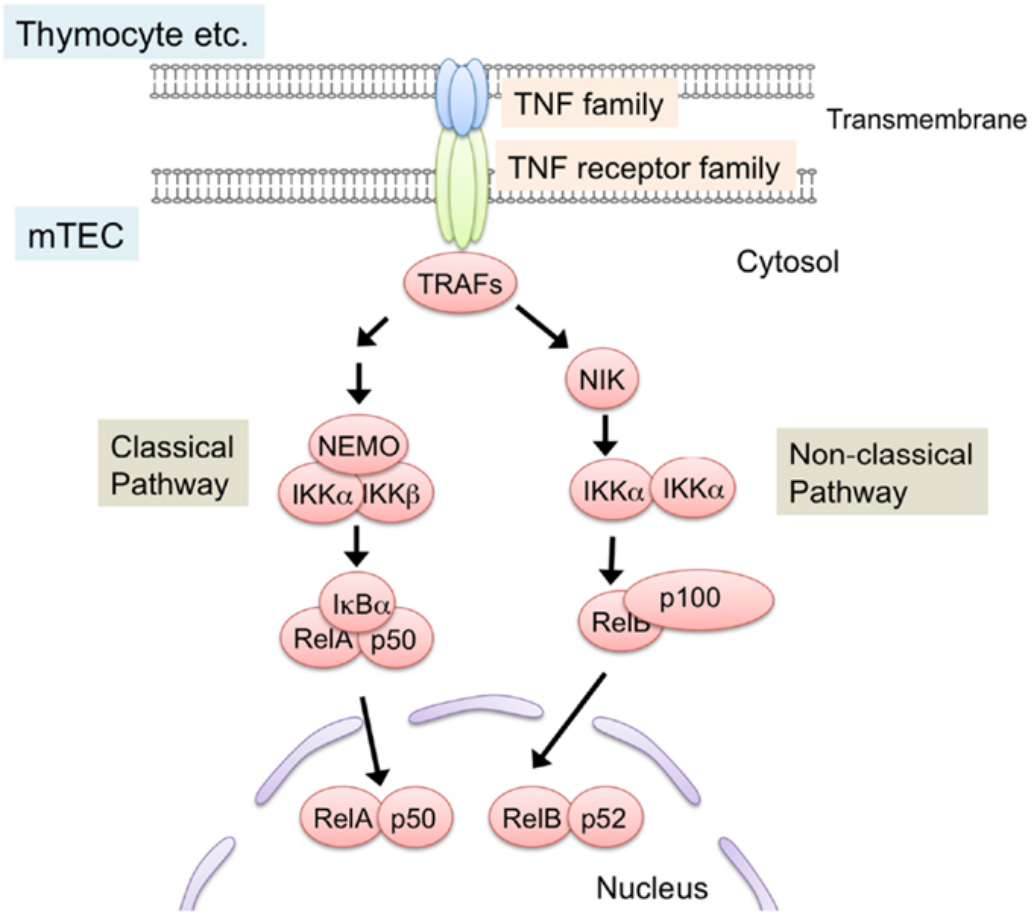


Figure 3: The NF-κB pathway and its activation by TNF family. Interaction between the ligand and its receptor induce the binding of TRAF-family proteins to the cytoplasmic domains of TNF receptors. TRAF-family proteins in turn activate downstream serine/threonine kinase cascade. These kinases trigger the degradation of inhibitory proteins that sequester NF-κB in cytosol, thereby leading to the translocation and transcriptional activation of NF-κB members. NF-κB pathways are classified into classical and non-classical pathway. From (Akiyama, 2013)[19]

seamlessly integrating the advantages of both microscopy (single-cell precision) and genomics (extensive attribute measurement). All these concepts

can be categorized as offshoots of systems immunology, an innovative field at the forefront of immunological research. This field merges biology, data science, and computational techniques to comprehend the immune system's complexity as a dynamic network. By scrutinizing high-dimensional data and interactions among immune components, systems immunology offers deeper insights into immunity. This, in turn, leads to novel approaches in understanding diseases and developing therapeutic interventions.

1.2.1 High-throughput technologies

Systems immunology relies heavily on high-throughput technologies, such as next-generation sequencing, microarrays, mass spectrometry, and flow cytometry, which allow researchers to analyze large amounts of data simultaneously (**Figure 4**). These techniques enable the generation of vast datasets that capture the diverse elements and interactions within the immune system. Over the last century, numerous effective experimental approaches have emerged, playing a pivotal role in delineating distinct cell types and states within the immune system. These strategies have unveiled fundamental molecular and functional elements of immunity and have elucidated cause-and-effect connections underlying transcriptional and functional sequences driving immune activation

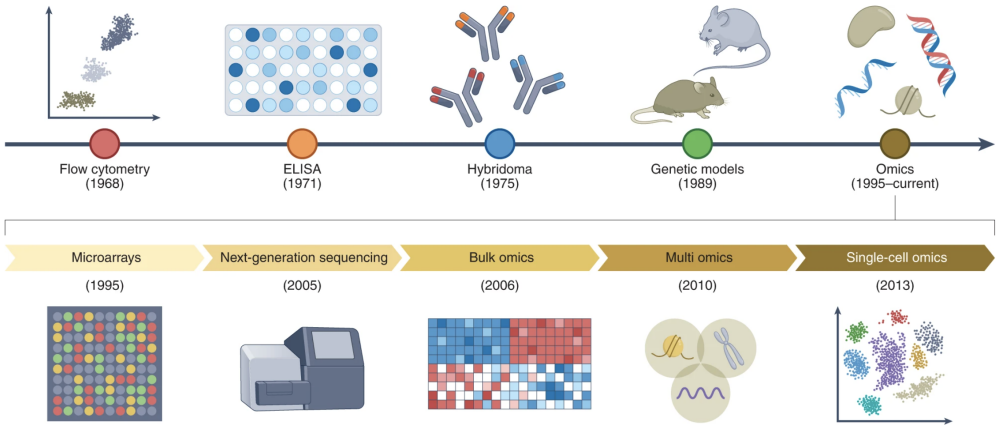


Figure 4: Chronology highlighting the pivotal technological advancements in the realm of immunology research. From (Bonaguro, 2022)[20]

1.2.2 Balancing exploratory data analysis and hypothesis-driven science in omics-based immunological research

While exploratory analysis of data plays a crucial role in initially comprehending data patterns and uncovering potential biases, it is advocated for researchers to adhere to established principles of hypothesis-driven science, as outlined in the proposed systems-immunology cycle (**Figure 5**)[21]. This cyclical approach entails the utilization of multi-omics technologies in conjunction with the formulation of hypotheses or research questions, combined with traditional experimental designs (such as loss-of-function or gain-of-function experiments, specific clinical cohorts, and trials like vaccine or immunotherapy studies). These endeavors aim to establish insights into immune functionality, molecular phenotypes, as well as predictions related to immunotherapy outcomes.

Although some view hypotheses as restrictive, it is emphasized their pivotal role in guiding research. While it is true that hypotheses in omics-based immunological investigations can sometimes be vaguely formulated, such as proposing broad transcriptional disparities across multiple peripheral immune cells in a case–control study of an inflammatory disease, it is contended that an approach centered around hypotheses aids scientists in formulating and concentrating on core questions [22]. Importantly, this approach doesn't impede the potential for autonomous discovery and the generation of novel hypotheses during the secondary utilization of data.

The key distinction between this comprehensive strategy and traditional reductionist experimentation lies in the necessity for mathematical and computational modeling of extensive datasets. This phase of the process might be labeled as 'data-guided.' Yet, even with advanced multi-omics technologies, the absence of a well-defined hypothesis and a robust experimental framework poses a potential challenge to achieving meaningful outcomes. In contrast, approaches grounded in hypotheses and carefully planned experimental designs yield high-resolution omics data that furnish valuable, and sometimes unexpected, biological insights. Such an approach mitigates the likelihood of setbacks, facilitates the selection of subsequent inquiries, and supports validation studies.

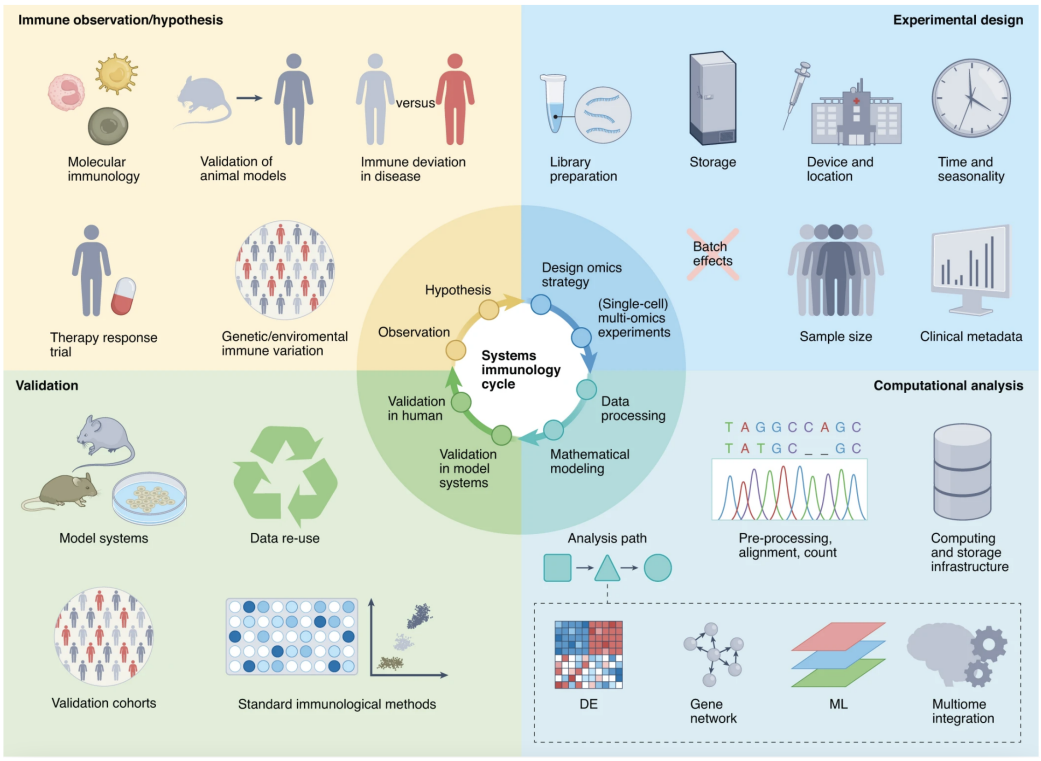


Figure 5: From (Bonaguro, 2022)[20]

Overall, systems immunology has revolutionized the field of immunology by providing a holistic approach to understanding the complexities of the immune system. It has the potential to lead to significant advances in the diagnosis, treatment, and prevention of various immune-related diseases.

2 Autoimmunity

The immune system possesses potent mechanisms to combat various pathogens, but if misdirected towards the host, it can lead to tissue damage. Autoimmunity, proposed by Paul Ehrlich, involves specific immune responses against self antigens, resulting in chronic autoimmune diseases. While autoimmune disorders affect around 5% of Western populations, the immune system has evolved mechanisms to prevent self-injury, primarily by distinguishing self from non-self. Initial distinctions occur during lymphocyte development, promoting self-tolerance through central tolerance mechanisms. Mature lymphocytes in the periphery are activated by certain signals, preventing self-reactivity. For deeper notions, I recommend the book "The Autoimmune Diseases" by N. Rose, widely regarded as the father of autoimmune disease research[23].

2.1 Navigating the realm of autoimmunity terminology

Bellow, I will provide a concise yet comprehensive overview of fundamental concepts related to autoimmunity. Recognizing that delving into these concepts all at once can be overwhelming, I aim to present a non-exhaustive list of definitions that will serve as a foundation for better understanding the intricacies of autoimmunity.

- **Antigens** (Ag) are molecules that can trigger an immune response in the body. They are recognized by the immune system as foreign or non-self, and the immune system produces specific antibodies or activates immune cells to target and eliminate them. Autoantigens are a specific subset of antigens that are derived from the body's own tissues or cells. In autoimmune diseases, the immune system mistakenly identifies these self-antigens as foreign and mounts an immune response against them, leading to damage and inflammation within the body's own tissues. In the realm of autoimmune disease, a crucial inquiry revolves around determining whether the process operates independently or is influenced by antigens. If it is the latter case, the key point of interest is to ascertain whether the antigen responsible is of self-origin or foreign in nature.

- An **antibody**, also known as an immunoglobulin, is the Y-shaped protein produced by the immune system in response to the presence of anti-

gens in the body. Antibodies play a vital role in the immune response by recognizing and binding to specific antigens, such as those found on pathogens like bacteria, viruses, or toxins. This binding marks the antigens for destruction by other immune cells, effectively neutralizing or eliminating the threat. Antibodies are a crucial component of the adaptive immune system and provide long-lasting protection against infections and diseases. On the other hand, an autoantibody is an antibody that mistakenly targets and attacks the body's own healthy cells and tissues. In autoimmune diseases, the immune system produces autoantibodies against self-antigens, which are components of the body's own cells. This immune response leads to inflammation and damage to various organs and tissues, causing the characteristic symptoms of autoimmune disorders.

- In order to ensure protection against infectious agents, the immunological repertoire of T cells and B cells must be diverse enough to recognize all foreign antigens (Ags). The plasticity of the T cell receptor (TCR) and B cell receptor (BCR) for Ags follows somatic gene modification steps for these receptors. As previously described, the first step involves the recombination of V, D, and J gene segments encoding the variable domains of TCR and BCR during T and B lymphocyte maturation. The second step pertains solely to BCR and occurs in the periphery following antigenic encounter. This is known as somatic hypermutation, which enables the production of higher-affinity receptors. The trade-off for this diversity is the production of TCRs and BCRs that can recognize self-antigens (self-Ags). Thus, between 20 and 50 % of TCRs and BCRs potentially recognize self-Ags. **Immunological tolerance** is a characteristic of the adaptive immune system, signifying the absence of specific immune reactivity towards a recognized Ag[24]. Inactivation of auto-reactive clones involves mechanisms of immune tolerance that affect both T and B cells. Classically, tolerance mechanisms are classified based on their anatomical location into two main categories: central tolerance, occurring in central lymphoid organs (thymus for T cells and bone marrow for B cells), and peripheral tolerance, taking place in secondary lymphoid organs (spleen, lymph nodes, mucosa-associated lymphoid tissue).

  - Central Tolerance: This occurs during the development of lym-

phocytes. Lymphocytes that react strongly against self-antigens are either eliminated (clonal deletion) or rendered non-functional (anergy) to prevent them from causing autoimmune responses.

- Peripheral Tolerance (negative selection): This mechanism operates in the mature lymphocytes found in the peripheral tissues. Regulatory T cells (CD4+) play a vital role in maintaining self-tolerance by suppressing the activity of autoreactive lymphocytes. Additionally, peripheral tissues may have low levels of costimulatory molecules, making it less likely for lymphocytes to be activated by self-antigens.

Breakdown of self-tolerance can occur when these mechanisms fail, leading to the activation of immune responses against the body's own tissues. This can result in autoimmune diseases where the immune system attacks healthy cells, tissues, or organs, causing inflammation and damage. The exact reasons for the breakdown of self-tolerance can vary and might involve genetic predisposition, environmental factors, or dysregulation of the immune system's checks and balances.

Together, these mechanisms ensure that the immune system can differentiate between self and non-self antigens, preventing harmful autoimmune responses while maintaining effective defense against foreign invaders.

By establishing this essential groundwork, we are better equipped to explore the intricate details of autoimmunity that lie ahead, fostering a deeper comprehension of this fascinating field.

2.2 Causes of autoimmunity

For the immune system to distinguish self from nonself is not trivial, as it requires a delicate balance between safeguarding against autoimmune disease and maintaining immune effectiveness. The causes of autoimmune diseases remain unclear. Genetic risk factors, like certain alleles of MHC class II molecules and gene mutations, play a role, but not all predisposed individuals develop the disease. Environmental factors also contribute, but their impact is not well understood. Toxins, drugs, infections, and molecular mimicry are potential triggers. Research is needed to identify specific environmental

influences, possibly a combination of factors or chance events. Below, I list a few examples can lead to autoimmunity

2.2.1 External events can initiate autoimmunity

The distribution of autoimmune diseases across geographical regions demonstrates significant heterogeneity across continents, countries, and ethnic groups. An illustrative example is the varying disease incidence in the Northern Hemisphere, which tends to decline from north to south. This phenomenon is particularly conspicuous in Europe for conditions like multiple sclerosis and type 1 diabetes, where higher incidence rates are observed in northern nations compared to Mediterranean areas. A plethora of epidemiological and genetic connections strongly imply that this pattern is potentially influenced, at least in part, by levels of vitamin D [25]. Vitamin D's active form is generated in the skin upon sunlight exposure, and its availability decreases with latitude. This nutrient possesses multiple immune-regulatory roles that impact both innate and adaptive immune cells, including the inhibition of TH17 cell development [26]. Furthermore, research indicates an elevated prevalence of autoimmunity in more developed countries [27], although the precise underlying factors remain enigmatic. In addition to vitamin D levels, a multitude of other non-genetic factors contribute to these geographical discrepancies, encompassing aspects like socioeconomic status and dietary habits. The influence of these non-genetic factors becomes evident in situations involving genetically identical mice, where variations in the rates and severity of autoimmunity occur.

2.2.2 Infection can lead to autoimmune disease via molecular mimicry

Modified or altered genetic material from infections can trigger processes that lead to inflammation-induced cell death, such as pyroptosis and NETosis (Neutrophil extracellular traps). These infections can also cause the release of the host's nuclear autoantigens into the surrounding environment, and these antigens can then be identified by the immune system in a way that stimulates both the innate and adaptive immune responses [28]. Additionally, bacterial infections can result in the release of bacterial DNA along with other bacterial components. These complexes can initiate autoimmune reactions by serving as triggers for pattern recognition receptors, and they can activate self-reactive B cells through a mechanism similar to molecular mimicry.

2.2.3 Drugs and toxins can cause autoimmune syndromes

The most evident indication of external factors contributing to human autoimmunity is found in the impact of specific medications that trigger autoimmune responses in a small subset of patients. An example of this is Procainamide, a medication employed for treating heart arrhythmias, which stands out for its ability to generate autoantibodies resembling those seen in Systemic Lupus Erythematosus (SLE), although these autoantibodies are seldom pathogenic [29]. Other examples can be found with hydralazine [30] (used to treat high blood pressure or hypertension) and isoniazid [31] (used to treat tuberculosis). Additionally, a number of drugs are linked to the emergence of autoimmune hemolytic anemia [32], a condition where autoantibodies directed against surface components of red blood cells generate immune complexes, resulting in the destruction of these cells [33].

Autoimmunity can also be triggered by environmental toxins. For instance, when administered to susceptible strains of mice, heavy metals like mercury, silver, or gold can induce a consistent autoimmune syndrome characterized by the production of autoantibodies [34]. While the degree to which heavy metals contribute to autoimmunity in humans is a topic of debate, the evidence from animal models underscores the potential involvement of environmental factors, such as toxins, in certain syndromes.

2.2.4 An impaired cellular communication network

Aberrations in cytokine generation or signaling pathways can precipitate the onset of autoimmunity. Genetic investigations, primarily conducted in animal models, have unveiled certain signaling pathways linked to autoimmunity. The **Figure 6** from Murphy's book [1] encompasses the repercussions stemming from the excessive or insufficient expression of select cytokines and intracellular signaling components implicated in this context.

Defects in cytokine production or signaling that can lead to autoimmunity		
Defect	**Cytokine, receptor, or intracellular signal**	**Result**
Overexpression	TNF-α	Inflammatory bowel disease, arthritis, vasculitis
	IL-2, IL-7, IL-2R	Inflammatory bowel disease
	IL-3	Demyelinating syndrome
	IFN-γ	Overexpression in skin leads to SLE
	IL-23R	Inflammatory bowel disease, psoriasis
	STAT4	Inflammatory bowel disease
Underexpression	TNF-α	SLE
	IL-1 receptor agonist	Arthritis
	IL-10, IL-10R, STAT3	Inflammatory bowel disease
	TGF-β	Ubiquitous underexpression leads to inflammatory bowel disease. Underexpression specifically in T cells leads to SLE

Figure 6: Defective Cytokine Production and Signaling Associated with Autoimmunity. This table presents a comprehensive overview of cytokine-related dysregulation linked to the development of autoimmune diseases. From (Murphy, 2017)[1]

In conclusion, the intricate landscape of autoimmunity reveals a symphony of external triggers and influences that collectively contribute to the intricate development of autoimmune diseases. Geographical disparities in disease incidence, often linked to levels of vitamin D [25] and other non-genetic factors, underscore the role of environment in shaping autoimmunity. Infections, through mechanisms like molecular mimicry and the release of autoantigens [28], can ignite autoimmune responses, adding yet another layer of complexity. Additionally, medications and toxins exhibit the power to in-

duce autoimmune reactions, as exemplified by drugs like Procainamide and environmental factors such as heavy metals. This intricate interplay between genetic predisposition, environmental factors, and immune responses highlights the multifaceted nature of autoimmunity, offering a deeper understanding of its origins and potential therapeutic avenues. Further investigation is required to delineate the precise roles of environmental factors in the development of autoimmune diseases. It is conceivable that, in the majority of cases, identifying a solitary environmental trigger responsible for initiating disease might not be feasible . Instead, it could be the culmination of various triggers or even stochastic occurrences that play pivotal roles in this complex process.

2.3 Common autoimmune diseases

Shoenfeld et al. [35] have identified more than one hundred autoimmune diseases. The classification of these conditions remains uncertain, especially when a thorough understanding of their causal mechanisms is lacking. Rather than discrete entities, these diseases can be seen along a spectrum, distinguishing between 'organ-specific' autoimmune diseases targeting specific body organs, and 'systemic' autoimmune diseases affecting multiple tissues throughout the body. Both categories tend to become chronic due to persistent autoantigens, except for exceptions (e.g., Hashimoto's thyroiditis[36]). The classification of autoimmune diseases is not a simplistic endeavor. As emphasized by Barturen and colleagues (2018)[37], it is necessary to distinguish between "classification criteria" and "diagnostic criteria". Classification criteria are not designed to be used for diagnostic purposes, but rather in an epidemiological context or in studies where homogeneity and comparability among patient populations are of great importance. Nevertheless, it is common for classification criteria to be widely employed as diagnostic criteria. A recent notable development is the PreciseSADS cohort initiative led by Barturen et al. (2021)[38]. This initiative focuses on identifying molecular clusters to reclassify seven different systemic autoimmune diseases independently of clinical diagnosis. Using integrated transcriptome and methylome data, the study revealed four stable clusters – "inflammatory," "lymphoid," "interferon," and a low-activity cluster including healthy controls. Longitudinal analysis confirmed cluster stability over time. This innovative stratification holds implications for clinical trials and understanding treatment non-response, reshaping our perception of autoimmune diseases.

In the following section, concise explanations of ten prevalent autoimmune disorders – Rheumatoid arthritis (RA), Multiple sclerosis (MS), Psoriasis (PsO), Inflammatory bowel disease (IBD), Systemic sclerosis (SSc), Systemic Lupus Erythematous (SLE), and Sjögren's disease – will be provided, with detailed expansion in section 3.

2.3.1 Rheumatoid arthritis

(RA) is characterized by synovial inflammation, joint damage, and systemic effects [39, 40]. This inflammation leads to joint destruction, disability, and shortened life span. Genetic predisposition, immune dysregulation, and environmental factors contribute to its development. Immune cells orchestrate synovial inflammation through cytokines like TNF-α and IL-6. Autoantibodies, including rheumatoid factor (RF) and anti-citrullinated protein antibodies, are diagnostic markers. Disease-modifying antirheumatic drugs (DMARDs), targeting diverse pathways, are pivotal in managing RA. Despite advancements, treatment challenges persist. Emerging therapies, such as JAK inhibitors, offer new avenues.

2.3.2 Multiple sclerosis

(MS) is a prevalent neurological condition affecting young adults. Its rising global incidence is linked to genetic and environmental factors like vitamin D exposure, Epstein–Barr virus infection, obesity, and smoking[41]. While historically seen as a T-cell autoimmune disorder, effective B-cell targeted therapies challenge this understanding. MS is traditionally categorized into relapsing–remitting and progressive stages. Advanced biological treatments, aiming for no evident disease activity (NEDA), are improving long-term outcomes. Promising trials of disease-modifying therapies offer hope for slowing progression and preserving function in progressive MS. These developments challenge the conventional two-stage concept of MS progression.

2.3.3 Psoriasis

(PsO) is a prevalent, persistent skin disorder with global occurrence that affects people of all ages and imposes a significant societal and individual burden[42]. It is linked to significant health conditions like depression, psoriatic arthritis, and cardiometabolic syndrome. The most frequent form,

psoriasis vulgaris or chronic plaque psoriasis, arises from genetic susceptibility, especially the HLA-C*06:02 risk allele, and environmental triggers like streptococcal infection, stress, smoking, obesity, and alcohol use. Diverse phenotypes exist, with research distinguishing pustular from chronic plaque variants. Studies on immunology and genetics pinpoint IL-17, IL-23, and TNF-α as crucial drivers. Biologic therapies targeting these factors have transformed care for severe chronic plaque cases. While no cure exists, early treatment, managing associated health issues, lifestyle adjustments, and personalized care are vital for minimizing physical and psychological harm.

2.3.4 Inflammatory bowel disease

(IBD) is a chronic and potentially life-threatening inflammatory condition affecting the gastrointestinal tract, marked by episodes of intestinal inflammation[43]. The development of IBD involves intricate processes. Recent research has significantly enhanced our understanding of IBD's underlying mechanisms, leading to notable progress in both its diagnosis and treatment. This comprehensive review examines the pathogenesis of IBD, emphasizing recent breakthroughs in host genetic influences, gut microbiota, environmental factors, and particularly, aberrant innate and adaptive immune reactions and their interplay. These discoveries offer potential insights into discovering new predictive or prognostic biomarkers and innovative therapeutic approaches.

2.3.5 Systemic sclerosis

(SSc) is a multifaceted autoimmune connective tissue disorder marked by gradual and chronic tissue and organ fibrosis, varying between individuals[44]. Recognized risk factors encompass persistent Raynaud's phenomenon, hormonal imbalances, certain chemicals, thermal or other injuries. Genetically predisposed individuals are influenced by endogenous and/or exogenous environmental triggers, inducing epigenetic changes. Disease progression begins with microvascular alterations and endothelial dysfunction, leading to myofibroblast transformation. A complex autoimmune reaction, involving both innate and adaptive immunity with specific autoantibodies, characterizes SSc. Irreversible damage to skin and internal organs arises from progressing fibrosis and ischemia. Progenitor cells, growth factors, and cytokines contribute to disease spread and evolution. Emerging therapies target epige-

netic, vascular, and immunological factors underlying systemic fibrosis.

2.3.6 Systemic lupus erythematosus

(SLE) is a chronic, potentially life-threatening autoimmune disorder that affects various organ systems, predominantly striking women between puberty and menopause. Over almost a century, it was realized that SLE, originally considered a skin-focused ailment, is, in fact, a systemic affliction involving multiple organs, driven by an aberrant autoimmune response[45]. The disease presents a wide spectrum of clinical manifestations due to defects across the immune cascade, leading to heterogeneity in its presentation. Delayed diagnosis exacerbates organ damage. SLE's development is influenced by a combination of genetic and environmental factors. While genetic heritability plays a role, it is not the sole determinant of SLE's complex phenotype. Some genetic associations, like C1Q and C4 gene defects, are strong links, while others, such as interferon regulatory factor 5 (IRF5) mutations, contribute to risk. Environmental triggers include UV light exposure, Epstein-Barr virus infection, retroviral sequences, and certain drugs. Hormonal factors, especially estrogen, are implicated due to the higher prevalence in females. The intricate nature of SLE is evident in its diverse clinical and laboratory features, including varied organ involvement, hematological changes, and autoantibody elevation. Subsets like cutaneous and drug-induced lupus further complicate diagnosis. Comorbidities, such as antiphospholipid syndrome and cardiovascular disease, further contribute to the disease's complexity, damage, and mortality risk. As understanding grows, the hope is that new therapeutic strategies will emerge to better manage SLE's multifaceted nature[46].

In conclusion, the realm of autoimmune diseases reveals a broad landscape encompassing over a hundred distinct conditions. The challenge of classification persists, particularly when comprehensive insights into their underlying causative mechanisms remain elusive. These diseases, rather than isolated entities, can be visualized along a continuum, distinguishing between 'organ-specific' afflictions that target specific body organs, and 'systemic' ailments that impact multiple tissues across the body. This chronicity arises from the sustained presence of autoantigens, with some exceptions. The complexity of classification is underscored by initiatives like the PreciseSADS cohort led by Barturen and collaborators, which employs molecular clustering to redefine systemic autoimmune diseases. As we delve into succinct explanations

of prevalent autoimmune disorders in the ensuing section, the intricacies of each condition will further illuminate the intricate tapestry of autoimmunity.

3 Primary Sjögren's Disease

Sjögren's disease (previously named Sjögren's syndrome, before 2023) is named after Henrik Sjögren, a Swedish ophthalmologist. He first described the condition in a doctoral thesis in 1933. While significant strides have been made in understanding its pathophysiology, including the role of immune cells attacking the body's own tissues, there remain gaps in pinpointing its precise triggers and progression. Current treatment options focus on symptom management, often involving artificial tears, saliva stimulants, and immunosuppressive medications in more severe cases. The field is increasingly exploring innovative approaches, such as biologic therapies targeting specific immune pathways, with the hope of not only alleviating symptoms but also addressing the underlying mechanisms to improve patients' overall health and quality of life.

3.1 Overview of a complex systemic autoimmune disease

3.1.1 Description and diagnosis

Primary Sjögren Disease (pSD) is a chronic, disabling inflammatory autoimmune disease characterized by lymphoid infiltration of exocrine glands leading to dryness of the mucosal surfaces, such as the mouth and eyes and by the production of specific auto-antibodies [47–49]. Long-term complications include ocular and dental diseases, systemic involvement, organ damages and increased risk of lymphoma with excess mortality [50, 51]. This pathology is affecting between 0.05% and 0.4% of the adult population [52–55] and is the second most common systemic autoimmune disease [56]. It affects women more often than men (9:1) and the peak frequency of the disease is around fifty years of age [57]. Secondary Sjögren Disease is diagnosed when it is associated with specific organ-specific autoimmune diseases such as thyroiditis, primary biliary cholangitis (PBC), cholangitis, or other systemic autoimmune diseases like RA, SLE, scleroderma, or even dermatomyositis.

The classification criteria for pSD, proposed in 2016 by the American College of Rheumatology and the EULAR, are based on the combination of objective criteria for ocular and/or oral dryness, histological signs assessing glandular lymphocytic infiltration on a salivary gland biopsy, and the presence of anti-SSA autoantibodies (refer to **Figure 7**). However, it important

to note that the diagnosis can pose considerable difficulties as distinct clinical symptoms are often absent during the initial disease phases, compounded by the absence of noninvasive diagnostic techniques boasting both high specificity and sensitivity. This can result in noteworthy treatment delays and exacerbate overall clinical results.

Among autoimmune diseases, pSD presents a particularly interesting study model. This disease resides at the intersection of organ-specific autoimmune diseases and systemic diseases.

The terms ESSDAI and ESSPRI refer to two commonly used assessment tools in the context of Sjögren's disease. ESSDAI: The EULAR Sjögren's Syndrome Disease Activity Index (ESSDAI), takes into account various systemic manifestations of the disease, such as glandular, articular, cutaneous, hematological, and other organ involvement. Each of the 12 domains (or manifestation) is scored based on its severity and impact on the patient's well-being. The total ESSDAI score provides an overall assessment of disease activity. ESSPRI: The EULAR Sjögren's Syndrome Patient Reported Index (ESSPRI) is used to assess the patient's subjective experience of the disease and its impact on their quality of life. It focuses on the severity of dryness-related symptoms, including ocular and oral symptoms, as well as fatigue and pain. The ESSPRI score is based on patient self-assessment through questionnaires and provides insight into the patient's perspective on their symptoms and their effect on daily life. Both ESSDAI and ESSPRI are valuable tools - and currently used by physicians - for evaluating different aspects of Sjögren's disease, helping to monitor disease activity and its impact on patients' lives.

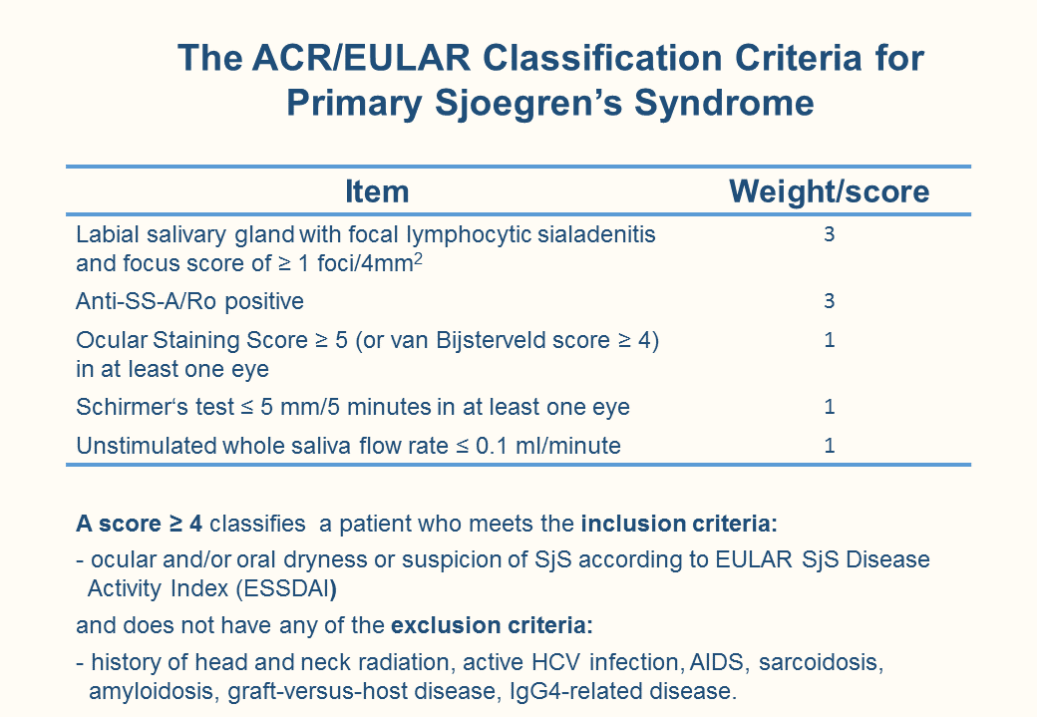


Figure 7: Classification criteria for pSD proposed in 2016 by the American College of Rheumatology and the EULAR. The classification criteria for pSD are applicable to individuals with a score ≥ 4 by adding the following criteria, in the absence of exclusion criteria (history of cervical irradiation, HCV infection, HIV infection, sarcoidosis, amyloidosis, graft-versus-host disease, IgG4-related disease).

3.1.2 Etiology of pSD

The precise origins of Sjögren's disease remain unidentified. Nevertheless, a correlation to genetic factors appears evident. The severity of the autoimmune disease may be influenced by HLA genes in humans, with suggestions that individuals possessing DQ1/DQ2 alleles might experience more severe autoimmune manifestations compared to those with different allelic combinations at the HLA locus[48]. Furthermore, environmental elements are implicated, for instance, infections of exocrine glands, like salivary or lacrimal glands, could potentially result in damage to salivary glands and expose their

cellular components such as DNA, RNA, and histones to circulating immune cells. This is particularly attributed to the fact that salivary gland tissues serve as reservoirs for latent viral infections. The array of viruses implicated in Sjögren's disease is extensive and encompasses viruses from the Herpesviridae family[58], such as Epstein–Barr virus (EBV) and human herpesvirus 6 (HHV6).

3.1.3 Pathophysiology of pSD

The pathophysiology of Sjögren's syndrome revolves around a dysregulated immune response that results in chronic inflammation and tissue damage, predominantly targeting the exocrine glands.

Autoantibodies production This dysfunction process leads to the activation of the innate and adaptive immune systems with the secretion of autoantibodies. Autoantibodies serve as vital biological markers for autoimmune diseases, often emerging up to two decades before diagnosing Sjögren's Syndrome. Antinuclear antibodies (ANAs), targeting cell nucleus and cytoplasm, are prevalent (over 80%) in Sjögren's patients, aiding identification, particularly in primary care. Rheumatoid factor(RF) antibodies (specific to IgG's Fc fragment) are found in half of Sjögren's patients, linked to key disease features. While ANAs and RF detection assists in diagnosing autoimmune diseases, they lack specificity for Sjögren's in current classification criteria due to their presence in various autoimmune conditions.

Anti-Ro/SSA antibodies, ANAs targeting Ro52 and Ro60 proteins associated with RNA molecules, are detected in around 70% of Sjögren's patients. Recent studies suggest different clinical associations for anti-Ro52 and anti-Ro60 autoantibodies, warranting separate detection for suspected Sjögren's diagnosis. Anti-Ro52 antibodies closely correlate with primary clinical, immune, and histopathological Sjögren's features.

Anti-La/SSB antibodies target the La/SSB protein involved in RNA metabolism, often co-existing with anti-Ro/SSA antibodies in Sjögren's patients. Detecting both anti-La/SSB and anti-Ro/SSA autoantibodies links to higher ANA positivity and systemic Sjögren's activity. Isolated anti-La/SSB antibodies without anti-Ro/SSA antibodies occur in only 2.3% to 7% of Sjögren's cases. Patients with anti-Ro/SSA antibodies tend to display more main clinical and immune Sjögren's features than those with only anti-La/SSB antibodies. However, patients with anti-La/SSB antibodies often

exhibit certain Sjögren's clinical traits, such as dry mouth and ANA-specific antibodies, at a higher frequency than those negative for both anti-Ro/SSA and anti-La/SSB antibodies [48].

The interferon pathway The IFN signature also plays a crucial role in the underlying mechanisms of pSD. Various transcriptomic studies of accessory salivary glands and PBMCs have revealed elevated expression of IFN-induced genes in pSD patients compared to controls [59–61]. Plasmacytoid dendritic cells (pDCs) are pivotal in producing type-I interferons (IFNs) [62]. Hillen et al. discovered multiple patterns within pDCs of primary Sjögren's patients, signifying their activation and heightened IFN-related gene activity [63]. Exploring IFN signatures continues to capture the attention of research groups globally, and understandably so. The emergence of therapies targeting IFNs holds the potential to profoundly alter the prognosis of individuals with IFN-related conditions [64].

For instance, Emamian et al., in peripheral blood analysis of 21 pSD patients and 23 controls, and independently in 17 patients and 22 controls, the authors demonstrated that the expression levels of most IFN-induced genes were positively correlated with anti-SSA and anti-SSB antibodies, suggesting a link between innate immunity and B lymphocyte activation [65]. Another study by Trutschel et al. in a French cohort, evaluated the expression of certain genes induced by type I interferon: IFI27, IFI44, and OAS3. The authors demonstrated that the presence of a high type I IFN signature was associated with an earlier onset of the disease and a higher ESSDAI compared to patients with a weaker IFN signature [60].

In parallel, an interesting study driven by Brkic et al. [66] aimed to assess the prevalence of elevated expression of interferon (IFN) type I inducible genes within CD14 monocytes among 69 primary Sjögren's patients and 44 healthy controls (HC). The research sought to establish a correlation between this signature, disease manifestations, and the expression of B cell activating factor (BAFF). The findings revealed that an IFN type I signature was detected in 55% of pSS patients, whereas only 4.5% of HC exhibited this signature. Patients with the IFN type I signature displayed several characteristics: (a) elevated EULAR Sjögren Disease Activity Index scores; higher levels of anti-Ro52, anti-Ro60, and anti-La autoantibodies; increased rheumatoid factor levels; higher serum IgG levels; decreased C3 levels; lower absolute

lymphocyte and neutrophil counts; (b) heightened BAFF gene expression within monocytes. Moreover, the serum from patients with a positive signature induced increased BAFF gene expression within monocytes.

Roles of B and T lymphocytes Over the past decade, numerous research studies have been dedicated to unraveling the role of the B-cell activating factor (BAFF) cytokine in primary Sjögren's disease (pSD). Initially, elevated levels of BAFF in the serum were observed, revealing a correlation with autoantibody levels (Anti-SSA, Anti-SSB, and RF) in pSD patients [67]. Moreover, heightened levels of BAFF were detected within the salivary glands of individuals with pSD. Importantly, it was found that BAFF secretion did not only originate from conventional BAFF-producing cell types such as monocytes, macrophages, and dendritic cells (DCs), but also from T cells and B cells.

An excessive activation of B cells is described throughout the disease progression. Various clinical and biological aspects of the condition mirror this activation, including the presence of autoantibodies, polygammopathy, elevated serum free light chain levels, and a heightened risk of lymphoma in patients.

In addition to B lymphocytes, T cells play a significant role in the pathogenesis. As previously noted, genetic investigations have proposed the involvement of Th1 cells in pSD. However, evidence has also hinted at the polarization of Th1 cells. Notably, a substantial elevation in Th1-related cytokines has been observed in both mouse models and human studies. More recently, Th17 cells have garnered attention due to their emerging relevance in autoimmune processes [68].

In the review article titled "*One year in review 2020: pathogenesis of primary Sjögren's syndrome*" by M. Bombardieri et al., it is mentioned that extensive exploration has been conducted into the roles of B and T cell subsets, with a notable emphasis on Tfh cells and their counterpart T follicular regulatory cells. Recent findings suggest an increased Tfr to Tfh cell ratio in patients with Sjögren's syndrome compared to healthy controls, indicating a potential imbalance between pro-inflammatory and immunoregulatory pathways in the disease.

Ultimately, the proposed pathophysiological hypothesis proposes that an initial activation of the innate immune system, triggered by environmental

factors, leads to the production of interferon (IFN) by pDCs (type I IFN), CD8 T cells, and Natural Killer cells. Factors stimulating innate immunity (such as viruses or Toll-like receptor activators) are also capable of activating epithelial cells. Subsequently, type I IFN induces the release of the BAFF cytokine, which plays a pivotal role in the activation of B lymphocytes. Consequently, the BAFF cytokine emerges as a critical juncture where innate and adaptive immune system activation intersect (refer to **Figure 8**)."

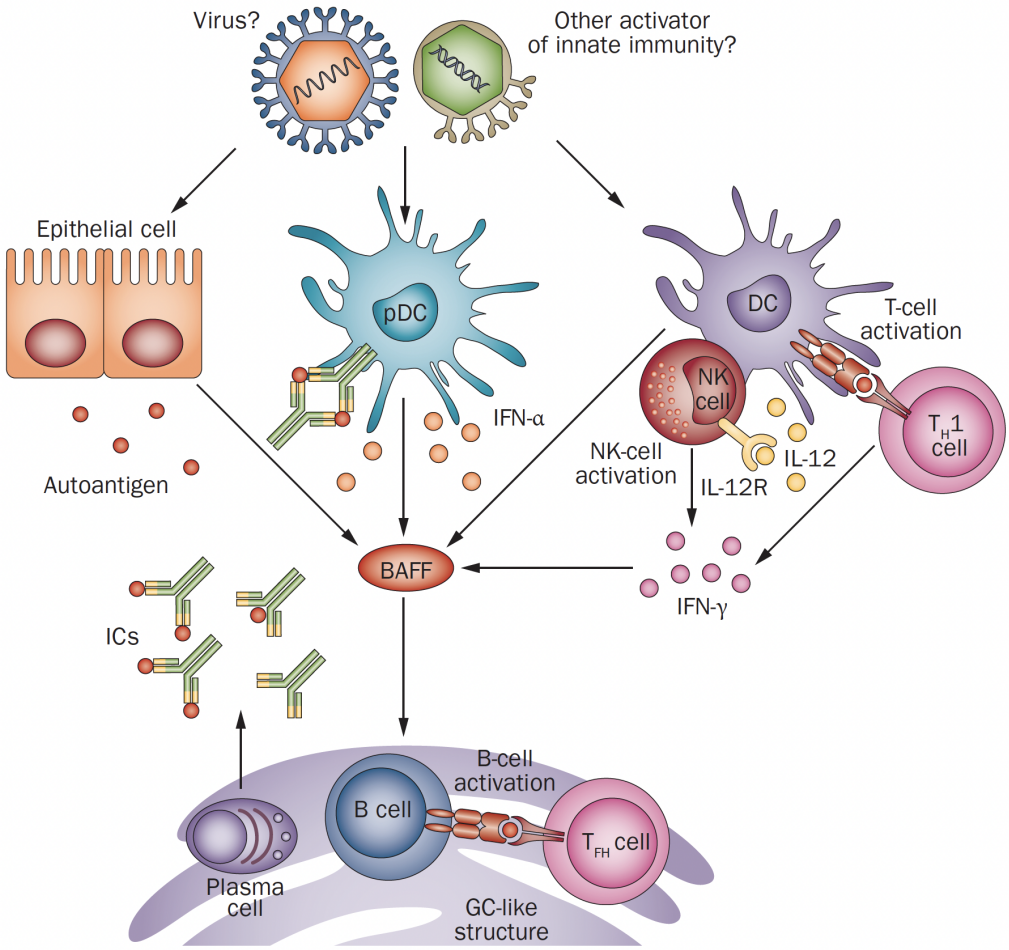


Figure 8: Schematic representation of the pathophysiological hypothesis of pSD (adapted from G. Nocture and X. Mariette [67]

3.2 Molecular taxonomies of Sjögren's disease

Sjögren's disease displays heterogeneity due to its capacity to exhibit a wide spectrum of symptoms, varying degrees of severity, and the potential to impact different organ systems in diverse manners Additionally, Sjögren's disease can exert its effects on various organ systems beyond the exocrine glands, encompassing the salivary and tear glands. Some individuals may primarily encounter symptoms related to these glands, while others may develop broader systemic manifestations. Certain individuals may experience mild symptoms that remain stable over time, whereas others might confront swiftly advancing and severe iterations of the disease.

The intricate array of clinical manifestations, coupled with involvement across multiple organ systems, renders the precise categorization or prediction of the disease course in any given individual a difficult challenge. This complexity adds intricacy to the diagnostic process. Moreover, within the context of clinical trials, responses to therapies can be divergent; some patients may favor therapies targeting specific symptoms, while others may necessitate more objective approaches. Consequently, the pressing need to better stratify patients becomes evident.

Recent research endeavors have been undertaken, including explorations into symptoms-based classifications and molecular classifications.

3.2.1 Symptom-based classification

In a study conducted by Tarn and colleagues [69], a comprehensive investigation was undertaken utilizing the UK Primary Sjögren's Syndrome Registry (UKPSSR). This registry encompasses a national observational cohort of extensively characterized patients (n=608) diagnosed with primary Sjögren's disease based on the 2002 American European Consensus Group (AECG) classification criteria. The research team carried out a robust unsupervised hierarchical cluster analysis, considering patient-reported symptoms such as pain, fatigue, dryness, anxiety, and depression. As a result of this analysis, the study discerned four distinctive patient clusters. These clusters were categorized as follows: low symptom burden (LSB), high symptom burden (HSB), dominant dryness with fatigue (DDF), and dominant pain with fatigue (PDF).

The LSB and DDF subgroups shared numerous objectively measured laboratory characteristics, including diminished lymphocyte counts and height-

ened IgG concentrations. Additionally, they were more likely to exhibit positivity for anti-SSA and anti-SSB antibodies compared to the HSB and PDF subgroups. As anticipated, the DDF subgroup exhibited the most compromised objectively measured glandular function. Notably, variations were also detected in the transcriptomic modular profiles between the LSB and DDF subgroups.

Furthermore, within the UKPSSR cohort, the prevalence of lymphoma was highest in the DDF subgroup. This aligns with the elevated serum concentrations of CXCL13 (linked to lymphoma in primary Sjögren's patients) and heightened levels of β2-microglobulin (a prognostic marker for unfavorable outcomes in malignant lymphoma) in the DDF subgroup. Moreover, altered expression of genes associated with B-cell signaling, germinal centers, lymphoproliferative disease, and oxidative stress was observed in this subgroup. Contrasting this, the LSB subgroup exhibited the highest levels of IFN and T cell transcriptomic modular activities.

3.2.2 Molecular classification

In the beginning of my PhD journey, I had the privilege of engaging in an ongoing project. The project aimed to better understand the heterogeneity of Sjögren's patients within the PreciseSADS project [38]. Within this study, aimed at identifying a molecular classification of Sjögren disease to stratify patients based on molecular characteristics, possibly overcoming the clinical heterogeneity of the disease, we introduce a molecular classification framework tailored for individuals affected by Sjögren, using data from a cross-sectional cohort (c.f. **Annex 1**). This framework is established through a comprehensive multi-omic profiling of whole blood samples, employing a European cohort encompassing over 300 patients, along with a corresponding number of age and gender-matched healthy volunteers. By analyzing information spanning transcriptomics, genomics, epigenetics, cytokine expression, flow cytometry, and clinical parameters, we discern distinct patterns of immune dysregulation among four patient groups. These groupings shed light on diverse immune dysregulation profiles, thereby enriching our holistic comprehension of Sjögren's syndrome.

A distinct investigation led by Trutschel et al. [60] delves into a cohort sourced from the French multicenter 5-year prospective Assessment of Systemic Signs and Evolution of Sjögren's Syndrome (ASSESS) study. This cohort encompassed 395 patients who all met the criteria set by the Ameri-

Authors	Cohort	Patients samples	Material	Main features
Tarn et al	UKPSSR	608	Symptom-based	- Low Symptom Burden (LSB) - High Symptom Burden (HSB), - Dominant Dryness with Fatigue (DDF) - Dominant Pain with Fatigue (PDF)
Soret et al	PreciseSADS	300	Whole blood transcriptome	- Interferon - Neutrophils / Inflammation - Lymphoid lineage
Trutschel et al	ASSESS	395	Whole blood transcriptome	- Interferon - Erythrocytes

Table 1: Summary table of stratification models proposed in the literature

can–European Consensus Group for primary Sjögren's disease (pSD). Within this investigation, the researchers also unveiled four distinctive patient clusters. These clusters were differentiated based on the characteristics of two specific modules: IFN-stimulated genes (ISGs) and the erythroid module (ERM). Among these clusters, cluster T1 displayed heightened expression levels in both of these modules. Conversely, cluster T2 exhibited a contrastingly low expression of ISGs alongside elevated ERM expression. Cluster T3 stood out for its elevated ISG expression combined with lower ERM expression, while cluster T4 was characterized by reduced expression levels in both ISGs and ERM.

These studies (summarized in **Table 1**) provided a clear cut proof that patients exhibit dissimilar behaviors at both the molecular and symptomatic levels. Despite the emergence of somewhat similar outcomes, a key concern within these classifications is the requirement for a consensus classification. Furthermore, a significant challenge lies in comprehending the linkages between molecular and clinical data—such as disease severity and response to treatment. Addressing these challenges could potentially be facilitated by employing more refined or tailored technologies, such as single-cell technologies.

3.3 Treatment options for Sjogren's Disease

As of now, neither the Food and Drug Administration (FDA) nor the European Medicines Agency (EMA) has approved any disease-modifying treatment for Sjögren's syndrome. Available treatments aim to manage symptoms, reduce inflammation, and improve patients' overall quality of life. Since Sjögren's is a chronic condition without a known cure, the focus is on controlling the symptoms and preventing complications such as artificial tears and moisturizing agents, salivary stimulants, anti-inflammatory medications,

corticosteroids may and immunosuppressive drugs.

3.3.1 Hydroxychloroquine

Hydroxychloroquine (HCQ) is an immunomodulatory and disease-modifying antirheumatic drug (DMARD). Its effectiveness involves several mechanisms of action: inhibition of chemotaxis and phagocytosis of polymorphonuclear cells, macrophages, and monocytes; inhibition of T lymphocyte proliferation and the production of numerous cytokines such as IL1, IL2, IL6, IL17, IL22, IFN-α, IFN-γ, and finally, inhibition of certain receptors of innate immunity, particularly TLR7 and TLR9 [70, 71]. Hydroxychloroquine indeed inhibits the acidification of late endosomal compartments, preventing the interaction between TLR7, TLR9, and their respective ligands, single-stranded RNA, and double-stranded DNA. Administration of HCQ has been tested in Sjögren's disease. However, a randomized controlled trial (JOQUER study) conducted on pSD patients treated with hydroxychloroquine (400mg/day) for 24 weeks did not show improvement in ESSDAI and ESSPRI compared to the placebo arm[72]. Following this study, a recently published work demonstrated that HCQ treatment led to a reduction in the IFN score in the blood of treated patients compared to placebo. The IFN score was defined based on the relative expression of 5 genes: IFI44, IFI44L, IFIT1, IFIT3, and MXA. Furthermore, this study showed that HCQ decreased IgG and IgM levels in patients [73].

3.3.2 Leflunomide

Leflunomide (LEF) is also a DMARD. It is primarily used in the treatment of autoimmune diseases, particularly rheumatoid arthritis. LEF helps to reduce inflammation and slow down the progression of the disease by targeting the underlying immune system dysfunction. LEF works by inhibiting an enzyme called dihydroorotate dehydrogenase, which is involved in the production of pyrimidines, essential components of DNA and RNA. By interfering with this enzyme, LEF hampers the proliferation of immune cells, such as T cells. LEF also exerts a direct influence on B cells, resulting in a decrease in their growth and the synthesis of antibodies [74]. LEF has been investigated in context of pSD, especially with patients showing joint involvement, preliminary open trial conducted in individuals with pSD, LEF (20mg/day) demonstrated only a trend advantage for 15 patients who had been recently diagnosed with pSD

[75].

3.3.3 Repurpss-1 clinical trial, a leflunomide and hydroxychloroquine combination therapy

Drawing upon the synergistic effects of leflunomide and hydroxychloroquine in quelling the activation of pivotal immune cells in primary Sjögren's syndrome, van der Heijden and colleagues (2020) [**Heijden**] proposed a randomized phase 2A clinical trial, distinguished by its placebo-controlled, double-blind framework. This trial took place at the University Medical Center Utrecht in the Netherlands. The study enrolled individuals aged 18 to 75, presenting a European League Against Rheumatism (EULAR) Sjögren's syndrome disease activity index (ESSDAI) score of 5 or higher, as well as a lymphocytic focus score of 1 or higher in biopsied labial salivary gland samples. Patients were assigned randomly (2 1) using block randomization (with a block size of six) to either receive leflunomide(20 mg) and hydroxychloroquine(400 mg daily) - or a placebo -for a duration of 24 weeks. The primary objective was to ascertain the intergroup variation in ESSDAI score changes from week 0 to week 24. Notably, the combined application of HCQ and LFU yielded a reduction in ESSDAI scores at the 24 week mark in comparison to the placebo group: -4.35 points (95% CI -7.45 to –.25, p=0.0078). These promising outcomes provide a compelling rationale for the initiation of further trials involving a more extensive cohort of patients.

3.3.4 Rituximab

Rituximab is a chimeric monoclonal antibody against CD20. CD20 is present on all mature B cells but is not expressed on pre-B lymphocytes or plasma cells. The binding of rituximab to CD20 leads to the lysis of B cells through mechanisms such as antibody-dependent cellular cytotoxicity (ADCC) involving the Fc fragment of immunoglobulins, complement activation, or direct apoptosis. Therapeutic strategies aimed at controlling B cell hyperactivation have proven effective in the treatment of rheumatoid arthritis and ANCA-associated vasculitis [76]. Two large-scale randomized trials were conducted but did not show clinical efficacy of rituximab. Primary outcome measures, such as the reduction of dryness and fatigue reported by patients, were not achieved [77, 78]. This lack of efficacy could be linked to the presence of elevated levels of BAFF in patients [79].

In conclusion, despite the fact that I did not list all the treatment available, Sjögren's syndrome lacks an approved disease-modifying treatment, leaving available options focused on symptom management and enhancing patients' quality of life. Combining HCQ and LEF, a repurposing clinical trial exhibited a reduction in disease activity scores, suggesting a promising avenue for future research. Furthermore, rituximab, a monoclonal antibody targeting CD20 on B cells, demonstrated efficacy in related conditions but fell short in large-scale trials for Sjögren's disease. The presence of elevated BAFF levels might contribute to this lack of efficacy. While current treatments provide symptomatic relief, ongoing research seeks more effective approaches to address the root causes of Sjögren's syndrome and improve patients' long-term outcomes.

3.4 The IMI 2 NECESSITY European consortium

In 2019, the IMI 2 NECESSITY European consortium was established, and it played a pivotal role in inspiring and furnishing the essential resources for the research presented in this manuscript. The primary objective is a multifaceted endeavor aimed at revolutionizing the landscape of research and treatment for primary Sjögren's disease. The consortium's first objective entails an exhaustive re-analysis of data derived from all available primary Sjögren's disease randomized controlled trials. The overarching goal is to discern clinically relevant outcome measures that can effectively distinguish between patients treated with the drug and those administered placebos. This rigorous re-analysis forms the foundation for the development of a groundbreaking composite responder index known as the Sjögren's Syndrome Tool for Assessing Response (STAR). STAR will be seamlessly integrated with established indices like ESSDAI (EULAR Sjögren's Syndrome Disease Activity Index) and ESSPRI (EULAR Sjögren's Syndrome Patient Reported Index), serving as pivotal endpoints for future studies. Concurrently, the consortium aims to pioneer innovative tools for comprehensively assessing the multifaceted nature of the disease. Furthermore, the second objective revolves around identifying and assessing discriminative biomarkers capable of stratifying primary Sjögren's Syndrome patients, ultimately predicting organ involvement and disease progression. To achieve this, the consortium will harness cutting-edge "omics" technologies on biological samples from existing clinical trial cohorts, unraveling the potential of these "smart biomarkers" for disease stratification and prediction. Finally, the third objective involves designing

and executing a clinical trial to validate the newly defined clinical endpoints and biomarkers. This original and innovative "multi-arm multi-stage platform trial" is designed to accommodate diverse patient types, a variety of drugs, and distinct methodologies. The consortium also strives to garner consensus among key stakeholders, including health authorities, payers, and patient advocacy groups, to ensure the integration of these novel approaches into regulatory approval and reimbursement processes. In doing so, Necessity endeavors to bring about transformative advancements in the management of primary Sjögren's Syndrome, engaging with entities like Health Technology Assessment (HTAs), the European Medicines Agency (EMA), payers, and organizations representing patients with primary Sjögren's Syndrome to achieve these goals collaboratively.

4 Precision Medicine

Precision medicine is an approach to healthcare that tailors medical decisions and treatments to individual characteristics, such as genetics, lifestyle, and environment. By personalizing prevention and therapy, precision medicine aims to improve patient outcomes and optimize healthcare interventions [80–84]. In this chapter, I will concentrate on the concepts within the broad scope of precision medicine that are specifically relevant to my project.

4.1 Unleashing the Power of Transcriptomic Data in Precision Medicine

Omics data encompasses extensive datasets derived from state-of-the-art technologies like genomics, transcriptomics, proteomics, and metabolomics. These techniques enable researchers to explore biological molecules and their interactions on a large scale, providing valuable insights into cellular functions and mechanisms. Omics data plays a crucial role in advancing biomedical research, our understanding of various diseases, and personalized medicine. In this subsection, I will be focusing on transcriptomic data, which involves the study of a sample's complete set of RNA transcripts, shedding light on gene expression patterns and co-expression networks. Transcriptomic analyses offer crucial information about how genes are activated or silenced in different conditions, facilitating a deeper comprehension of cellular processes and disease mechanisms.

4.1.1 Exploring transcriptomics

Transcriptomics is a branch of molecular biology that focuses on the study of an sample's entire set of RNA transcripts, collectively known as the transcriptome. These RNA molecules serve as intermediaries between genes and the proteins they encode. By analyzing the transcriptome, researchers gain valuable insights into the gene expression patterns and regulatory mechanisms within cells, tissues, or organisms.

Gene expression analysis is performed using advanced techniques such as microarrays and RNA sequencing (RNA-seq). These methods allow to comprehensively profile the transcriptome and identify changes in gene expression levels under different experimental conditions or disease states. With the help of bioinformatics tools and computational analyses, vast amounts of

transcriptomic data can be processed, integrated, and interpreted to reveal meaningful biological insights.

4.1.2 Importance of transcriptomic data in precision medicine

Transcriptomic data plays a crucial role in precision medicine by providing valuable insights into the gene expression patterns and regulatory mechanisms that underlie various diseases and individual responses to treatments. Here are some key aspects highlighting the importance of transcriptomic data in precision medicine:

1. Identification of distinct molecular subtypes of diseases. This allows for unbiased disease classification, enabling clinicians to tailor treatments based on the specific molecular characteristics of a patient's condition. Subtyping helps ensure that patients receive the most effective therapies, reducing the risk of treatment resistance and improving overall outcomes.

2. Identification of specific gene expression signatures (or biomarkers) associated with diseases. These biomarkers can be used for early detection, predicting disease progression, and monitoring treatment response. Having biomarkers enables targeted and timely interventions, increasing the chances of successful outcomes.

3. A support in the discovery of potential drug targets by pinpointing genes and pathways that are dysregulated in specific diseases. This information allows researchers to develop targeted therapies that focus on the molecular drivers of a disease, leading to more effective and less harmful treatments.

4. A way to monitor changes in gene expression levels in response to treatments. This provides real-time feedback on the effectiveness of a chosen therapy, allowing for treatment adjustments or the switch to alternative treatments if necessary.

5. Clinical Trial Design: Transcriptomic data is increasingly used in designing clinical trials for precision medicine. By stratifying patients based on their gene expression profiles, researchers can create more homogeneous study groups, leading to more informative and successful clinical trials.

Overall, transcriptomic data is a powerful tool in precision medicine, enabling a deeper understanding of the molecular basis of diseases and individual variability in treatment responses. By integrating transcriptomic information with other omics data and clinical data, precision medicine approaches can deliver more targeted and effective treatments, ultimately improving patient outcomes and healthcare efficiency.

4.2 Unraveling the Transcriptome: Analysis and information extraction from transcriptomic data

Finding valuable information in transcriptomic data requires a systematic and thorough approach, involving various data analysis techniques and tools. Here are some key steps to extract meaningful insights from transcriptomic data:

4.2.1 Preprocessing and Quality Control

Before diving into analysis, it is crucial and mandatory to preprocess the raw transcriptomic data (FPKM, TPM or counts) to remove noise and artifacts. This step may involve background correction, normalization, and filtering out low-quality or unreliable data points. Throughout my PhD, I consistently had access to pre-aligned data. As a result, the subsequent steps outline the main approach I followed.

- Perform variance stabilizing (or log tansformation) of the raw count matrix to account for differences in sequencing depth, alleviate heteroskedasticity and other technical biases. Popular methods include TPM (Transcripts Per Million), RPKM (Reads Per Kilobase Million), or DESeq normalization [85].
- Examine the distribution of gene expression values before and after normalization.
- Check for batch effects if the data comes from multiple experiments or sequencing runs.
- Filtering Low-Expression Genes, most of the genes, are not involved in the biological process studied and will be expressed at a similar level by all the cells. Such genes are not informative and can therefore be

removed before performing any analysis. To do so the variance and the mean of each gene is computed, and a local polynomial regression model is fitted to account for the dependency of the variance toward the mean, a characteristic of the negative binomial distribution. Residuals of the regression are computed and are considered as the 'corrected variance'. The genes with the highest corrected variance are usually selected and kept for downstream analysis.

- If batch effects are identified during quality control, a batch correction would be necessari, many methods are available in the litterature, for instance, ComBat from limma R package [86] .

- Outlier Detection and Sample Clustering: Perform sample clustering or Principal Component Analysis (PCA) to identify potential outliers or groupings of samples.

Rigorous preprocessing and quality control of RNAseq data lay the foundation for reliable and accurate analyses.

4.2.2 Differential Gene Expression Analysis

One of the primary goals of transcriptomic data analysis is to identify genes that are differentially expressed between different experimental conditions or disease states. This analysis can be performed using statistical tests to determine which genes show significant changes in expression levels. A myriad of methods and packages are accessible in diverse programming languages for conducting this analysis. Below are some of the R packages that I relied on during my PhD:

- DESeq [85] is an R package widely used. It uses a negative binomial distribution model to account for the count-based nature of RNA-seq data and provides robust statistical methods to identify differentially expressed genes.

- Limma [86], Although primarily used for microarray data, this R package can also be applied to RNA-seq data. It utilizes linear models and empirical Bayes methods to identify differentially expressed genes.

It is essential to consider that the selection of a method and package might be influenced by various factors related to the data, such as the number of

replicates, distributional assumptions, and the computational resources at hand.

4.2.3 Functional Enrichment Analysis

After identifying differentially expressed genes, functional enrichment analysis emerges as a crucial tool for unraveling the intricate relationships between these genes and their involvement in biological processes, molecular functions, or cellular components. This analysis provides invaluable insights into the underlying biological pathways and mechanisms impacted in a specific condition. Two major components are essential for conducting this analysis. Firstly, the selection of an appropriate method is critical. Among the plethora of available options, a two-tailed Fisher-exact test can be applied [87] against different sources of gene modules or pathways and Gene Set Enrichment Analysis (GSEA) [88]. Secondly, the choice of the gene signatures database plays a vital role. For a comprehensive understanding, utilizing multiple and complementary databases is recommended. Broad databases like Gene Ontology [89], KEGG [90] or MsigDB [91] can be employed, while for specific fields, using more specialized databases is advisable. Despite the availability of numerous methods, a clear understanding of their appropriate usage remains essential. Addressing the inherent limitations in these approaches is crucial to ensure the robustness and reliability of the findings [92].

4.2.4 Clustering and Classification

Applying clustering algorithms to transcriptomic data can help identify distinct subgroups of samples with similar gene expression patterns. This can aid in disease subtyping and patient stratification for precision medicine approaches. Classification algorithms can also be employed to predict disease outcomes or treatment responses based on gene expression profiles. This subject will be further discussed in the subsequent section (c.f. **Section 4.3**).

4.2.5 Gene co-expression networks

Transcriptomic data can be used to construct gene co-expression networks, where genes with similar expression patterns are grouped together. Analyzing these networks can reveal modules of functionally related genes and

potential key regulators of biological processes. This analysis constitutes a central component of my project and will be further elucidated in the subsequent section (c.f. **Section 4.3.3**).

4.2.6 Multi-omic, integration with other data types

Integrating transcriptomic data with other omics data, such as genomics or proteomics, can provide a more comprehensive understanding of biological processes. It can help identify potential regulatory mechanisms and interactions between different molecular layers [93, 94].

4.2.7 Drug repurposing

Drug repurposing, or the repositioning of existing drugs for new therapeutic indications, has emerged as a promising strategy in pharmaceutical research and development. This approach capitalizes on the extensive knowledge and safety profiles of drugs that have already received regulatory approval for one purpose, expediting the drug discovery process and reducing associated costs. Notable examples of successful drug repurposing include the use of the anti-malarial drug chloroquine for the treatment of autoimmune diseases like RA and SLE [71]. A prominent instance of drug repurposing involves the utilization of sildenafil (commonly known as Viagra) for the treatment of erectile dysfunction. Originally, Pfizer developed sildenafil in the 1980s as a treatment of coronary artery disease [95]. Additionally, thalidomide, infamous for its teratogenic effects, has been repurposed to effectively treat multiple myeloma [96]. These examples illustrate the potential of drug repurposing to uncover new therapeutic avenues, enhance patient care, and optimize the utilization of existing pharmaceutical agents, offering an efficient and innovative approach to drug discovery and development. During the challenging period of the COVID-19 crisis, it presented a unique blend of stress and scientific curiosity. Throughout this time, I took the initiative to develop and implement a transcriptomic analysis pipeline for COVID-19 data. The primary objective was to identify genes that were differentially expressed in COVID-infected cell lines compared to mock samples. The outcomes of this analysis subsequently contributed to the enhancement of an internal molecular repositioning tool within Servier (Patrimony). This improvement allowed us to propose hypotheses for potential drug candidates aimed at mitigating severe lung inflammation in COVID-19. For more de-

tailed information, please refer to the article provided in **Annex 2**.

4.2.8 Visualization Techniques

Effective visualization of transcriptomic data is crucial not only for gaining profound insights but also for effectively communicating research findings to both scientific peers and wider audiences. In the realm of transcriptomic data analysis, several visualization methods prove indispensable. Among these, heatmaps stand as a powerful tool, enabling the comprehensive depiction of gene expression patterns across samples and conditions, facilitating the identification of clusters and trends within the data. Additionally, volcano plots offer a dynamic way to visualize differentially expressed genes, highlighting the statistical significance and magnitude of gene expression changes. Furthermore, pathway diagrams provide an intuitive representation of how genes are interconnected within biological pathways, shedding light on potential molecular mechanisms underlying observed changes in gene expression. These visualization techniques collectively empower researchers to unravel complex patterns and narratives hidden within transcriptomic data, fostering a deeper understanding of the biological systems under investigation.

4.2.9 Validation and Reproducibility

Ensuring the reliability of findings from transcriptomic data analysis is of utmost importance due to the high-dimensional nature of the data, which inherently carries a risk of generating false positive findings. To mitigate this risk, it becomes essential to validate and replicate the findings using independent datasets or experimental validations.

Overall, by following a rigorous and comprehensive analytical pipeline, researchers can identify biomarkers, potential therapeutic targets, and pathways associated with diseases, thereby unlocking valuable information hidden within transcriptomic data. These insights can significantly impact precision medicine, leading to more personalized and effective healthcare strategies.

4.2.10 Computational pipeline for transcriptomic analysis

4.3 Identification of patterns in transcriptomic data through clustering

Clustering is a fundamental technique in the field of unsupervised learning, designed to organize similar items into groups. The main goal is to group together objects that share similarities, while at the same time, ensuring that dissimilar objects are placed in distinct clusters. In this section, we will go through various clustering methods, exploring their strengths and weaknesses. Additionally, we will delve into two compelling applications that harness the power of clustering in gene expression data analysis. Our first application will focus on the identification of patient subgroups, clinically known as disease stratification, allowing for more personalized and effective medical treatments. Our second application will delve into the realm of gene module identification, unraveling complex patterns of gene interactions.

4.3.1 Overview of unsupervised machine learning algorithms

There are several clustering methods employed in data analysis, each offering distinct approaches to grouping similar data points.

K-means is a popular partition-based clustering algorithm that iteratively assigns data points to the nearest cluster centroid based on distance measures, aiming to minimize the within-cluster sum of squares. Its strengths include simplicity, efficiency on large datasets, and effectiveness in producing compact, well-separated clusters. However, K-means requires the number of clusters to be specified beforehand and may converge to local optima, leading to suboptimal results.

On the other hand, **hierarchical clustering** forms a tree-like structure of nested clusters by iteratively merging or splitting clusters based on distance measures. Its main strength lies in providing a visual representation of hierarchical relationships in data, allowing users to explore different levels of granularity. Yet, hierarchical clustering can be computationally expensive for large datasets and lacks the flexibility to adjust cluster assignments once merging or splitting occurs.

Spectral clustering, a dimension reduction based clustering method, leveraging the eigenvectors of similarity matrices, excels in capturing complex relationships among data points and performing well on non-convex clusters.

Nevertheless, spectral clustering might require careful tuning of parameters and can be sensitive to noise and outliers.

Louvain Clustering, a community detection algorithm used to identify cohesive groups or communities within a network or graph. The algorithm operates in two phases: the first phase optimizes the modularity metric to detect small communities, and the second phase aggregates the communities found in the first phase to form larger communities. The algorithm starts with each node belonging to its own community. In the first phase, it iteratively moves each node to its neighbor's community or its own community, aiming to maximize the modularity of the graph. Modularity measures the strength of the division of the graph into communities by maximizing the number of edges between vertices within the same community and minimizing edges between vertices in distinct communities, and the goal is to find a partition that maximizes this metric. Once the first phase is complete and no further improvement in modularity can be achieved, the second phase begins. In this phase, the communities detected in the first phase are considered as new nodes, and a new graph is constructed where each node represents a community. The edges between the new nodes are weighted based on the sum of the weights of the edges between the nodes in the corresponding communities. The first and second phases are repeated until no further increase in modularity is possible.

It is crucial to retain focus on two significant points from the clustering analysis.

- Choosing the appropriate clustering method depends on the nature of the data and the specific requirements of the analysis, making it essential to consider their respective strengths and weaknesses for successful applications in various domains.

- The optimization of the number of cluster. The validation of clustering structures is the most difficult and frustrating part of cluster analysis.

it is important to remember that clustering is an exploratory technique, and there may not be a clear-cut answer for the optimal number of clusters. it is a good practice to try multiple methods and cross-validate the results to choose the most appropriate K for your specific analysis. Additionally, the choice of clustering algorithm can also influence the optimal number of clusters, so you may want to try different algorithms as well.

4.3.2 Patient stratification

Patient stratification is a crucial approach in contemporary healthcare that customizes medical interventions according to individual patient characteristics. This process involves analyzing various types of data, such as clinical data, omics data, and images. In this particular section, we will specifically explore the concept of transcriptomic-based stratification and examine some noteworthy examples from the literature.

In a wide range of medical conditions, from cancer to autoimmune diseases, the traditional one-size-fits-all treatment approach may not be optimal. Thus, there is a growing recognition that separating patients based on their unique characteristics can offer valuable insights and lead to better-tailored treatments.

Oncology The initial classifications for breast cancer and B cell lymphoma were established during the early 21st century. Specifically, a significant milestone in the understanding of non-Hodgkin's lymphoma occurred in 1994 when a classification based on morphological and molecular parameters was published [97]. However, it was noted that the subtype diffuse large B cell lymphoma (DLBCL), although common, encompassed distinct diseases without a clear subclassification.

In a breakthrough study by A. Alizadeh *et al.* [98], DNA microarray cluster analyses were conducted on various lymphoma samples and normal hematopoietic cells. This research revealed that DLBCL was not only molecularly different from other lymphomas like chronic lymphoid leukemia and follicular lymphoma but also could be separated into two distinct molecular subgroups.

Another remarkable example, J. Guinney *et al.* successfully established a consensus gene expression–based subtyping classification system for colorectal cancer [99]. Subsequently, Becht *et al.* [100] conducted a comprehensive analysis, revealing a strong correlation between colorectal cancer molecular subgroups and microenvironmental signatures. This significant finding has opened the door for personalized immunotherapies.

Autoimmunity As discussed in the "Autoimmune" section, systemic autoimmune diseases exhibit considerable heterogeneity, which poses challenges in finding effective treatments. Bancherau *et al.* conducted a comprehensive study on 158 systemic lupus erythematosus (SLE) patients over a 4-year period [101], employing clinical and transcriptional profiling. They identified seven subgroups of SLE patients based on patient-specific modules. Each

subgroup displayed a distinct combination of five immune signatures related to SLEDAI: erythropoiesis, interferon response, myeloid lineage (including neutrophils), plasmablasts, and lymphoid lineage. Notably, the interferon signature was widespread across the subgroups, and the plasmablast signature emerged as the most robust biomarker of disease activity. Interestingly, the plasmablast signature showed higher levels in African American patients, who demonstrated better responsiveness to rituximab compared to white patients, possibly indicating differences in B cell biology between ethnic groups. In a 2019 study by Petri *et al.* [102], researchers used microarray analysis to investigate the stability of gene signatures over time in SLE patients. They also examined whether these signatures were linked to specific types of activity and if changes in the gene signature affected disease activity. The findings revealed that patients' gene-expression signatures remained relatively stable over time.

By leveraging transcriptomic information, researchers can gain deeper insights into the molecular signatures and genetic expression patterns of individual patients.

4.3.3 Gene Modules and Their Role in Precision Medicine

The core of my main project lies in the development of a fixed module repertoire, which originated from two papers. The first paper, authored by Damien Chaussabel and Nicole Baldwin [103] featured in Nature Reviews Immunology, explores how modular transcriptional repertoire analyses democratize systems immunology by facilitating gene expression pattern comparisons among researchers. The second paper, by Matthew C. Altman et al. [104], introduces a standardized framework for analyzing and interpreting blood transcriptome data, ensuring consistency and reliability in immunological research . This novel approach empowers the scientific community to gain valuable insights and enhance our understanding of complex immune-related processes.

As genes do not work alone but coordinately in regulatory pathways, a natural way to model the relationship between them is to derive Gene Co-expression Networks (GCNs). Such networks can be for instance used to assess the distance between nodes or sets of nodes. Also, the identification of nodes exhibiting more frequent interactions with other nodes (i.e. hubs), and nodes having frequent interactions with each other (i.e. modules, clusters, or communities) are generally of particular interest as they tend to correspond

to biological functions that may play a key role in biological or pathological processes. The detection of gene modules in GCNs has found number of strategic biomedical applications: for instance to better understand a disease, characterize patients, derive biomarkers (for classification, follow-up, response-to-treatment) and to develop therapeutic hypotheses [61, 80, 105]. By reducing the number of variables from ∼20000 genes to a much smaller number of modules, it operates as a dimension reduction approach and opens an opportunity to simplify the study of complex pathophysiological processes.

The process of identifying gene modules in Gene Co-Expression networks (GCNs), commonly known as network clustering, presents real-world obstacles. To address this, researchers have put forward multiple algorithms that can be categorized into two groups: GCN-specific algorithms like WGCNA [106] and more versatile algorithms applicable to diverse contexts, such as K-means.

5 Hypothesis, Objectives, and Strategies

In the introduction, I presented the intricate complexity not only within the realm of the immune system and autoimmune disorders, but more specifically in relation to Sjögren's syndrome. Furthermore, our attention was drawn to the potential of precision medicine approaches to contribute to an enhanced understanding of diseases. While numerous oncological conditions have greatly benefited from the timely implementation of precision medicine, the same level of success has not yet been achieved in rheumatic diseases. Despite the utilization of increasingly advanced technologies, along with sophisticated bioinformatic and machine-learning algorithms, the achievements seen in precision medicine for oncology have yet to be replicated in the context of rheumatic diseases. Initial initiatives have been deployed, notably through attempts to stratify patients (c.f. the PreciseSADS project).

More specifically, in context of Primary Sjögren's disease, physicians have noted the presence of distinct patient profiles. Given the diverse range of symptoms, it is apparent that individualized treatment is necessary, highlighting the complexity of the issue. Thus, Tarn et al. introduced a stratification model based on highly frequent symptoms in Sjögren context. Consequently, variations in interferon and B lymphocytes activity emerged across these stratified groups. In contrast, Soret et al. and Trutschel et al. proposed alternative stratification models. They utilized comprehensive transcriptomic patient blood samples to delineate subgroups within the patient population. Both investigations identified four patient clusters (refer to Table 1). Nevertheless, a challenge presents itself due to the divergence in results, stemming from the utilization of distinct cohorts in these models. This prompted me to conceive the notion rather of a gene-centered stratification approach through a consensus strategy, inspired by the work of Chaussabel. This entails leveraging multiple datasets to stratify genes, shifting the focus from patients to gene space. The overarching objective was to generate a collection of gene clusters applicable across diverse independent cohorts. This pursuit serves multiple practical purposes:

- Constructing a comprehensive chart of distinct molecular processes and cell categories throughout the disease.
- Formulating hypotheses for deeper comprehension of the disease.
- Stratifying patients based on identified patterns.

- Monitoring disease progression and activity.

A secondary objective is to implement these gene consensus modules within the framework of clinical trials (previously negative and ongoing trials). Specifically, our focus is on analyzing longitudinal data to assess whether we can predict treatment response by examining baseline information.

5.1 Methodological objective: define a pipeline to identify gene modules across multiple cohort

Methodologically, my objective is to establish a robust pipeline for the identification of gene modules across diverse - and independent - cohorts. This entails developing a systematic approach that can effectively detect and categorize groups of genes exhibiting coordinated behavior (e.g. highly correlated genes). By designing such a pipeline, we aim to enhance our ability to extract meaningful insights from complex transcriptomic data.

5.1.1 Integration of datasets

Aggregating multiple transcriptomic datasets can be challenging due to the inherent complexity and heterogeneity of biological samples. Each dataset may originate from different laboratories, platforms (micro-array or RNAseq), or experimental conditions, making direct comparisons of gene expression levels problematic. To address this issue, we opted for a strategy centered around transforming the data into correlation space, which simplifies comparisons across datasets. By focusing on correlations between genes rather than raw expression values, we can establish a common framework for integration. Once this correlation-based approach is applied, datasets can be merged cohesively. However, the process of merging is not as straightforward as taking a simple arithmetic mean, as biological data often involves intricate relationships that may not be accurately captured by basic averaging methods. Consequently, I decided to use a method proposed by Wang et al. [107], Similiarity Network Fusion, an iterative cross-diffusion algorithm that allowed to merge the four independant datasets aimed to eliminate the risk of filtering out any potentially significant features, including those with low signals. The SNF method involves a series of key steps. Initially, it computes pairwise similarities between samples within each data source to construct separate similarity networks. Next, it normalizes these similarity matrices

to ensure they are consistently scaled. A critical stage follows, wherein the normalized similarity matrices are combined to create an affinity matrix (using a sigmoid function) that captures sample relationships across all data sources. Finally, the authors recommend applying spectral clustering. Due to its multi-step nature, this method does require hyperparameter tuning.

5.1.2 Clustering of correlation matrix

Numerous methods exist for clustering correlation matrices. However, given our context of a graph-like structure, the Louvain clustering method is suited for the data structure. This method is particularly well-suited for automatically detecting clusters in a graph [108] and therefor in this context, gene context in a graph of co-expressed genes.

5.1.3 Annotation of clusters

The annotation of gene modules takes center stage, as it played a pivotal role in enabling robust downstream analyses. Recognizing the importance of this task, I adopted a comprehensive approach that combines multiple sources of gene modules and employs various annotation methods. This strategy encompasses gene enrichment analysis, also known as pathway analysis, leveraging different gene module repertoires to uncover the functional significance of these modules. Additionally, I incorporated the Microenvironment Cell Populations-counter (MCP-counter) method[100], a powerful tool for quantifying the absolute abundance of eight distinct immune cell populations, enhancing the depth and accuracy of our annotations. To further ensure the reliability of our results, a critical validation step was integrated, which can involve cross-referencing with matched flow cytometry data and cytokine measurements. This multifaceted approach to gene module annotation strengthens the foundation for rigorous and insightful downstream analyses.

5.2 Disease understanding objective: unifying immune and molecular classifications in pSD

Following the identification of gene modules, the objective was to elucidate the functional roles of these modules in patients. To achieve this, we conducted a comprehensive re-analysis of previously defined patient subgroups.

Our initial findings confirmed the significance of the IFN pathway, Inflammation and B lymphocytes as a pivotal determinants in patient stratification. However, we extended and enriched this classification by incorporating additional informative factors such as the involvment of Monocytes and Erythrocytes modules. This approach enabled us to gain deeper insights into the intricate interactions among these diverse gene modules, contributing to a more comprehensive understanding of the disease.

5.3 Clinical objective: A Retrospective exploration of clinical trials

Reassessing clinical trials (even negative), can hold significant value. Understanding the reasons for failures is crucial. Yet, accessing such data can be a complex endeavor. In my research, I am privileged to have access to an ongoing clinical trial, Repurpss-1, investigating the combined effects of hydroxychloroquine and leflunomide. Working closely with clinician from the projet, our goal is to utilize gene modules to pinpoint distinct cellular or functional dimensions that can facilitate the prediction of treatment responses.

5.4 Limitations of the work

Several limitations are inherent in this study. Although a deconvolution-like method is being employed, it is worth noting that utilizing bulk transcriptomics data has its constraints. The incorporation of single-cell RNA sequencing (scRNA) could potentially offer more robust and detailed insights. Additionally, the current research relies on whole blood transcriptomic profiles, whereas exploring transcriptomics within salivary or lacrimal glands, where the autoimmune reactions predominantly occur, could provide more targeted insights into the underlying mechanisms. Furthermore, applying the findings to clinical trials necessitates a larger patient cohort to establish a validation subset. This would enable the verification or refinement of hypotheses generated through computational algorithms, enhancing the reliability and generalizability of the results. One notable limitation of my study is that the developed transcriptomic profiling pipeline has been exclusively applied to patients with pSS, lacking analysis on healthy control subjects which could provide valuable insights for comparative analysis.

5.5 Importance of the study and potential impact in the research field

Despite its limitation, this study holds significant importance in advancing our understanding of complex biological systems. By constructing a repertoire of gene consensus modules and elucidating various axes encompassing immune cells and biological processes, this research contributes to unveiling the intricate interplay between genes within a broader context. The identified gene modules offer a novel approach to deciphering the molecular intricacies that underlie diseases, ultimately facilitating the development of more precise diagnostic tools and targeted therapeutic interventions. As a result, this study's impact extends beyond theoretical insights, with implications that could pave the way for transformative breakthroughs in the field of medical research and personalized medicine.

5.6 Article 1

Consensus gene modules strategy identifies candidate blood-based biomarkers for primary Sjögren's disease

Cheïma Boudjeniba, Perrine Soret, Diana Trutschel, Antoine Hamon, Valentin Baloche, Bastien Chassagnol, Emiko Desvaux, Antoine Bichat, Audrey Aussy, Philippe Moingeon, Céline Lefebvre, Sandra Hubert, Marta Alarcón-Riquelme, Wan-Fai Ng, Jacques-Eric Gottenberg, Benno Schwikowski, Michele Bombardieri, Joel A.G. van Roon, Xavier Mariette, Mickaël Guedj, Etienne Birmele, Laurence Laigle, Etienne Becht.

Submitted.

Consensus gene modules strategy identifies candidate blood-based biomarkers for primary Sjögren's disease

Cheïma Boudjeniba[1,2,3], Perrine Soret[1], Diana Trutschel[3], Antoine Hamon[4], Valentin Baloche[9], Bastien Chassagnol[1], Emiko Desvaux[1], Antoine Bichat[1], Audrey Aussy[1], Philippe Moingeon[1], Céline Lefebvre[1], Sandra Hubert[1], Marta Alarcón-Riquelme[5], Wan-Fai Ng[6], Jacques-Eric Gottenberg[7], Benno Schwikowski[3], Michele Bombardieri[8], Joel A.G. van Roon[9], Xavier Mariette[10], Mickaël Guedj[1], Etienne Birmele[11], Laurence Laigle[1] and Etienne Becht[1]

[1] Translational Medicine, Servier, Research and Development, Gif-Sur-Yvette, France
[2] Laboratoire MAP5 UMR 8145, Université Paris Cité, Paris, France
[3] Computational Systems Biomedicine Lab, Institut Pasteur, Université Paris Cité, F-75015 Paris, France
[4] Lincoln, Research and development, Paris, France
[5] GENYO, Centre for Genomics and Oncological Research. Pfizer, University of Granada, Spain
[6] Translational and Clinical Research Institute, Newcastle University, Newcastle upon Tyne, NE2 4HH, UK
[7] Rhumatologie, hôpitaux universitaires Strasbourg, CHU de Strasbourg, Strasbourg, France
[8] Centre for Experimental Medicine and Rheumatology, William Harvey Research Institute, Barts and the London, School of Medicine and Dentistry, Queen Mary University of London, Charterhouse Square, London, EC1M 6BQ, UK
[9] Department of Rheumatology and Clinical Immunology and Center for Translational Immunology, University Medical Center Utrecht, Utrecht University, Utrecht, Netherlands
[10] Department of Rheumatology, Université Paris-Saclay, INSERM UMR1184, AP-HP, Hôpital Bicêtre, Le Kremlin Bicêtre, France
[11] Institut de Recherche Mathématique Avancée, UMR 7501 Université de Strasbourg et CNRS, Strasbourg, France

Corresponding author: etienne.becht@servier.com

Abstract *Primary Sjögren disease (pSD) is an autoimmune disease characterized by lymphoid infiltration of exocrine glands leading to dryness of the mucosal surfaces and by the production of autoantibodies. The pathophysiology of pSD remains elusive and no treatment with demonstrated efficacy is available yet. To better understand the biology underlying pSD heterogeneity, we aimed at identifying Consensus gene Modules (CMs) that summarize the high-dimensional transcriptomic data of whole blood samples in pSD patients. We performed unsupervised gene classification on four data sets and identified thirteen CMs. We annotated and interpreted each of these CMs as corresponding to cell type abundances or biological functions by using gene set enrichment analyses and transcriptomic profiles of sorted blood cell subsets. Correlation with independently measured cell type abundances by flow cytometry confirmed these annotations. We used these CMs to reconcile previously proposed patient stratifications of pSD. Importantly, we showed that the expression of modules representing lymphocytes and erythrocytes before treatment initiation is associated with response to hydroxychloroquine and leflunomide combination therapy in a clinical trial. These consensus modules will help the identification and translation of blood-based predictive biomarkers for the treatment of pSD.*

Keywords Precision Medicine, Sjögren Disease, Unsupervised learning, Integrated analysis.

Introduction

Primary Sjögren Disease (pSD) is a chronic, disabling inflammatory autoimmune disease characterized by lymphoid infiltration of exocrine glands leading to dryness of the mucosal surfaces, such as the mouth and eyes and by the production of specific auto-antibodies[1–3]. Long-term complications include ocular and dental diseases, systemic involvement, organ damages and increased risk of lymphoma with poor prognosis[4, 5]. This pathology is affecting between 0.05% and 0.4% of the adult population[6–9] and is the second most common systemic autoimmune disease[10]. It affects women more often than men (9:1) and the peak frequency of the disease is around fifty years of age[11].

The advent of new technologies has provided a path towards the development of classification criteria for autoimmune diseases that are based on molecular patterns representing disease mechanisms and molecular pathways[12, 13]. By applying computational methodologies to clinical and multi-omic datasets, several pSD disease taxonomies have recently been proposed. Indeed, Tarn et al. proposed a symptom-based stratification of patients with pSD[14], while Soret et al.[15] and Trutschel et al.[16] proposed a molecular classification of pSD based on whole blood transcriptomic profiles of pSD patients. These classifications may provide useful clinical insights on disease subtypes of pSD patients but remain limited in the characterization of the biology underlying the disease in each patient subgroup. Indeed, pathogenesis of autoimmunity involves dysfunction of the entire immune system, and many cellular or functional components, including neutrophils, dendritic cells (DCs), macrophages, T and B cells, cytokine signaling pathways or autoantibodies[17, 18].

The clinical manifestations and biological disturbances associated with pSD are indeed highly heterogeneous among individuals which complicates its diagnosis. Mechanistically, the pathophysiology of pSD remains elusive[19]. No targeted therapy is therefore currently approved and only symptomatic treatments are offered[20, 21]. Precision Medicine approaches designed to better address the needs of patients based on the specific biological mechanisms underlying their symptoms would greatly improve the management of patients suffering from pSD.

The IMI 2 NECESSITY European consortium was launched in 2019 to identify a new composite clinical endpoint, biomarkers for stratifying patients and predictive biomarkers of treatment response for pSD, and test them in a prospective clinical trial. To achieve these goals, members of the NECESSITY consortium share clinically-annotated datasets, including whole blood transcriptomic datasets of pSD patients. These transcriptomes allow the identification of biological heterogeneity across pSD patients and its potential link with response to treatments, but were produced using diverse transcriptomic technologies, making their combined analysis challenging.

In order to jointly analyze independent whole blood transcriptomic datasets of pSD patients, we used a graph theoretical approach to unify four correlation networks into a consensus graph linking positively correlated genes. By clustering this unified representation of multiple cohorts, we identified 13 consensus transcriptomic gene modules that summarize the pathophysiology of pSD at the blood level. We annotated each of these modules for correspondence with cell types or molecular pathways, and validated these biological interpretation with matching flow cytometry data or cytokine measurements whenever available. We used these modules to better characterize and reconcile previously-published pSD patient stratifications[15, 16]. Importantly, we investigated clinical trial data to decipher the impacts of treatments on the peripheral blood of patients and propose a model predictive of the response to leflunomide-hydroxychloroquine combination therapy.

Results

Identification of thirteen consensus gene modules (CMs) from whole blood transcriptomes of pSD patients

We analyzed four whole blood transcriptomic datasets from pSD patients. Three were provided by the NECESSITY consortium: ASSESS[22] ($n = 371$), PreciseSADS[12] ($n = 341$) and UKPSSR[23] ($n = 144$). We also included the publicly-available GSE84844[24] dataset ($n = 30$). Our goal was to identify consistent signals across these four sources, and in particular consensus gene modules (CMs) of coexpressed genes. Transcriptomic data sets are however high dimensional which can hamper the correct identification of gene modules. Indeed, spurious correlations may appear due to the size and noisiness of the data: 20,000 protein coding genes indeed correspond to 400×10^6 correlation coefficients. To ensure that the CMs we identify were reproducible across a large range of blood transcriptomic data sets (from distinct pSD cohorts), we used a dedicated analysis workflow summarized in **Figure 1A**. We first converted each cohort's gene expression matrix to an affinity matrix (gene co-expression network). This affinity is non-linearly and monotonically linked to the observed correlation between two genes and shrinks low correlation coefficients towards 0 (See **Methods** and Wang *et al.*[25]). We applied Similarity Network Fusion (SNF)[25], a computational method designed for the merging of multiple affinity matrices, generating a consensual representation of genes' pairwise

similarities in the blood of pSD patients across these four independent cohorts (**Figure 1B**). We pruned the consensual affinity matrix to obtain a sparse weighted graph with edges corresponding to highly co-expressed genes (**Supplementary Figure 1**). Finally, Louvain clustering[26] of the sparse graph (see **Methods**) identified 13 CMs (**Supplementary Table 1**). We confirmed a posteriori that these CMs are reproducible groups of highly co-expressed genes that are reproducible across the four datasets (**Figure 1C**).

Biological interpretation of the CMs

The 13 CMs represent the main axes of heterogeneity of the blood transcriptome across pSD patients and can therefore facilitate the interpretation of high dimensional transcriptomic data by summarizing it using 13 dimensions. In order to biologically interpret these 13 axes of variation, we annotated each of them as corresponding to cell types or biological functions by using gene set enrichment analyses using gene sets from the Gene Ontology[27] and Altman *et al.*[28] databases (**Figure 2A, 2B**), as well as their average expression in transcriptomic profiles of sorted blood cell subsets[29] (**Figure 2C**).

CM1 was enriched in Interferon related as well as response to viruses pathways, and we interpreted it as representing type 1 IFN signaling. CM7 was enriched in cell cycle-related genes, and we interpreted it as a transcriptomic signature of mitosis within blood cells.

Out of the 11 other modules, 9 represent different cell types. We found four modules corresponding to lymphoid cells: CM4, CM5 and CM11 were respectively enriched in pathways associated with T cells, NK cells and B cells functions **(Figure 2A, 2B)** and that were overexpressed in the transcriptome of the corresponding purified cell types **(Figure 2C)**. CM8 was enriched in genes associated with gene transcription and overexpressed across the transcriptomes of purified lymphocytes (T, B and NK cells) and therefore represents a shared gene transcription signature across all lymphocytes **(Figure 2C)**. We found six modules (CM2, CM6, CM9, CM10, CM12, CM13) representing myeloid cell subsets. CM2 was enriched in erythrocytes-annotated gene sets and CM10 in platelets-annotated gene sets. Module CM6 was overexpressed in the transcriptome of eosinophils. CM9 and CM13 were enriched in inflammation and neutrophil-related gene sets and overexpressed in the transcriptome of purified granulocytes and neutrophils. CM13 was in addition enriched in genes from the I-κB kinase/NF-κB signaling pathway, an inflammatory transcription factor expressed by neutrophils[30]. Finally, CM12 was enriched in gene sets related to monocytes and overexpressed in the transcriptome of cells derived from monocytes.

Among the 13 CMs, CM3, which contains the highest number of genes (n=1247), was the least co-expressed, had the lowest absolute expression levels **(Supplementary Figure 2)** module and showed inconsistent characterization results **(Figure 2A, 2B)**. We therefore did not take it into consideration for further analysis. In summary, we interpreted CM1 as type 1 interferon (IFN) activation, CM2 as representing the frequency of erythrocytes within the blood, CM3 as residual variance, CM4, CM5, CM6 as the frequencies of respectively T cells, NK cells and Eosinophils, CM7 as a signature of cell proliferation, CM8, CM10, CM11 and CM12 as the frequencies of respectively lymphocytes, platelets, B cells and monocytes, and CM9 and CM13 as representing neutrophils.

Validation of the biological interpretations of the CMs

To confirm the biological interpretations of the CMs representing cell types, we compared their average expressions (**Material and Methods**) to the corresponding cellular frequencies measured by flow cytometry in matching samples whenever available (**Figure 3A**). For functional modules, we compared them to previously-published gene signatures (**Figure 3B**) or cytokines concentrations (**Figure 3C**).

For all the cellular modules for which we had matching cytometry data, we observed a high and significant correlation of the average module expression with the frequency among live single cells measured by flow cytometry (**Figure 3A**). More precisely, we observed correlation coefficients of 0.71 between the CM4 module and the frequency of T cells, of 0.51 between the CM5 modules and NK cells, of 0.39 between CM6 and eosinophils, 0.75 (respectively 0.64) between CM9 (respectively CM13) and

neutrophils, 0.84 between CM11 and B cells, 0.67 between CM12 and monocytes, and 0.62 between CM8 and lymphocytes (all p-values $< 2 \times 10^{-12}$).

For functional modules, we observed a strong correlation (Pearson's $r > 0.94$) of the CM7 with genes signatures corresponding to phases of the mitotic cycle identified with single cell RNA-sequencing data[31]. The other functional module CM1 was highly correlated with the concentration of type 1 IFN (measured by SIMOA) in the blood ($r = 0.65, p = 3.3 \times 10^{-11}$) (**Figure 3C**). Collectively, these analyses confirm the interpretation of the CMs derived from gene set enrichment analyses.

The consensus gene modules identify consistency and heterogeneity across pSD patient stratifications

Three studies have proposed pSD patients stratifications according to molecular and clinical features of the disease[14–16]. Two methods were based on blood transcriptomic profiles of pSD patients on two distinct cohorts[15, 16]. Both studies identified four clusters of patients hereafter referred to as S1, S2, S3 and S4 (respectively T1, T2, T3 and T4) for the Soret (respectively Trutschel) classification. These stratifications were established using unsupervised clustering methods. Algorithmic classifiers to stratify new pSD cohorts according to these classification systems are however currently lacking, and no direct comparison has been performed so far.

Briefly, from Soret *et al.*, cluster S1 exhibited high levels of interferon (IFN) activity and an increased frequency of B lymphocytes in the blood. Cluster S2 showed a similar expression profile to that of healthy volunteers. Cluster S3 displayed a high IFN signature, along with a more prominent involvement of B cell components compared to other clusters, including an increased frequency of B cells in the blood. Lastly, cluster C4 was characterized by an inflammatory signature driven by monocytes and neutrophils. Confirming the findings of *et al.*[15], our analysis confirmed the defining characteristics of these patient clusters. We consistently observed an upregulation of the Interferon module CM1 in S1 patients, the Neutrophils module CM9 in S4 patients, and the B cell module CM11 in S3 patients (Figure 4A). Our analysis further revealed that S3 is defined by a high abundance of lymphocytes (B, T, and NK cells represented by the CM11, CM4, and CM5 modules, respectively) associated with cell proliferation (CM7). Cluster S4 is characterized by a high abundance of platelets (CM10), erythrocytes (CM2), and neutrophils (CM9 and CM13). S1 is distinguished by high activation of type 1 IFN (CM1), while S2, described as normal-like by Soret *et al.*, has fewer monocytes (CM12) and more T cells (CM4) compared to the cohort's averages.

In a separate study by Trutschel *et al.*, four patient clusters were also identified. These clusters were based on two modules: IFN-stimulated genes (ISGs) and the erythroid module (ERM). Cluster T1 showed high expression of both these modules, while cluster T2 had low ISG expression but high ERM expression. Cluster T3 had high ISG expression and low ERM expression, and cluster T4 had low expression in both ISGs and ERM. We observed a high interferon signature (CM1) in clusters T1 and T3, with cluster T1 exhibiting a higher platelet presence compared to cluster T3 (Figure 4B). Cluster T2 had a lower abundance of monocytes (CM12), while cluster T4 had a high neutrophil signature (CM13). Cluster T1 had a high presence of erythrocytes, cluster T3 had fewer eosinophils (CM6), and clusters T3 and T4 had a higher abundance of lymphocytes (CM8).

To formally study the correspondence between the Soret and Trutschel classification systems, we computed Pearson correlation coefficients across centroids computed on mean-centered and unit variance-scaled module expression scores. This comparison highlighted a very high concordance between cluster S2 and T2 (r = 0.9), good concordance between clusters S1 and T1 (r = 0.6), moderate across clusters S3 and T3 (r = 0.4), and poor concordance across clusters S4 and T4 (r = 0) (**Figure 4E**). This analysis shows that there is a substantial overlap between the two classification systems, especially in the identification of T2 patients.

It therefore appears that cluster S1 of the Soret classification corresponds to cluster T1 of the Trutschel classification, marked by high type 1 IFN signaling (CM1) **(Figure 4C, 4D)**. Cluster S3 matches cluster T3, as identified by high type 1 IFN signaling (CM1) in the context of a lower abundance of platelets (CM10) and erythrocytes (CM2). Cluster S2 matches cluster T2, with the lowest type 1 IFN signature (CM1). Cluster S4 in resembles cluster T4, as both have the highest

expression of the Neutrophil activation module (CM13), although other modules such as platelets (CM10) and erythrocytes (CM2) had discordant expression levels across the two patient classification systems. In general, there were no differences in the lymphoid modules (CM4, CM5 and CM11) across Trutschel clusters.

Tarn et al. propose a stratification model based on patient-reported symptoms and identified four clusters of patients: Low symptom burden (LSB), high symptom burden (HSB), dryness dominant with fatigue (DDF), and pain dominant with fatigue (PDF). We were unable to see any significant difference in the level of expression of any CM across the four subgroups of patients **(Supplementary Figure 3)**. Consistently, we observed -in the PreciseSADS and ASSESS cohorts- weak correlations of the CMs expression scores with the ESSDAI[32] and ESSPRI[33] disease activity scores **(Supplementary Figure 4)**. We however noted that unlike other components of the ESSDAI and ESSPRI disease activity scores, the presence of autoantibodies (anti-SSA, anti-SSB, PFLC, IgG) was positively-associated with the CM1 module representing type 1 IFN **(Supplementary Figure 5)**. These observations suggest that among pSD clinical manifestations, the presence of autoantibodies is the most associated with a specific blood transcriptomic profile.

CM8 and CM2 are associated with response to hydroxychloroquine and leflunomide combination

Many clinical trials for Sjögren's patients have shown poor results especially for response to treatment[34–37] but, negative clinical trials can still provide valuable information about the efficacy of a particular treatment and can help guide future research. However, positive trials provide a unique opportunity to compare responder and non-responder patients' characteristics. Within the IMI2 NECESSITY, data from both positive and negative clinical trials are available for exploratory retrospective analyses. RepurpSS-1[38] is a placebo-controlled, double-blinded, phase 2A randomized clinical trial that evaluated the combination therapy of hydroxychloroquine and leflunomide and is one of the first positive clinical trials in pSD.

Firstly, we validated the co-expression of the genes within each CM on this cohort independent of those used for the identification of the modules, highlighting the reproducibility and generalizability of the CMs to independent pSD blood transcriptomic datasets **(Supplementary Figure 7)**.

Secondly, we looked at the evolution of the expression of each module between treatment initiation and completion. We observed that lefluonomide-hydroychloroquine combination led to a decrease in the expression of CMs representing T cells, platelets and B cells, and an increase expression of the CMs representing monocytes and neutrophils, thus suggesting that this treatment combination favored the number of myeloid immune cells over lymphoid immune cells in the blood (**Figure 5A**). While treatments received by patients before blood transcriptomic profiling were more heterogeneous in the PreciseSADS cohort, we consistently observed an influence of the type of treatment received on the expression level of the CMs (**Supplementary Figure 6**).

Finally, we examined whether the heterogeneity of the patients encompassed in the modules could help identify responders in the RepurpSS-1 trial before treatment initiation. To do so, we focused on the recently developed STAR clinical endpoint[39]. The CM8 Lymphoid Lineage module was significantly overexpressed in responders before treatment initiation ($q = 0.013$) (**Figure 5B, 5C, Supplementary Figure 8)**. Conversely, a trend for higher expression in non-responders of the CM2 module representing erythrocytes was also found ($q = 0.055$). By combining CM2 and CM8, we were able to perfectly separate responders and non-responders in this clinical trial (**Figure 5D**). These analyses suggest that these cell populations could represent biomarkers predictive of therapeutic efficacy of this treatment combination.

Discussion

Primary Sjögren's disease (pSD) is a debilitating and clinically heterogeneous disease with no well-established causal mechanism, nor approved targeted therapy. There is therefore an urgent need to identify biomarkers able to inform treatment selection as well as to stratify patients in clinical trials in the context of personalized medicine. High throughput transcriptomic profiling is an appealing

technology for biomarker discovery as it allows the interrogation of tens of thousands of genes for differential expression across groups of patients, such as responders and non-responders to a drug in a clinical trial. The interpretation of transcriptomic profiles is however difficult, as groups of differentially expressed genes may represent dysregulation of functional pathways or changes in the cellular composition of samples, or both. In addition, the very high dimensionality of whole transcriptome assays makes difficult distinguishing true and replicable biological signal from noise.

To overcome these difficulties in the interpretation of the transcriptome in the context of pSD, we jointly analyzed four independent transcriptomic datasets profiling whole blood samples from pSD patients. We used clustering methods to identify the main axes of variation across these four datasets. As clustering algorithms are sensitive to noise, we implemented a method to perform a gene clustering analysis on a joint representation of the pairwise gene correlations matrix across the four datasets, rather than on each dataset separately. To do so, we recast the four observed matrices of pairwise gene correlations as graphs and used the SNF[25] algorithm to obtain a consensus graph representation of the gene correlation network across the four cohorts, on which we applied the Louvain graph clustering algorithm. We importantly showed that the gene modules we identified are reproducible across the four cohorts on which they were discovered **(Figure 1C)** as well as on an independent cohort (**Supplementary Figure 7**). These modules therefore represent the main biological features contained in the transcriptomic profile of the whole blood in pSD patients, therefore facilitating its interpretation for translational research.

In order to make the CMs more biological meaningful, we interpreted them using distinct public databases of pathways and blood cells transcriptomes[29]. This allowed us to identify both functional modules (interferon signaling or cell proliferation) or modules reflecting the cellular composition of the patients' blood. Importantly, we observed highly significant correlations between the expression of the gene modules and corresponding cellular frequencies or cytokine levels, thus validating these computationally derived biological interpretations. In the recent years, so called transcriptomic deconvolution methods have been proposed in order to infer cellular proportions from transcriptomic measurements[40]. Most of these methods rely on a reference averaged transcriptomic profiles of cell types, usually derived from purified cells from the blood of healthy donors and use genes that are discriminative across cell populations in a given context, such as cancer[29]. In contrast, our approach is driven by the observed variations in the blood of pSD patients across multiple cohorts, ensuring that the gene signatures of the identified cell types are valid in this context. In addition, this data driven approach allowed us to define gene modules indicative of rare cell populations such as eosinophils or signatures of non-immune cell types such as erythrocytes or platelets which are not typically quantified by deconvolution algorithms[41]. Moreover, we found functional modules (CM1 type 1 IFN and CM7 Cell Cycle) that do not correspond to variations in the frequencies of blood cell types. The consensus gene modules described herein therefore could help understanding the complex pathophysiology of pSD as they represent biologically meaningful, reproducible, and sensitive sources of heterogeneity in the blood transcriptome of pSD patients.

The gene modules that we identified can serve as a building block for translational research in pSD, by providing a concise list of potential biomarkers provided by whole blood transcriptomic profiling. Multiple independent studies have recently focused on the stratification of the disease into discrete patient subgroups, based on whole blood transcriptomic profiles[15, 16] or clinical characteristics[14]. These classifications systems may become relevant in future clinical trials, as new treatments may benefit only to a restricted subset of patients. Our approach complements these classifications by highlighting the functional and cellular composition differences across patient subgroups, as well as highlighting the consensus and differences across classification systems. Our analyses notably suggest that the patient subgroups in published transcriptomic-based patient stratification systems can be distinguished based on the measurement of three variables: the frequency of neutrophils in the peripheral blood, the concentration of type 1 IFN, as well as the frequency of either erythrocytes or platelets within the blood (**Figure 4C, 4D**). These observed differences across patient subgroups may provide clinically actionable biomarkers for disease stratification in settings where whole blood transcriptomic

profiling is impractical. Indeed, these key features of pSD drive disease heterogeneity and altogether may be useful predictors of response.

Some medications are designed to target specific genes or proteins, altering their activities and ultimately leading to changes in cellular behavior. Understanding the complex relationship between medications and gene expression is an important area of research that includes Drug Repurposing computational activities and may eventually lead to the definition of more effective treatment strategies for a wide range of diseases and conditions. Our analyses showed that the CMs can be used to understand the effect of drugs on the composition and functional orientation of the peripheral blood **(Figure 5, Supplementary Figure 7)**. We also confirmed, in two independent cohorts, the correlation between the presence of anti-SSA and anti-SSB autoantibodies and the level of type 1 IFN in the peripheral blood. The pathogenic role of the IFN pathway has been extensively described: type I IFN signature is correlated with the development of systemic extra-glandular manifestations, and a substantial production of autoantibodies and inflammatory cytokines[42]. Moreover, in the context of systemic autoimmune manifestations, pSD patients may present with hematologic abnormalities including anaemia, leukopenia (mainly neutropenia or lymphopenia), and thrombocytopenia[43, 44]. These three components are indeed evaluated in the haematological domain of the ESSDAI scale. As these patient characteristics are recapitulated by our CMs, whole blood transcriptomic profiling thus appears informative in the context of pSD translational research.

The CMs we identified indeed provide a succinct list of candidate blood-based biomarkers that recapitulate whole transcriptome profiles in a biologically interpretable manner. These modules can therefore be examined in exploratory and clinical research for their potential association with the response to a treatment or to study drug mechanism of action. We exemplified this idea by retrospectively analyzing data from the RepurpSS-1 phase IIa clinical trial[38] which evaluated a combination of leflunomide and hydroxychloroquine for the treatment of pSD. Longitudinal whole blood transcriptomic profiling allowed us to show that this combination led to a decreased expression of CMs corresponding to T cells, platelets and B cells, and an increase in modules representing monocytes and neutrophils. Our results therefore show that this combination of treatments influence the cellular composition of the peripheral blood in pSD patients.

Importantly, we investigated the relationship between each CM expression levels before treatment initiation and the observed clinical response upon completion of the clinical trial. Our results show that responders to this treatment combination featured higher expression of the module representing lymphocytes and a trend for lower expression of the module representing erythrocytes. These observations are consistent with the mechanism of action of leflunomide, an immunomodulatory drug known to inhibit de novo synthesis of pyrimidine, preventing lymphocytes from expanding in inflammatory context[45]. While the mechanism of hydroxychloroquine is less clear considering its initial use as an antimalarial drug, this molecule has widely been used in rheumatic autoimmune diseases such as systemic lupus erythematosus[46]. Studies have shown that hydroxychloroquine can contribute to regulate inflammation by blocking Toll-like receptors (TLR) leading to type I IFN pathway inhibition[47]. Hydroxychloroquine has also demonstrated inhibitory effect on platelet activation[48], in accordance with modulations seen on CM relating to platelets in the RepurpSS-1 clinical trial. Our results suggest that clinical efficacy for this treatment combination may be restricted to patients with high lymphoid frequency and low erythrocytes frequency, thus providing new hypotheses guiding the treatment strategy of pSD patients and the design of future clinical trials.

Our work is therefore expected to facilitate translational and clinical research on primary Sjögren's disease by presenting a set of reproducible and annotated gene modules that capture the major variations in the blood transcriptome of patients, which will open up the path for identifying biomarkers in clinical trials for this disease that is still poorly managed.

Acknowledgements

Funding: This project has received funding from the Innovative Medicines Initiative 2 Joint Undertaking (JU) under grant agreement number 806975. JU receives support from the European Union's

Horizon 2020 research and innovation program and EFPIA. The present article reflects only the authors' view and JU is not responsible for any use that may be made of the information it contains.

The UKPSSR is established with the funding provided by the Medical Research Council (G0800629), with additional infrastructural support from the British Sjogren's syndrome association, NIHR Newcastle Clinical Research Facility and the NIHR Newcastle Biomedical Research Centre.

Author contributions

Conceptualization: C.B., M.G., E.Be, E.Bi, L.L.

Methodology: C.B., M.G., E.Be, E.Bi, A.H., B.C.

Validation: D.T, P.S, E.D.

Formal analysis: C.B, A.H, B.C., A.B, E.Be

Writing – Original Draft: C.B., E.Be, E.Bi, L.L.

Writing – Review and Editing: P.S., D.T., A.H., B.C, C.L., A.A., S.H., P.M., E.D, A.B, M.A.R, W.F.N, J.R, J-E.G., B.S., X.M, M.G.

Resources: J-E.G, M.E.A.R, W.F.N, J.R.

Supervision: M.G., E.Be, E.Bi, L.L.

We, the authors of this manuscript, confirm that we have collectively agreed to submit this work for publication. We have read and approved the final draft and take full responsibility for its content, including the accuracy of the data presented. We have also ensured that the statistical analysis, where applicable, was conducted appropriately and accurately. As authors, we are committed to upholding the highest standards of scientific integrity and ethical conduct, and we affirm that this work represents our best efforts to contribute to the advancement of knowledge in our field.

Declaration of Interests

While engaged in the research project, C.B., B.C. and E.D. were phD students financed by Institut de Recherches Internationales Servier when they contributed to the research project. P.S., A.B., A.A., P.M., M.G., L.L., C.L., S.H., and E.Be were employees Institut de Recherches Internationales Servier when they contributed to the research project. W.F.N. has provided consultation for Novartis, GlaxoSmithKline, Abbvie, BMS, Sanofi, MedImmune, Argenx, Janssen, Resolves Therapeutics, Astella and UCB.

Figures

Figure 1 A) Schematic summary of the work. pSD = primary Sjögren Disease **B)** Heatmap of the consensus pairwise gene affinity computed by Similarity Network Fusion (SNF). Side annotations represent gene modules.**C)** Heatmaps of Pearson's correlation matrices of the four input datasets, with genes grouped by their consensus gene modules.

Figure 2 A) For each module, the two most significantly-enriched pathways in the Chaussabel database[28]. B) Most significantly-enriched pathways in the GO database[27] C) Average expression of modules in transcriptomes of purified cells

Figure 3 A) Significant Pearson's correlations between the average expression of the CMs and cell types abundances measured by flow cytometry. Scatter plots of average CMs expression and matching cellular frequencies. B) Scatter plots illustrating the average expression of CM7 versus averages of cell cycle signatures C) Scatter plot of the average expression of CM1 type 1 IFN and dosage of type 1 IFN

Figure 4 CMs scores across patient subgroups of A) the Soret classification B) the Trutschel classification. Average expression of the CM1 type 1 IFN, CM2 Erythrocytes, CM10 Platelets and CM13 Neutrophils.2 CMs in the C) Soret classification and D) Trutschel classification. E) Correlation across cluster centroids of the two stratification systems.

Figure 5 A) Boxplots illustrating the evolution of the modules significantly differentially-expressed at baseline (BL) versus Week 24 for treated patients B) Heatmap of baseline average gene expression of the CMs. Patients are split by their responder status according to the STAR clinical endpoint. Right side annotations indicate FDR corrected p-value (qvalue) C) Avegage expression of CM8 and CM2 at baseline in responders versus non-responders D) Dotplot of average expression of the CM8 and CM2 modules, colored by response statuses.

Material and Methods

Data collection

Gene expression and associated clinical and biological data was obtained through tranSMART, the NECESSITY consortium data sharing platform for the ASSESS (Assessment of Systemic complications and Evolution in Sjögren's Syndrome) cohort[22], PRECISESADS[12] and UKPSSR[23] cohort. Data from the fourth cohort was downloaded from the Gene Expression Omnibus repository, under the accession number GSE84844[24].

Transcriptomic data pre-processing

The UKPSSR RNA-seq count data was transformed as in[14]. RNA-seq data from the PreciseSADS cohort was normalized as in Soret *et al.*[15]. The ASSESS Affymetrix Clariom S microarray data were normalized as in[16].

The GSE84844 Affymetrix Human Genome U133 Plus 2.0 Array data was pre-treated by filtering out probesets indistinguishable from background noise. For that purpose, we modeled probesets expression after applying a $log2(x+1)$ transformation by a two component Gaussian mixture model[**dempster˙maximum˙1977**] with the first peak corresponding to unexpressed genes, and the second peak to expressed genes. We retrieved the parameters of the mixture distribution using the function *normalmixEM* from the *mixtools* R package. The 0.95^th quantile of the first component of the distribution was used as a threshold. Probesets whose expression were below that threshold in more than 95% of the samples were removed. Finally, the fRMA function from the fRMA R Package[**McCall˙Bolstad˙Irizarry˙1970**] was used to normalize probesets intensities across samples.

Finally, to have comparable data sets, the intersection of the 80% most varying common genes across all the data sets was selected (5443 genes).

Integrated affinity network

The construction of the integrated network involves two steps: First, gene affinity ($affi$) is computed independently on each data set as follow : for each pair of genes (x, y), we consider the affinity between x and y as $affi_{(x,y)} = exp((1 - cor(x, y))/\sigma)$ where cor is the Pearson correlation coefficient and $\sigma = 3$, as suggested by Wang *et al.*[25]. The four networks are then merged into an integrated affinity network by using the Similarity Network Fusion (SNF) method[25], with 30 neighbours per gene and 20 iterations. The SNF algorithm produces a weighted fully connected graph with $5000^2 = 2.5 \times 10^6$ edges. Visual inspection of the distribution of the weights showed that their distribution was bimodal, with a largely preponderant low weight peak **[Supplementary Figure 1]**. To convert the fully connetected output of the SNF algorithm to a sparse graph, we removed edges below the $0.9775th$ quantile of the weights distribution **(Supplementary Figure 1)**.

Consensus modules identification

Consensus gene modules were identified by applying the Louvain clustering algorithm[26] on the fused and truncated graph of pairwise gene affinities. This method is based on a modularity optimization algorithm that aims to partition genes into communities with high within-group affinity and low between-group affinity. The modularity score of a community structure is calculated as the difference between the weighted proportion of intra-community edges and the expected weighted proportion of such edges if the edges were randomly distributed.

Gene modules summarization

We used the mean expression the genes contained in a module to represent that module's expression as performed in Becht et al[29].

Gene set enrichment analysis

Enrichment analysis is performed by applying a Fisher-exact tests on the human blood-derived transcriptomic modules of Altman *et al.*[28] as well as the Gene Ontology database[27]. P-values were corrected using the Benjamini-Hochberg procedure to select pathways by controlling the false discovery rate at a 0.05 level.

Mapping with purified and sorted immune cells

To identify modules representing the abundances of blood cell types, we used the GSE86362 dataset[29], which consists of 1936 gene expression profiles from immune cell populations, non-immune non-malignant cell populations and non-hematopoietic cancer cell lines. For consistency with our sample types, we only retained samples corresponding to blood cell populations ($n = 1095$).

Correlation between CMs and cell type abundances measured by Flow Cytometry

On the PreciseSADS cohort, proportions of relevant cell types using flow cytometry custom marker panels were analyzed for samples where matched transcriptomic profiles and cytometry data were available. Correlations were performed between summarized CM expression levels and log-frequencies of the corresponding cell populations among live single cells, as previously described[29]. We corrected the p-values by Benjamini-Hochberg (BH) procedure by controlling the False Discovery Rate (FDR) at a 0.05 level.

Correlation between CMs and cytokines

On the PreciseSADS cohort, relevant cytokines were measured as in[12]. A log transformation was applied on the concentrations. Finally, we computed correlations tests between the average expression of the CMs and the cytokines levels we corrected the p-value by controlling the FDR at a 0.05 level (BH procedure).

Application to clinical trial

RepurpSS-1 (registered under trial number EudraCT, 2014–003140–12) was a phase II a placebo-controlled clinical trial testing a combination of Leflunomide and Hydroxychloroquine[38]. Gene expression and associated biological and clinical data for the RepurpSS-1 trial was obtained through the NECESSITY consortium. Transcriptomes of samples with a RIN < 6 or DV200 > 70 were excluded, resulting in the analysis of 16 patients. Pre-treatment and post-treatment (at week 24) CM expression levels were compared using paired t-tests with Benjamini-Hochberg correction. Responder status was determined based on the STAR clinical composite endpoint[39]. Patients with a STAR score of 5 or above were classified as responders. Difference in CM expression levels between responders and non-responders were assessed using univariate t-tests with BH FDR correction.

Supplementary materials

Table 1. List of genes (SYMBOL) in each Concensus Modules (CMs)

Supplementary Fig1. Histogram showing the distribution of weights in the SNF matrix. The x-axis denotes the weight range (logged) and the y-axis represents the frequency of weights. A vertical red line indicates the discretization threshold corresponding to the 0.975^{th} quantile (for better visualization).

Supplementary Fig2. A) Average correlation of the 4 input datasets **B)** Average of average correlation matrices **C)** Average gene expression levels for each CM in cohorts profiled by RNA-sequencing

Supplementary Fig3. CMs scores across patient subgroups of the Tarn classification in UKPSSR cohort

Supplementary Fig4. Pearson's correlation between average CMs expression and ESSDAI and ESSPRI scores in A) PRECISESADS and B) ASSESS cohorts

Supplementary Fig5. Pearson's correlation between average CMs expression and autoantibodies levels in A) PRECISESADS and B) ASSESS cohorts

Supplementary Fig6. A)T-test between average CMs expression and treatment. q = corrected p-value **B)**CMs expression scores across patients stratified by treatments received. AM = Antimalarials, STD = Steroids, IS = Immunosupressors **C)**Significant diffrences observed in treated versus untreated patients.

Supplementary Fig7. Correlation matrix in REPURPSS-1 cohort, sorted by CMs.

Supplementary Fig9. Boxplots of average expression of the CMs at baseline versus after treatment splitting patients by treatment and placebo.

Supplementary Fig8. Boxplots of average expression of the CMs versus response status.

References

1. Mariette, X. & Criswell, L. A. Primary Sjögren's Syndrome. *New England Journal of Medicine* **378** (ed Solomon, C. G.) 931–939 (Mar. 2018).
2. Brito-Zerón, P. *et al.* Sjögren syndrome. *Nature Reviews Disease Primers* **2** (July 2016).
3. Parisis, D., Chivasso, C., Perret, J., Soyfoo, M. S. & Delporte, C. Current State of Knowledge on Primary Sjögren's Syndrome, an Autoimmune Exocrinopathy. *Journal of Clinical Medicine* **9,** 2299 (July 2020).
4. Solans-Laqué, R. *et al.* Risk, Predictors, and Clinical Characteristics of Lymphoma Development in Primary Sjögren's Syndrome. *Seminars in Arthritis and Rheumatism* **41,** 415–423 (Dec. 2011).
5. Nocturne, G., Pontarini, E., Bombardieri, M. & Mariette, X. Lymphomas complicating primary Sjögren's syndrome: from autoimmunity to lymphoma. *Rheumatology* (Mar. 2019).
6. Narváez, J., Sánchez-Fernández, S. Á., Seoane-Mato, D., Dıaz-González, F. & Bustabad, S. Prevalence of Sjögren's syndrome in the general adult population in Spain: estimating the proportion of undiagnosed cases. *Scientific Reports* **10** (June 2020).
7. Mavragani, C. P. & Moutsopoulos, H. M. The geoepidemiology of Sjögren's syndrome. *Autoimmunity Reviews* **9,** A305–A310 (Mar. 2010).
8. Anagnostopoulos, I. *et al.* The prevalence of rheumatic diseases in central Greece: a population survey. *BMC Musculoskeletal Disorders* **11** (May 2010).
9. Maldini, C. *et al.* Epidemiology of Primary Sjögren's Syndrome in a French Multiracial/Multiethnic Area. *Arthritis Care &amp Research* **66,** 454–463 (Feb. 2014).
10. Vivino, F. B. Sjogren's syndrome: Clinical aspects. *Clinical Immunology* **182,** 48–54 (Sept. 2017).
11. Qin, B. *et al.* Epidemiology of primary Sjögren's syndrome: a systematic review and meta-analysis. *Annals of the Rheumatic Diseases* **74,** 1983–1989 (June 2014).
12. Barturen, G., Beretta, L., Cervera, R., Vollenhoven, R. V. & Alarcón-Riquelme, M. E. Moving towards a molecular taxonomy of autoimmune rheumatic diseases. *Nature Reviews Rheumatology* **14,** 75–93 (Jan. 2018).
13. Barturen, G. *et al.* Integrative Analysis Reveals a Molecular Stratification of Systemic Autoimmune Diseases. *Arthritis &amp Rheumatology* **73,** 1073–1085 (Apr. 2021).
14. Tarn, J. R. *et al.* Symptom-based stratification of patients with primary Sjögren's syndrome: multi-dimensional characterisation of international observational cohorts and reanalyses of randomised clinical trials. *The Lancet Rheumatology* **1,** e85–e94 (Oct. 2019).
15. Soret, P. *et al.* A new molecular classification to drive precision treatment strategies in primary Sjögren's syndrome. *Nature Communications* **12** (June 2021).
16. Trutschel, D. *et al.* Variability of Primary Sjögren's Syndrome Is Driven by Interferon-α and Interferon-α Blood Levels Are Associated With the Class II HLA–DQ Locus. *Arthritis &amp Rheumatology* **74,** 1991–2002 (Nov. 2022).
17. Fu, X., Liu, H., Huang, G. & Dai, S.-S. The emerging role of neutrophils in autoimmune-associated disorders: effector, predictor, and therapeutic targets. *MedComm* **2,** 402–413 (July 2021).
18. Negrini, S. *et al.* Sjögren's syndrome: a systemic autoimmune disease. *Clinical and Experimental Medicine* **22,** 9–25 (June 2021).

19. Bombardieri, M. *et al.* One year in review 2020: pathogenesis of primary Sjögren's syndrome. *Clinical and experimental rheumatology* **38 Suppl 126,** 3–9. ISSN: 0392-856X (4 **Jul-Aug** 2020). ppublish.
20. Saraux, A., Pers, J.-O. & Devauchelle-Pensec, V. Treatment of primary Sjögren syndrome. *Nature Reviews Rheumatology* **12,** 456–471 (July 2016).
21. Ritter, J., Chen, Y., Stefanski, A.-L. & Dörner, T. Current and future treatment in primary Sjögren's syndrome – A still challenging development. *Joint Bone Spine* **89,** 105406 (Nov. 2022).
22. Gottenberg, J.-E. *et al.* Serum Levels of Beta2-Microglobulin and Free Light Chains of Immunoglobulins Are Associated with Systemic Disease Activity in Primary Sjögren's Syndrome. Data at Enrollment in the Prospective ASSESS Cohort. *PLoS ONE* **8** (ed Re, V. D.) e59868 (May 2013).
23. Ng, W.-F., Bowman, S. J. & and, B. G. United Kingdom Primary Sjogren's Syndrome Registry–a united effort to tackle an orphan rheumatic disease. *Rheumatology* **50,** 32–39 (Aug. 2010).
24. Tasaki, S. *et al.* Multiomic disease signatures converge to cytotoxic CD8 T cells in primary Sjögren's syndrome. *Annals of the Rheumatic Diseases* **76,** 1458–1466 (May 2017).
25. Wang, B. *et al.* Similarity network fusion for aggregating data types on a genomic scale. *Nature Methods* **11,** 333–337 (Jan. 2014).
26. Blondel, V. D., Guillaume, J.-L., Lambiotte, R. & Lefebvre, E. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment* **2008,** P10008 (Oct. 2008).
27. Ashburner, M. *et al.* Gene Ontology: tool for the unification of biology. *Nature Genetics* **25,** 25–29 (May 2000).
28. Altman, M. C. *et al.* Development of a fixed module repertoire for the analysis and interpretation of blood transcriptome data. *Nature Communications* **12** (July 2021).
29. Becht, E. *et al.* Estimating the population abundance of tissue-infiltrating immune and stromal cell populations using gene expression. *Genome Biology* **17** (Oct. 2016).
30. Castro-Alcaraz, S., Miskolci, V., Kalasapudi, B., Davidson, D. & Vancurova, I. NF-κB Regulation in Human Neutrophils by Nuclear IκBα: Correlation to Apoptosis. *The Journal of Immunology* **169,** 3947–3953 (Oct. 2002).
31. Tirosh, I. *et al.* Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science* **352,** 189–196 (Apr. 2016).
32. Seror, R. *et al.* EULAR Sjögren's syndrome disease activity index: development of a consensus systemic disease activity index for primary Sjögren's syndrome. *Annals of the Rheumatic Diseases* **69,** 1103–1109 (June 2009).
33. Seror, R. *et al.* EULAR Sjögren's Syndrome Patient Reported Index (ESSPRI): development of a consensus patient index for primary Sjögren's syndrome. *Annals of the Rheumatic Diseases* **70,** 968–972 (Feb. 2011).
34. Devauchelle-Pensec, V. *et al.* Treatment of Primary Sjögren Syndrome With Rituximab. *Annals of Internal Medicine* **160,** 233–242 (Feb. 2014).
35. Bowman, S. J. *et al.* Randomized Controlled Trial of Rituximab and Cost-Effectiveness Analysis in Treating Fatigue and Oral Dryness in Primary Sjögren's Syndrome. *Arthritis &amp Rheumatology* **69,** 1440–1450 (June 2017).
36. Ship, J. A. *et al.* Treatment of Primary Sjogren's Syndrome with Low-Dose Natural Human Interferon-alpha Administered by the Oral Mucosal Route: A Phase II Clinical Trial. *Journal of Interferon &amp Cytokine Research* **19,** 943–951 (Aug. 1999).
37. Zandbelt, M. M. *et al.* Etanercept in the treatment of patients with primary Sjögren's syndrome: a pilot study. *The Journal of rheumatology* **31,** 96–101. ISSN: 0315-162X (1 Jan. 2004). ppublish.
38. Van der Heijden, E. H. M. *et al.* Leflunomide–hydroxychloroquine combination therapy in patients with primary Sjögren's syndrome (RepurpSS-I): a placebo-controlled, double-blinded, randomised clinical trial. *The Lancet Rheumatology* **2,** e260–e269 (May 2020).
39. Seror, R. *et al.* Development and preliminary validation of the Sjögren's Tool for Assessing Response (STAR): a consensual composite score for assessing treatment effect in primary Sjögren's syndrome. *Annals of the Rheumatic Diseases* **81,** 979–989 (Apr. 2022).

40. Finotello, F. & Trajanoski, Z. Quantifying tumor-infiltrating immune cells from transcriptomics data. *Cancer Immunology, Immunotherapy* **67,** 1031–1040 (Mar. 2018).
41. Sturm, G. *et al.* Comprehensive evaluation of transcriptome-based cell-type quantification methods for immuno-oncology. *Bioinformatics* **35,** i436–i445 (July 2019).
42. Papa, N. D. *et al.* The Role of Interferons in the Pathogenesis of Sjögren's Syndrome and Future Therapeutic Perspectives. *Biomolecules* **11,** 251 (Feb. 2021).
43. Stergiou, I. E., Kapsogeorgou, E. E., Tzioufas, A. G., Voulgarelis, M. & Goules, A. V. Clinical Phenotype and Mechanisms of Leukopenia/Neutropenia in Patients with Primary Sjögren's Syndrome. *Mediterranean Journal of Rheumatology* **33,** 99 (2022).
44. Wen, W. *et al.* Clinical and serologic features of primary Sjögren's syndrome concomitant with autoimmune hemolytic anemia: a large-scale cross-sectional study. *Clinical Rheumatology* **34,** 1877–1884 (Oct. 2015).
45. Breedveld, F. C. Leflunomide: mode of action in the treatment of rheumatoid arthritis. *Annals of the Rheumatic Diseases* **59,** 841–849 (Nov. 2000).
46. Shippey, E. A., Wagler, V. D. & Collamer, A. N. Hydroxychloroquine: An old drug with new relevance. *Cleveland Clinic Journal of Medicine* **85,** 459–467 (June 2018).
47. Kužnik, A. *et al.* Mechanism of Endosomal TLR Inhibition by Antimalarial Drugs and Imidazoquinolines. *The Journal of Immunology* **186,** 4794–4804 (Apr. 2011).
48. Erkan, D. *et al.* 14th International Congress on Antiphospholipid Antibodies Task Force Report on Antiphospholipid Syndrome Treatment Trends. *Autoimmunity Reviews* **13,** 685–696 (June 2014).

Figures

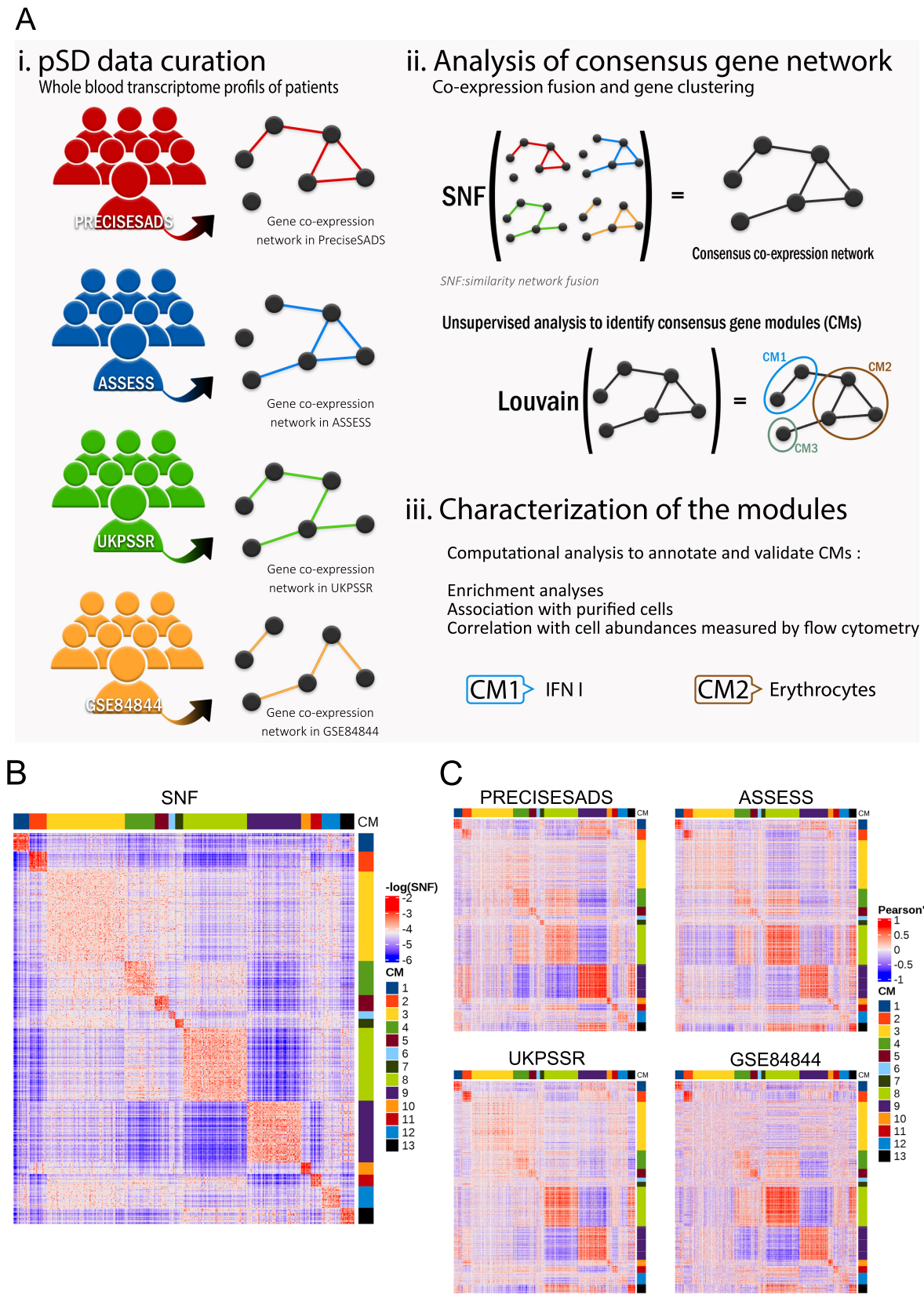


Fig. 1. A) Schematic summary of the work. pSD = primary Sjögren Disease B) Heatmap of the consensus pairwise gene affinity computed by Similarity Network Fusion (SNF). Side annotations represent gene modules. C) Heatmaps of Pearson's correlation matrices of the four input datasets, with genes grouped by their consensus gene modules.

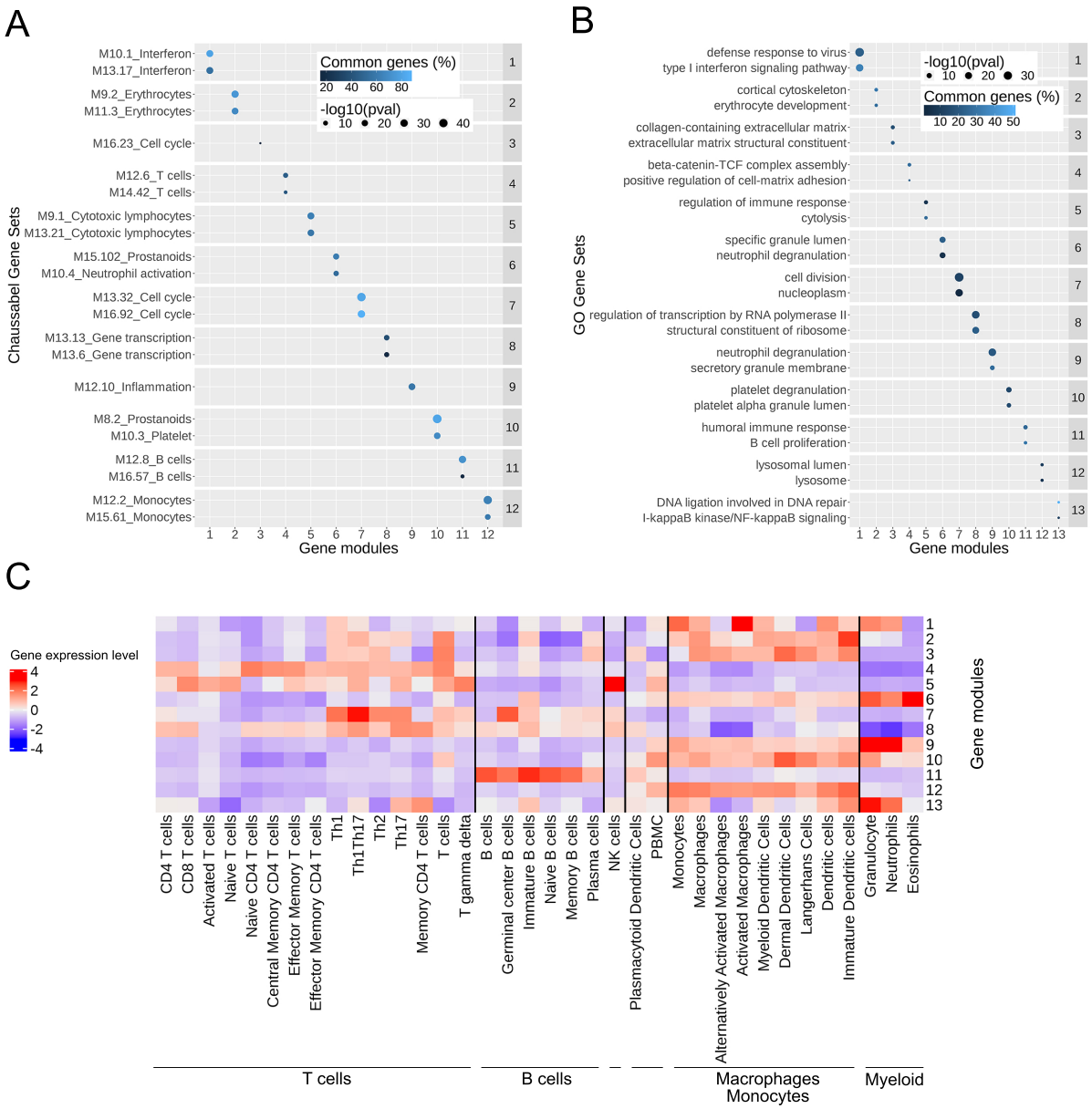


Fig. 2. A) For each module, the two most significantly-enriched pathways in the Chaussabel database[**?**]. B) Most significantly-enriched pathways in the GO database[**?**] C) Average expression of modules in transcriptomes of purified cells

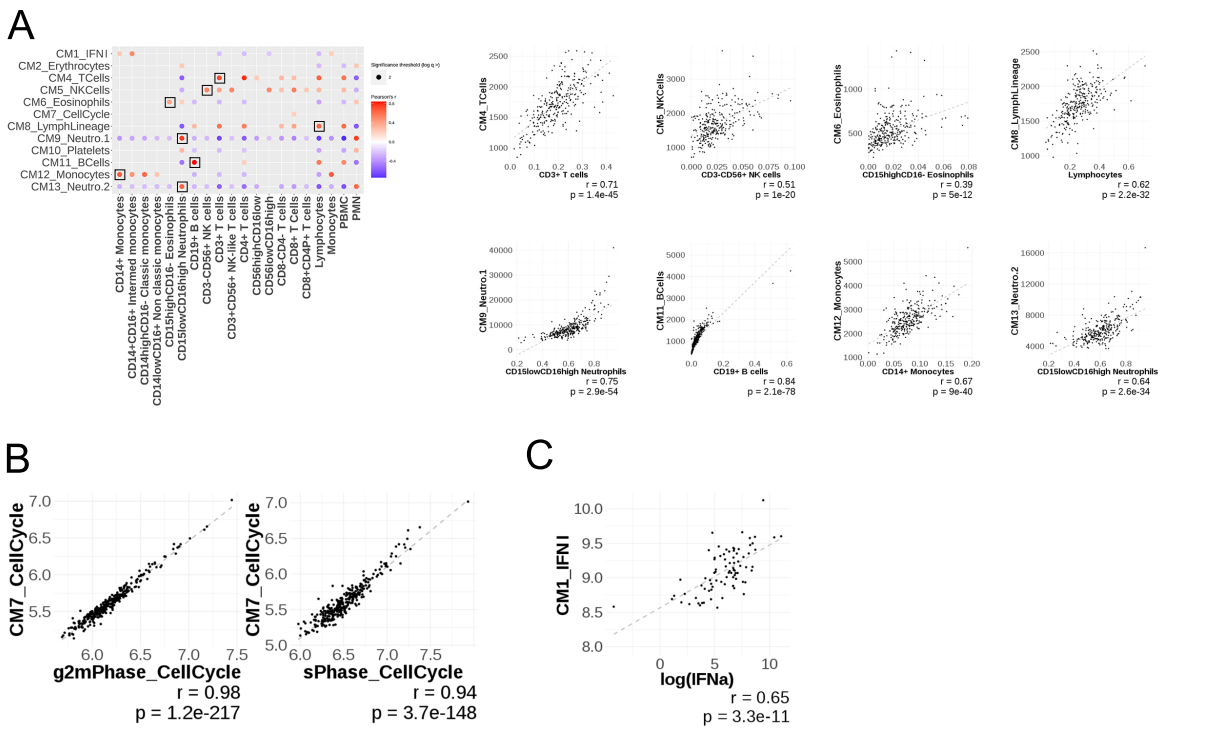


Fig. 3. A) Significant Pearson's correlations between the average expression of the CMs and cell types abundances measured by flow cytometry. Scatter plots of average CMs expression and matching cellular frequencies. B) Scatter plots illustrating the average expression of CM7 versus averages of cell cycle signatures C) Scatter plot of the average expression of CM1 IFN-α and dosage of IFN-α

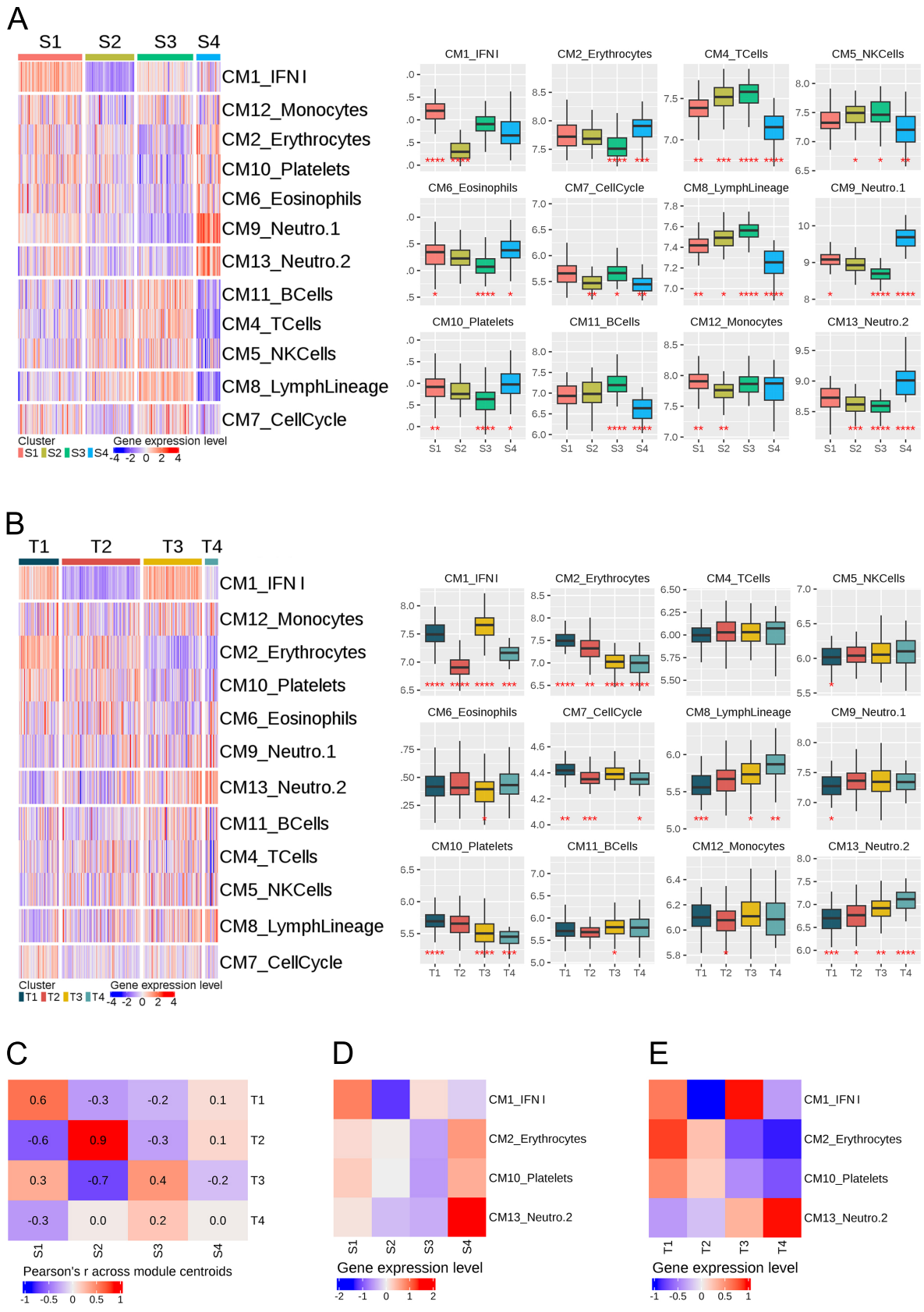


Fig. 4. CMs scores across patient subgroups of A) the Soret classification B) the Trutschel classification and ANOVA tests for each clusters. Average expression of the CM1 IFN-α, CM2 Erythrocytes, CM10 Platelets and CM13 Neutrophils.2 CMs in the C) Soret classification and D) Trutschel classification. E) Correlation across cluster centroids of the two stratification systems.

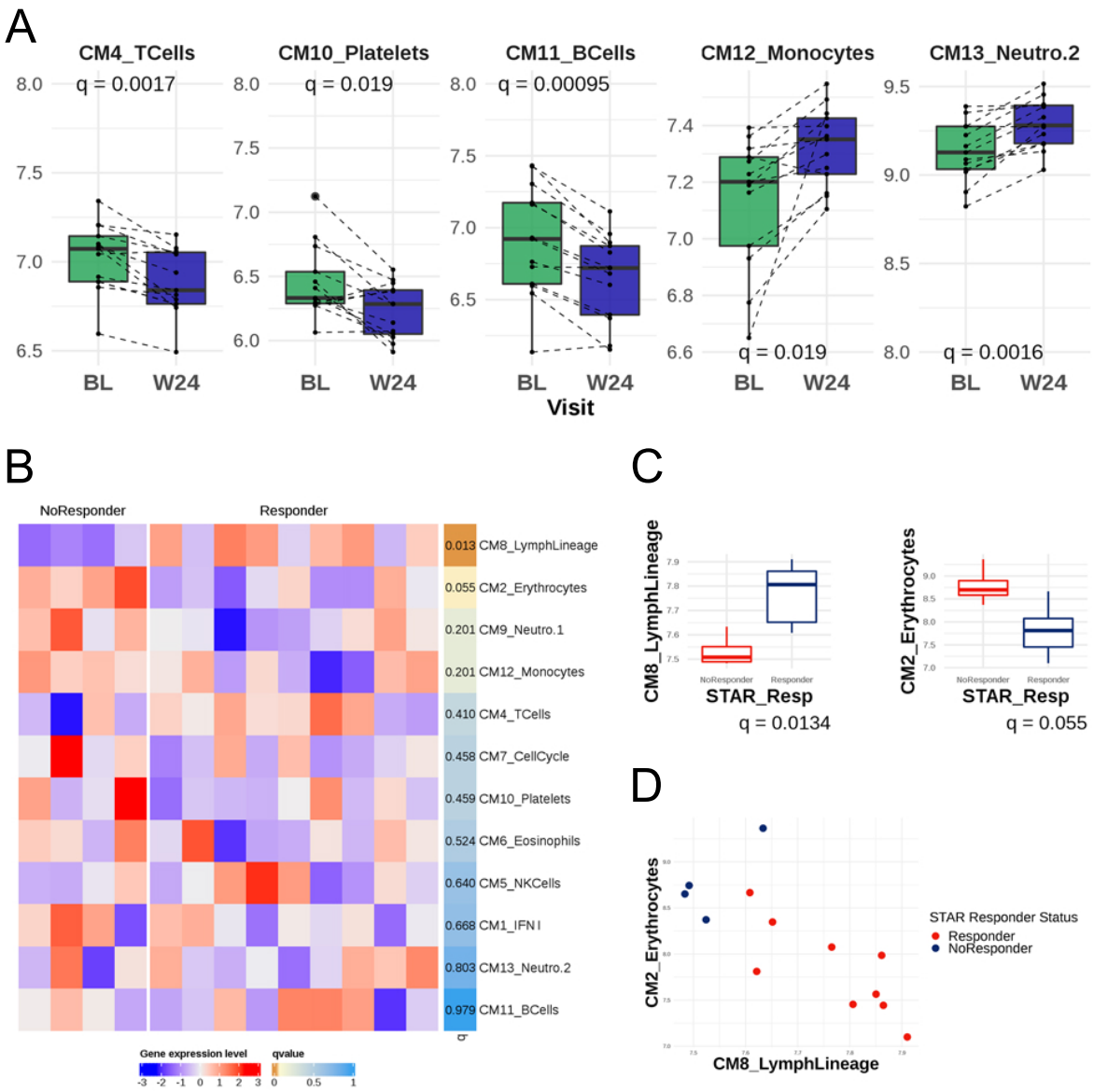


Fig. 5. A) Boxplots illustrating the evolution of the modules significantly differentially-expressed at baseline (BL) versus Week 24 for treated patients B) Heatmap of average gene expression of the CMs. Patients are split by their responder status according to the STAR clinical endpoint C) Avegage expression of CM8 and CM2 at baseline in responders versus non-responders D) Dotplot of average expression of the CM8 and CM2 modules, colored by response statuses.

6 Conclusion

In the ever-evolving landscape of medical research, the past two decades have borne witness to remarkable strides in molecular biology, catalyzed by groundbreaking techniques like next-generation sequencing. A paradigm shift has occurred, steering medicine away from the one-size-fits-all approach. Precision Medicine, a concept premised on the understanding that each individual's condition is distinct, has emerged. Genes and environment intertwine to influence health, symptoms, and treatment outcomes, signifying that tailored therapies are pivotal. An exemple is the Molecular cancer subtypes, coupled with dedicated treatments. This progress has been underscored by the realization of a FDA-approved prognostic signature, spotlighting Precision Medicine's transformative potential. However, as medical data expands exponentially, addressing the challenges posed by its volume, heterogeneity, and richness becomes essential.

In the realm of Precision Medicine, two key methodologies stand out: patient stratification and prediction of treatment responders. My PhD project resided within this dynamic context, with a focal point on Primary Sjögren's Disease (pSD). PSD, an autoimmune disorder, affects moisture-producing glands, leading to dryness of mouth and eyes, alongside other debilitating symptoms. Variability in both biological and clinical aspects among patients has posed substantial hurdles, leaving pSD patients without a definitive cure. Amidst this challenge, the IMI2 NECESSITY project presents a collaborative platform where academia and industry converge to deepen our comprehension of the ailment and unravel novel clinical and molecular markers for potential trials. Guided by this collective effort, the project undertakes dual trajectories. Firstly, the study capitalized on diverse stratification attempts of pSD patients (c.f. Tarn et al, Soret et al., and Trutschel et al.), hailing from independent cohorts and encompassing varied omics and clinical data. With a focus on attaining a consensus akin to successful endeavors in colorectal cancers, this approach navigates the complexity of pSD, seeking to establish robust patient clusters. Employing the intricate toolkit of Systems Immunology, the project delved into deciphering disrupted molecular networks, uncovering immunity's hidden drivers, and unraveling their downstream clinical mannifestations.

Secondly, the research endeavors to identify treatment response factors through historical clinical trial data within the IMI2 NECESSITY consortium. By pinpointing responder patients and refining clinical endpoints, this

initiative would contribute to bolstering sensitivity in future trials. The access to unique datasets and direct interaction with key stakeholders in the IMI 2 NECESSITY consortium and Servier internal initiatives underscores the project's interdisciplinary and collaborative nature.

This PhD ventured not only addresses the complexity of pSD but also resonates with the broader mission of advancing Precision Medicine, where data-driven insights pave the way for individualized care paradigms. As the interdisciplinary efforts converge, the potential for innovative therapeutic strategies in the field of autoimmune disorders becomes palpable. What lies ahead for primary Sjögren's disease research seems promising, with ample possibilities awaiting exploration through cutting-edge tools like scRNA-seq and/or spatial transcriptomics. Moreover, delving into more tissue-specific data holds the potential to significantly enhance our comprehension of action mechanisms at various cellular levels, potentially paving the way for novel therapeutic strategies in the course of managing Sjögren.

Simultaneously with these findings, I wish to emphasize my development of multiple computational tools. While I have not authored any standalone methodological papers, it is worth noting that RNA-seq data, as discussed earlier in this thesis, present complexity, size, and interpretational challenges. Consequently, collaboratively, I have designed and validated specific computational tools and methodologies tailored for the analysis of the data generated across my various projects.

Even though I initially embarked on a PhD in applied mathematics, I found myself delving into the field of immunology. This turned out to be a fortunate turn of events, considering that I intend to leverage my mathematical background to further my career in this captivating realm of systems immunology.

Bibliography

References

1. Murphy, K. M. & Weaver, C. *Janeway's Immunobiology* 924. ISBN: 9780815345053 (W. W. Norton & Company).
2. Rosales, C. Neutrophil: A Cell with Many Roles in Inflammation or Several Cell Types? *Frontiers in physiology* **9,** 113. ISSN: 1664-042X (2018). epublish.
3. Borregaard, N., Sørensen, O. E. & Theilgaard-Mönch, K. Neutrophil granules: a library of innate immunity proteins. *Trends in immunology* **28,** 340–345. ISSN: 1471-4906 (8 Aug. 2007). ppublish.
4. Guilliams, M., Mildner, A. & Yona, S. Developmental and Functional Heterogeneity of Monocytes. *Immunity* **49,** 595–613 (Oct. 2018).
5. Wynn, T. A., Chawla, A. & Pollard, J. W. Macrophage biology in development, homeostasis and disease. *Nature* **496,** 445–455. ISSN: 1476-4687 (7446 Apr. 2013). ppublish.
6. Cabeza-Cabrerizo, M., Cardoso, A., Minutti, C. M., da Costa, M. P. & e Sousa, C. R. Dendritic Cells Revisited. *Annual Review of Immunology* **39,** 131–166 (Apr. 2021).
7. Park, J.-E. *et al.* A cell atlas of human thymic development defines T cell repertoire formation. *Science* **367** (Feb. 2020).
8. Basso, K. & Dalla-Favera, R. Germinal centres and B cell lymphomagenesis. *Nature Reviews Immunology* **15,** 172–184 (Feb. 2015).
9. Spits, H. *et al.* Innate lymphoid cells — a proposal for uniform nomenclature. *Nature Reviews Immunology* **13,** 145–149 (Jan. 2013).
10. Uciechowski, P. & Dempke, W. C. M. Interleukin-6: A Masterplayer in the Cytokine Network. *Oncology* **98,** 131–137. ISSN: 1423-0232 (3 2020). ppublish.
11. Winer, H. *et al.* IL-7: Comprehensive review. *Cytokine* **160,** 156049. ISSN: 1096-0023 (Dec. 2022). ppublish.
12. Griffith, J. W., Sokol, C. L. & Luster, A. D. Chemokines and chemokine receptors: positioning cells for host defense and immunity. *Annual review of immunology* **32,** 659–702. ISSN: 1545-3278 (2014). ppublish.

13. Mollica Poeta, V., Massara, M., Capucetti, A. & Bonecchi, R. Chemokines and Chemokine Receptors: New Targets for Cancer Immunotherapy. *Frontiers in immunology* **10,** 379. ISSN: 1664-3224 (2019). epublish.

14. Liu, Y.-J. IPC: professional type 1 interferon-producing cells and plasmacytoid dendritic cell precursors. *Annual review of immunology* **23,** 275–306. ISSN: 0732-0582 (2005). ppublish.

15. Bondet, V. *et al.* Differential levels of IFNα subtypes in autoimmunity and viral infection. *Cytokine* **144,** 155533. ISSN: 1096-0023 (Aug. 2021). ppublish.

16. Siegal, F. P. *et al.* The nature of the principal type 1 interferon-producing cells in human blood. *Science (New York, N.Y.)* **284,** 1835–1837. ISSN: 0036-8075 (5421 June 1999). ppublish.

17. Kalie, E., Jaitin, D. A., Podoplelova, Y., Piehler, J. & Schreiber, G. The stability of the ternary interferon-receptor complex rather than the affinity to the individual subunits dictates differential biological activities. *The Journal of biological chemistry* **283,** 32925–32936. ISSN: 0021-9258 (47 Nov. 2008). ppublish.

18. Platanias, L. C. Mechanisms of type-I- and type-II-interferon-mediated signalling. *Nature reviews. Immunology* **5,** 375–386. ISSN: 1474-1733 (5 May 2005). ppublish.

19. Akiyama, T., Shinzawa, M., Qin, J. & Akiyama, N. Regulations of gene expression in medullary thymic epithelial cells required for preventing the onset of autoimmune diseases. *Frontiers in immunology* **4,** 249. ISSN: 1664-3224 (2013). epublish.

20. Bonaguro, L. *et al.* A guide to systems-level immunomics. *Nature Immunology* **23,** 1412–1423 (Sept. 2022).

21. Schultze, J. L. Teaching 'big data' analysis to young immunologists. *Nature immunology* **16,** 902–905. ISSN: 1529-2916 (9 Sept. 2015). ppublish.

22. Yanai, I. & Lercher, M. A hypothesis is a liability. *Genome biology* **21,** 231. ISSN: 1474-760X (1 Sept. 2020). epublish.

23. Rose, N. R. & Mackay, I. R. *The Autoimmune Diseases* 1304. ISBN: 9780123849298 (Academic Press).

24. Goodnow, C. C., Sprent, J., Fazekas de St Groth, B. & Vinuesa, C. G. Cellular and genetic mechanisms of self tolerance and autoimmunity. *Nature* **435,** 590–597. ISSN: 1476-4687 (7042 June 2005). ppublish.

25. Kriegel, M. A., Manson, J. E. & Costenbader, K. H. Does vitamin D affect risk of developing autoimmune disease?: a systematic review. *Seminars in arthritis and rheumatism* **40,** 512–531.e8. ISSN: 1532-866X (6 June 2011). ppublish.

26. Leray, E., Moreau, T., Fromont, A. & Edan, G. Epidemiology of multiple sclerosis. *Revue neurologique* **172,** 3–13. ISSN: 0035-3787 (1 Jan. 2016). ppublish.

27. Okada, H., Kuhn, C., Feillet, H. & Bach, J.-F. The 'hygiene hypothesis' for autoimmune and allergic diseases: an update. *Clinical and experimental immunology* **160,** 1–9. ISSN: 1365-2249 (1 Apr. 2010). ppublish.

28. Qiu, C. C., Caricchio, R. & Gallucci, S. Triggers of Autoimmunity: The Role of Bacterial Infections in the Extracellular Exposure of Lupus Nuclear Autoantigens. *Frontiers in immunology* **10,** 2608. ISSN: 1664-3224 (2019). epublish.

29. Atkins, C. J. & Hamilton, E. B. Procainamide-induced systemic lupus erythematosus. *Proceedings of the Royal Society of Medicine* **62,** 197–198. ISSN: 0035-9157 (2 Feb. 1969). ppublish.

30. REINHARDT, D. J. & WALDRON, J. M. Lupus erythematosus-like syndrome complicating hydralazine (apresoline) therapy. *Journal of the American Medical Association* **155,** 1491–1492. ISSN: 0002-9955 (17 Aug. 1954). ppublish.

31. Yung, R. L. & Richardson, B. C. Drug-induced lupus. *Rheumatic diseases clinics of North America* **20,** 61–86. ISSN: 0889-857X (1 Feb. 1994). ppublish.

32. Garratty, G. & Petz, L. D. Drug-induced immune hemolytic anemia. *The American journal of medicine* **58,** 398–407. ISSN: 0002-9343 (3 Mar. 1975). ppublish.

33. Domen, R. E. An overview of immune hemolytic anemias. *Cleveland Clinic journal of medicine* **65,** 89–99. ISSN: 0891-1150 (2 Feb. 1998). ppublish.

34. Rowley, B. & Monestier, M. Mechanisms of heavy metal-induced autoimmunity. *Molecular immunology* **42,** 833–838. ISSN: 0161-5890 (7 May 2005). ppublish.

35. Cervera, R., Gershwin, M. E. & Shoenfeld, Y. *Diagnostic criteria in autoimmune diseases* 593. ISBN: 9781603274272 (Humana Press, 2008).

36. Ralli, M. *et al.* Hashimoto's thyroiditis: An update on pathogenic mechanisms, diagnostic protocols, therapeutic strategies, and potential malignant transformation. *Autoimmunity reviews* **19,** 102649. ISSN: 1873-0183 (10 Oct. 2020). ppublish.

37. Barturen, G., Beretta, L., Cervera, R., Van Vollenhoven, R. & Alarcón-Riquelme, M. E. Moving towards a molecular taxonomy of autoimmune rheumatic diseases. *Nature reviews. Rheumatology* **14,** 180. ISSN: 1759-4804 (3 Feb. 2018). ppublish.

38. Barturen, G. *et al.* Integrative Analysis Reveals a Molecular Stratification of Systemic Autoimmune Diseases. *Arthritis & rheumatology (Hoboken, N.J.)* **73,** 1073–1085. ISSN: 2326-5205 (6 June 2021). ppublish.

39. McInnes, I. B. & Schett, G. The pathogenesis of rheumatoid arthritis. *The New England journal of medicine* **365,** 2205–2219. ISSN: 1533-4406 (23 Dec. 2011). ppublish.

40. Gibofsky, A. Epidemiology, pathophysiology, and diagnosis of rheumatoid arthritis: A Synopsis. *The American journal of managed care* **20,** S128–S135. ISSN: 1936-2692 (7 Suppl May 2014). ppublish.

41. Dobson, R. & Giovannoni, G. Multiple sclerosis - a review. *European journal of neurology* **26,** 27–40. ISSN: 1468-1331 (1 Jan. 2019). ppublish.

42. Griffiths, C. E. M., Armstrong, A. W., Gudjonsson, J. E. & Barker, J. N. W. N. Psoriasis. *Lancet (London, England)* **397,** 1301–1315. ISSN: 1474-547X (10281 Apr. 2021). ppublish.

43. Guan, Q. A Comprehensive Review and Update on the Pathogenesis of Inflammatory Bowel Disease. *Journal of immunology research* **2019,** 7247238. ISSN: 2314-7156 (2019). epublish.

44. Cutolo, M., Soldano, S. & Smith, V. Pathophysiology of systemic sclerosis: current understanding and new insights. *Expert review of clinical immunology* **15,** 753–764. ISSN: 1744-8409 (7 July 2019). ppublish.

45. Tsokos, G. C. Autoimmunity and organ damage in systemic lupus erythematosus. *Nature Immunology* **21,** 605–614 (May 2020).

46. Kaul, A. *et al.* Systemic lupus erythematosus. *Nature reviews. Disease primers* **2,** 16039. ISSN: 2056-676X (June 2016). epublish.

47. Mariette, X. & Criswell, L. A. Primary Sjögren's Syndrome. *The New England journal of medicine* **378,** 931–939. ISSN: 1533-4406 (10 Mar. 2018). ppublish.

48. Brito-Zerón, P. *et al.* Sjögren syndrome. *Nature reviews. Disease primers* **2,** 16047. ISSN: 2056-676X (July 2016). epublish.

49. Parisis, D., Chivasso, C., Perret, J., Soyfoo, M. S. & Delporte, C. Current State of Knowledge on Primary Sjögren's Syndrome, an Autoimmune Exocrinopathy. *Journal of Clinical Medicine* **9,** 2299 (July 2020).

50. Solans-Laqué, R. *et al.* Risk, Predictors, and Clinical Characteristics of Lymphoma Development in Primary Sjögren's Syndrome. *Seminars in Arthritis and Rheumatism* **41,** 415–423 (Dec. 2011).

51. Nocturne, G., Pontarini, E., Bombardieri, M. & Mariette, X. Lymphomas complicating primary Sjögren's syndrome: from autoimmunity to lymphoma. *Rheumatology* (Mar. 2019).

52. Narváez, J., Sánchez-Fernández, S. Á., Seoane-Mato, D., Dıaz-González, F. & Bustabad, S. Prevalence of Sjögren's syndrome in the general adult population in Spain: estimating the proportion of undiagnosed cases. *Scientific Reports* **10** (June 2020).

53. Mavragani, C. P. & Moutsopoulos, H. M. The geoepidemiology of Sjögren's syndrome. *Autoimmunity Reviews* **9,** A305–A310 (Mar. 2010).

54. Anagnostopoulos, I. *et al.* The prevalence of rheumatic diseases in central Greece: a population survey. *BMC Musculoskeletal Disorders* **11** (May 2010).

55. Maldini, C. *et al.* Epidemiology of Primary Sjögren's Syndrome in a French Multiracial/Multiethnic Area. *Arthritis Care &amp Research* **66,** 454–463 (Feb. 2014).

56. Vivino, F. B. Sjogren's syndrome: Clinical aspects. *Clinical Immunology* **182,** 48–54 (Sept. 2017).

57. Qin, B. *et al.* Epidemiology of primary Sjögren's syndrome: a systematic review and meta-analysis. *Annals of the Rheumatic Diseases* **74,** 1983–1989 (June 2014).

58. Lucchesi, D., Pitzalis, C. & Bombardieri, M. EBV and other viruses as triggers of tertiary lymphoid structures in primary Sjögren's syndrome. *Expert review of clinical immunology* **10,** 445–455. ISSN: 1744-8409 (4 Apr. 2014). ppublish.

59. Bombardieri, M. *et al.* One year in review 2020: pathogenesis of primary Sjögren's syndrome. *Clinical and experimental rheumatology* **38 Suppl 126,** 3–9. ISSN: 0392-856X (4 2020). ppublish.

60. Trutschel, D. *et al.* Variability of Primary Sjögren's Syndrome Is Driven by Interferon-α and Interferon-α Blood Levels Are Associated With the Class II HLA–DQ Locus. *Arthritis &amp Rheumatology* **74,** 1991–2002 (Nov. 2022).

61. Soret, P. *et al.* A new molecular classification to drive precision treatment strategies in primary Sjögren's syndrome. *Nature Communications* **12** (June 2021).

62. Shimizu, T., Nakamura, H. & Kawakami, A. Role of the Innate Immunity Signaling Pathway in the Pathogenesis of Sjögren's Syndrome. *International journal of molecular sciences* **22.** ISSN: 1422-0067 (6 Mar. 2021). epublish.

63. Hillen, M. R. *et al.* Dysregulated miRNome of plasmacytoid dendritic cells from patients with Sjögren's syndrome is associated with processes at the centre of their function. *Rheumatology (Oxford, England)* **58,** 2305–2314. ISSN: 1462-0332 (12 Dec. 2019). ppublish.

64. Nocturne, G. & Mariette, X. Interferon signature in systemic autoimmune diseases: what does it mean? *RMD open* **8.** ISSN: 2056-5933 (2 Dec. 2022). ppublish.

65. Emamian, E. S. *et al.* Peripheral blood gene expression profiling in Sjögren's syndrome. *Genes and immunity* **10,** 285–296. ISSN: 1476-5470 (4 June 2009). ppublish.

66. Brkic, Z. *et al.* Prevalence of interferon type I signature in CD14 monocytes of patients with Sjogren's syndrome and association with disease activity and BAFF gene expression. *Annals of the rheumatic diseases* **72,** 728–735. ISSN: 1468-2060 (5 May 2013). ppublish.

67. Nocturne, G. & Mariette, X. Advances in understanding the pathogenesis of primary Sjögren's syndrome. *Nature reviews. Rheumatology* **9,** 544–556. ISSN: 1759-4804 (9 Sept. 2013). ppublish.

68. Negrini, S. *et al.* Sjögren's syndrome: a systemic autoimmune disease. *Clinical and experimental medicine* **22,** 9–25. ISSN: 1591-9528 (1 Feb. 2022). ppublish.

69. Tarn, J. R. *et al.* Symptom-based stratification of patients with primary Sjögren's syndrome: multi-dimensional characterisation of international observational cohorts and reanalyses of randomised clinical trials. *The Lancet Rheumatology* **1,** e85–e94 (Oct. 2019).

70. Sperber, K. *et al.* Selective regulation of cytokine secretion by hydroxychloroquine: inhibition of interleukin 1 alpha (IL-1-alpha) and IL-6 in human monocytes and T cells. *The Journal of rheumatology* **20,** 803–808. ISSN: 0315-162X (5 May 1993). ppublish.

71. Schrezenmeier, E. & Dörner, T. Mechanisms of action of hydroxychloroquine and chloroquine: implications for rheumatology. *Nature reviews. Rheumatology* **16,** 155–166. ISSN: 1759-4804 (3 Mar. 2020). ppublish.

72. Gottenberg, J.-E. *et al.* Effects of hydroxychloroquine on symptomatic improvement in primary Sjögren syndrome: the JOQUER randomized clinical trial. *JAMA* **312,** 249–258. ISSN: 1538-3598 (3 July 2014). ppublish.

73. Bodewes, I. L. A., Gottenberg, J.-E., van Helden-Meeuwsen, C. G., Mariette, X. & Versnel, M. A. Hydroxychloroquine treatment downregulates systemic interferon activation in primary Sjögren's syndrome in the JOQUER randomized trial. *Rheumatology (Oxford, England)* **59,** 107–111. ISSN: 1462-0332 (1 Jan. 2020). ppublish.

74. Siemasko, K. F., Chong, A. S., Williams, J. W., Bremer, E. G. & Finnegan, A. Regulation of B cell function by the immunosuppressive agent leflunomide. *Transplantation* **61,** 635–642. ISSN: 0041-1337 (4 Feb. 1996). ppublish.

75. Van Woerkom, J. M. *et al.* Safety and efficacy of leflunomide in primary Sjögren's syndrome: a phase II pilot study. *Annals of the rheumatic diseases* **66,** 1026–1032. ISSN: 0003-4967 (8 Aug. 2007). ppublish.

76. Isenberg, D. A. B cell targeted therapies in autoimmune diseases. *The Journal of rheumatology. Supplement* **77,** 24–28. ISSN: 0380-0903 (May 2006). ppublish.

77. Devauchelle-Pensec, V. *et al.* Treatment of primary Sjögren syndrome with rituximab: a randomized trial. *Annals of internal medicine* **160,** 233–242. ISSN: 1539-3704 (4 Feb. 2014). ppublish.

78. Bowman, S. J. *et al.* Randomized Controlled Trial of Rituximab and Cost-Effectiveness Analysis in Treating Fatigue and Oral Dryness in Primary Sjögren's Syndrome. *Arthritis & rheumatology (Hoboken, N.J.)* **69,** 1440–1450. ISSN: 2326-5205 (7 July 2017). ppublish.

79. Cornec, D. *et al.* Blood and salivary-gland BAFF-driven B-cell hyperactivity is associated to rituximab inefficacy in primary Sjögren's syndrome. *Journal of autoimmunity* **67,** 102–110. ISSN: 1095-9157 (Feb. 2016). ppublish.

80. Desvaux, E. *et al.* Model-based computational precision medicine to develop combination therapies for autoimmune diseases. *Expert review of clinical immunology* **18,** 47–56. ISSN: 1744-8409 (1 Jan. 2022). ppublish.

81. Moingeon, P., Kuenemann, M. & Guedj, M. Artificial intelligence-enhanced drug design and development: Toward a computational precision medicine. *Drug discovery today* **27,** 215–222. ISSN: 1878-5832 (1 Jan. 2022). ppublish.

82. Lee, L. Y.-H. & Loscalzo, J. Network Medicine in Pathobiology. *The American journal of pathology* **189,** 1311–1326. ISSN: 1525-2191 (7 July 2019). ppublish.

83. Vamathevan, J. *et al.* Applications of machine learning in drug discovery and development. *Nature reviews. Drug discovery* **18,** 463–477. ISSN: 1474-1784 (6 June 2019). ppublish.

84. Pitzalis, C., Choy, E. H. S. & Buch, M. H. Transforming clinical trials in rheumatology: towards patient-centric precision medicine. *Nature reviews. Rheumatology* **16,** 590–599. ISSN: 1759-4804 (10 Oct. 2020). ppublish.

85. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome biology* **11,** R106. ISSN: 1474-760X (10 2010). ppublish.

86. Ritchie, M. E. *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic acids research* **43,** e47. ISSN: 1362-4962 (7 Apr. 2015). ppublish.

87. Gold, D. L., Coombes, K. R., Wang, J. & Mallick, B. Enrichment analysis in high-throughput genomics - accounting for dependency in the NULL. *Briefings in bioinformatics* **8,** 71–77. ISSN: 1467-5463 (2 Mar. 2007). ppublish.

88. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America* **102,** 15545–15550. ISSN: 0027-8424 (43 Oct. 2005). ppublish.

89. Consortium, G. O. Gene Ontology Consortium: going forward. *Nucleic acids research* **43,** D1049–D1056. ISSN: 1362-4962 (Database issue Jan. 2015). ppublish.

90. Kanehisa, M. *et al.* Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic acids research* **42,** D199–D205. ISSN: 1362-4962 (Database issue Jan. 2014). ppublish.

91. Liberzon, A. *et al.* Molecular signatures database (MSigDB) 3.0. *Bioinformatics (Oxford, England)* **27,** 1739–1740. ISSN: 1367-4811 (12 June 2011). ppublish.

92. Geistlinger, L. *et al.* Toward a gold standard for benchmarking gene set enrichment analysis. *Briefings in bioinformatics* **22,** 545–556. ISSN: 1477-4054 (1 Jan. 2021). ppublish.

93. Cantini, L., Medico, E., Fortunato, S. & Caselle, M. Detection of gene communities in multi-networks reveals cancer drivers. *Scientific reports* **5,** 17386. ISSN: 2045-2322 (Dec. 2015). epublish.

94. Olivier, M., Asmis, R., Hawkins, G. A., Howard, T. D. & Cox, L. A. The Need for Multi-Omics Biomarker Signatures in Precision Medicine. *International journal of molecular sciences* **20.** ISSN: 1422-0067 (19 Sept. 2019). epublish.

95. Boolell, M. *et al.* Sildenafil: an orally active type 5 cyclic GMP-specific phosphodiesterase inhibitor for the treatment of penile erectile dysfunction. *International journal of impotence research* **8,** 47–52. ISSN: 0955-9930 (2 June 1996). ppublish.

96. Moehler, T. M., Hillengass, J., Glasmacher, A. & Goldschmidt, H. Thalidomide in multiple myeloma. *Current pharmaceutical biotechnology* **7,** 431–440. ISSN: 1873-4316 (6 Dec. 2006). ppublish.

97. Harris, N. L. *et al.* A revised European-American classification of lymphoid neoplasms: a proposal from the International Lymphoma Study Group. *Blood* **84,** 1361–1392. ISSN: 0006-4971 (5 Sept. 1994). ppublish.

98. Alizadeh, A. A. *et al.* Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* **403,** 503–511. ISSN: 0028-0836 (6769 Feb. 2000). ppublish.

99. Guinney, J. *et al.* The consensus molecular subtypes of colorectal cancer. *Nature Medicine* **21,** 1350–1356 (Oct. 2015).

100. Becht, E. *et al.* Immune and Stromal Classification of Colorectal Cancer Is Associated with Molecular Subtypes and Relevant for Precision Immunotherapy. *Clinical Cancer Research* **22,** 4057–4066 (Aug. 2016).

101. Banchereau, R. *et al.* Personalized Immunomonitoring Uncovers Molecular Networks that Stratify Lupus Patients. *Cell* **165,** 551–565. ISSN: 1097-4172 (3 Apr. 2016). ppublish.

102. Petri, M. *et al.* Association between changes in gene signatures expression and disease activity among patients with systemic lupus erythematosus. *BMC medical genomics* **12,** 4. ISSN: 1755-8794 (1 Jan. 2019). epublish.

103. Chaussabel, D. & Baldwin, N. Democratizing systems immunology with modular transcriptional repertoire analyses. *Nature Reviews Immunology* **14,** 271–280 (Mar. 2014).

104. Altman, M. C. *et al.* Development of a fixed module repertoire for the analysis and interpretation of blood transcriptome data. *Nature Communications* **12** (July 2021).

105. Guedj, M. *et al.* A refined molecular taxonomy of breast cancer. *Oncogene* **31,** 1196–1206 (July 2011).

106. Zhang, B. & Horvath, S. A General Framework for Weighted Gene Co-Expression Network Analysis. *Statistical Applications in Genetics and Molecular Biology* **4** (Jan. 2005).

107. Wang, B. *et al.* Similarity network fusion for aggregating data types on a genomic scale. *Nature methods* **11,** 333–337. ISSN: 1548-7105 (3 Mar. 2014). ppublish.

108. Lancichinetti, A. & Fortunato, S. Community detection algorithms: a comparative analysis. *Physical review. E, Statistical, nonlinear, and soft matter physics* **80,** 056117. ISSN: 1550-2376 (5 Pt 2 Nov. 2009). ppublish.

7 Annexes

7.1 Annex 1

A new molecular classification to drive precision treatment strategies in primary Sjögren's syndrome

Perrine Soret, Christelle Le Dantec, Emiko Desvaux, Nathan Foulquier, Bastien Chassagnol, Sandra Hubert, Christophe Jamin, Guillermo Barturen, Guillaume Desachy, Valérie Devauchelle- Pensec, **Cheïma Boudjeniba**, Divi Cornec, Alain Saraux, Sandrine Jousse-Joulin, Nuria Barbarroja, Ignasi Rodríguez-Pinto, Ellen De Langhe, Lorenzo Beretta, Carlo Chizzolini, Laszlo Kovacs, Torsten Witte, PRECISESADS Clinical Consortium, PRECISESADS Flow Cytometry Consortium, Eléonore Bettacchioli, Anne Buttgereit, Zuzanna Makowska, Ralf Lesche, Maria Orietta Borghi, Javier Martin, Sophie Courtade-Gaiani, Laura Xuereb, Mickaël Guedj, Philippe Moingeon, Marta E. Alarcon-Riquelme, Laurence Laigle, Jacques-Olivier Pers.

Nat Commun. 2021 Jun 10;12(1):3523. doi: 10.1038/s41467-021-23472-7.

Summary

While numerous research studies have shed light on the pathophysiological roles of interferons (IFNs), the immunopathological mechanisms responsible for the clinical symptoms of primary Sjögren's disaese (pSD) remain largely unclear. As seen in the introduction, current treatments primarily target symptom management to enhance patients' quality of life, without significantly impacting disease progression. The limited success of various clinical trials can be attributed in part to the molecular diversity of pSD, which is often overlooked in clinical classifications. To address this issue, a collaborative effort within Servier International Research Institute and other academic institutes was initiated. Building on our previous involvement in the PRECISESADS IMI project [37], we harnessed data from a cohort of 304 pSD patients and 330 age and gender matched healthy volunteers. Disease diagnosis was confirmed through the presence of anti SSA/Ro autoantibodies or focal lymphocytic sialadenitis (inflammation of a salivary gland) with a focus score 1.

Each patient contributed multi-omics data encompassing genetics, epigenomics, and transcriptomics, alongside immuno-phenotypic analyses via flow cytometry and cytokine measurements.

We examined the transcriptomic data using a method previously applied to breast cancer [105], involving both supervised and unsupervised steps. Additionally, the patient samples were divided into two sets: one for identification (75% of patients) and the other for validation (remaining 25%). Three different clustering methods were employed to categorize the pSD patients into four distinct groups. A signature comprising 257 genes, identified during the supervised step, effectively segregated the 304 pSD patients into these four groups, which can be further broken down into three modules: M.a (105 genes), M.b (20 genes), and M.c (132 genes). Differentially expressed genes between patient and healthy volunteer groups were subjected to analysis using the Ingenuity Pathway Analysis program. This analysis was complemented by a repertoire from Chaussabel et al.[104] of 382 transcriptomic modules established in blood and further characterized by combining other omics and clinical data.

Through integrative analysis, we identified four groups of patients. Notably, three of these groups (clusters C1, C3, and C4) exhibited significant overexpression of genes associated with the IFN pathway, although they dif-

fered in their enrichment of specific types of IFNs (type I or type II). One cluster (cluster 1) demonstrated an intensified type I and II interferon (IFN) response, linked to a robust autoreactive response. Another cluster (cluster 3), resembling the first, showed IFN pathway activation and strong B cell hyperactivity. A smaller subset of patients (cluster 4) displayed a hyperinflammatory phenotype characterized by increased neutrophil counts, methylation irregularities, and an amplified pro-inflammatory cytokine profile, including type II IFN. Finally, the last cluster (cluster 2) was marked by a pronounced glandular component, without molecular dysfunctions at the peripheral level.

Moreover, we developed a composite predictive model to assign pSD patients to one of the four identified groups using a two-step machine learning approach, achieving an impressive overall accuracy of 95%. Initially, membership in the C4 group was determined based on the expression of 10 genes, followed by an assessment of the expression of 31 genes to allocate patients to the C1, C2, or C3 groups. An interpolation function employing six genes with consistent expression allowed for the application of this algorithm across various pSD transcriptomic databases.

In summary, this study provided a comprehensive characterization and stratification of pSD patients based on their detailed molecular profiles obtained from blood samples. These findings underscore the necessity of deepening our understanding of Sjögren's disease's pathophysiology, including the intricate relationships between molecular abnormalities, disease activity, and the efficacy of different treatment modalities in influencing its progression.

To be more specific, the role I had in this paper involved creating a pipeline for patient clustering. Initially, we performed calculations using various clustering methods and assessed their effectiveness.

nature COMMUNICATIONS

ARTICLE

https://doi.org/10.1038/s41467-021-23472-7 OPEN

A new molecular classification to drive precision treatment strategies in primary Sjögren's syndrome

Perrine Soret[1,29], Christelle Le Dantec[2,29], Emiko Desvaux[1,2], Nathan Foulquier[2], Bastien Chassagnol[1], Sandra Hubert[1], Christophe Jamin[2,3], Guillermo Barturen[4], Guillaume Desachy[1], Valérie Devauchelle-Pensec[2,3], Cheïma Boudjeniba[1], Divi Cornec[2,3], Alain Saraux[2,3], Sandrine Jousse-Joulin[2,3], Nuria Barbarroja[5], Ignasi Rodríguez-Pintó[6], Ellen De Langhe[7], Lorenzo Beretta[8], Carlo Chizzolini[9], László Kovács[10], Torsten Witte[11], PRECISESADS Clinical Consortium*, PRECISESADS Flow Cytometry Consortium*, Eléonore Bettacchioli[3], Anne Buttgereit[12], Zuzanna Makowska[12], Ralf Lesche[12], Maria Orietta Borghi[13], Javier Martin[14], Sophie Courtade-Gaiani[1], Laura Xuereb[1], Mickaël Guedj[1], Philippe Moingeon[1], Marta E. Alarcón-Riquelme[4], Laurence Laigle[1] & Jacques-Olivier Pers[2,3]✉

There is currently no approved treatment for primary Sjögren's syndrome, a disease that primarily affects adult women. The difficulty in developing effective therapies is -in part- because of the heterogeneity in the clinical manifestation and pathophysiology of the disease. Finding common molecular signatures among patient subgroups could improve our understanding of disease etiology, and facilitate the development of targeted therapeutics. Here, we report, in a cross-sectional cohort, a molecular classification scheme for Sjögren's syndrome patients based on the multi-omic profiling of whole blood samples from a European cohort of over 300 patients, and a similar number of age and gender-matched healthy volunteers. Using transcriptomic, genomic, epigenetic, cytokine expression and flow cytometry data, combined with clinical parameters, we identify four groups of patients with distinct patterns of immune dysregulation. The biomarkers we identify can be used by machine learning classifiers to sort future patients into subgroups, allowing the re-evaluation of response to treatments in clinical trials.

[1] Institut de Recherches Internationales Servier, Departments of Translational Medicine and Immuno-Inflammatory Diseases Research and Development, Suresnes, France. [2] LBAI, UMR1227, Univ Brest, Inserm, Brest, France. [3] CHU de Brest, Brest, France. [4] Department of Medical Genomics, Center for Genomics and Oncological Research (GENYO), Granada, Spain. [5] Reina Sofia Hospital, Maimonides Institute for Research in Biomedicine of Cordoba (IMIBIC), University of Cordoba, Cordoba, Spain. [6] Hospital Clinic, Institut d'Investigacions Biomèdiques August Pi i Sunyer, Barcelona, Catalonia, Spain. [7] Skeletal Biology and Engineering Research Center, KU Leuven and Division of Rheumatology, UZ Leuven, Belgium. [8] Scleroderma Unit, Referral Center for Systemic Autoimmune Diseases, Fondazione IRCCS Ca'Granda Ospedale Maggiore Policlinico di Milano, Milan, Italy. [9] Immunology & Allergy, University Hospital and School of Medicine, Geneva, Switzerland. [10] University of Szeged, Szeged, Hungary. [11] Klinik für Immunologie und Rheumatologie, Medical University Hannover, Hannover, Germany. [12] Pharmaceuticals Division, Bayer Pharma Aktiengesellschaft, Berlin, Germany. [13] Università degli studi di Milano, Milan, Italy. [14] Institute of Parasitology and Biomedicine López-Neyra, Consejo Superior de Investigaciones Científicas (IPBLN-CSIC), Granada, Spain. [29] These authors contributed equally: Perrine Soret and Christelle Le Dantec. *Lists of authors and their affiliations appear at the end of the paper. ✉email: pers@univ-brest.fr

Primary Sjögren's syndrome (pSS) is a chronic, disabling, complex systemic autoimmune disease that mostly affects adult women and still lacks a specific therapy. Although the involvement of salivary and lachrymal glands is the hallmark of the disease, during pSS progression, various organs and systems can be involved including joints, lungs, kidneys, liver, nervous and musculoskeletal system[1]. Thus, the clinical spectrum of the disease ranges from a benign slowly progressive autoimmune exocrinopathy to a severe systemic disorder with significant symptom heterogeneity and scattered complications. The diagnosis of pSS is currently based upon a combination of clinical, serological, histological, and functional parameters which are most often only satisfied at a late stage of the disease, i.e., when glandular dysfunction and symptoms already severely affect a patient's overall quality of life. Moreover, one fifth of pSS patients may present major organ involvement with potentially severe end-organ damage[2] and five percent of patients may also develop non-Hodgkin's lymphoma[3]. Primary SS is one of the few prototypic diseases to link autoimmunity, cancer development and infections, offering unique insights in many areas of basic science and clinical medicine. However, the pathogenesis of the disease remains elusive. Specifically, limited knowledge of existing pSS disease variants arguably represents the greatest obstacle to improve patients' diagnosis and identify patients' subsets in view of early stratification and personalized treatment[4]. It was recently shown in the PRECISESADS IMI JU project that systemic autoimmune diseases exhibit a diverse spectrum and a complex nuanced or overlapping molecular phenotype with four clusters identified, representing 'inflammatory', 'lymphoid', 'interferon' and 'healthy-like' patterns each including all diagnoses and defined by genetic, clinical, serological and cellular features[5]. Many of them share susceptibility genes[6] and an overexpression of interferon (IFN) inducible genes known as the IFN signature is observed in many of these patients[7]. Such autoimmune diseases are driven by numerous environmental factors, therefore displaying a marked variability in their natural course as it relates to their initiation, propagation and flares.

The present study was undertaken to establish a precise molecular classification of patients affected by pSS into more homogeneous clusters whatever their disease phenotypes, activity or treatment. We report herein on the integrated molecular profiling of 304 pSS patients compared to 330 matched healthy volunteers (HV) performed using high-throughput multi-omics data collected within the PRECISESADS IMI JU project (genetic, epigenomic, transcriptomic, combined with flow cytometric data, multiplexed cytokines, as well as classical serology and clinical data). We identify 4 groups of patients with distinct patterns of immune dysregulation. The Cluster 1 (C1), C3 and C4 display a high IFN signature reflecting the pathological involvement of the IFN pathway, but with various Type I and II IFN gene enrichment. C1 has the strongest IFN signature with both Type I and Type II gene enrichment when compared to C3 (intermediate) and C4 (lower). C4 has a Type II gene enrichment stronger than Type I and equivalent to C3 while C3 has the opposite composition. C2 exhibits a weak Type I and Type II IFN signature with no other obvious distinguishable profile relative to HV. We further characterized C1, C3 and C4 using multi-omics and clinical data. C1 patients present a high prevalence of SNPs, C3 patients an involvement of B cell component more prominent than in the other clusters and especially an increased frequency of B cells in the blood while C4 patients have an inflammatory signature driven by monocytes and neutrophils, together with an aberrant methylation status. Algorithms derived from machine learning discriminate the 4 clusters based on distinct biomarkers that can be easily used in a composite model to stratify patients in clinical trials. This composite model is validated by using an independent inception cohort of 37 pSS patients. In conclusion, this work provides a clear understanding of pSS heterogeneity providing clinically and immunopathologically relevant signatures to guide precision medicine strategies. Decision trees coming from this patient classification have an immediate application to re-evaluate response to treatments in clinical trials.

Results

Four functional molecular clusters of pSS patients were identified. Our initial study population comprised 382 pSS patients enrolled in the PRECISESADS cross-sectional study. Following complete quality control and diagnosis validation (each patient had to present either anti-SSA/Ro antibody positivity or focal lymphocytic sialadenitis with a focus score of ≥1 foci/mm^2), 78 patients were removed (Supplementary Fig. 1a–c). Patient characteristics are presented in Table 1. To perform the clustering of the remaining 304 samples, transcriptomics data were analyzed with a semi-supervised robust approach previously applied to breast cancer[8] that iterates unsupervised and supervised steps and relies on the concordance between 3 methods of clustering (see Methods). Samples were divided into a discovery set and an independent validation set, representing 75 and 25% of samples, respectively. The discovery set allowed to cluster patients in four groups, as confirmed in the validation set (Fig. 1a). When the two sets were merged, Cluster 1 (C1) contained 101 patients (33.2%), Cluster 2 (C2) 77 patients (25.3%), Cluster 3 (C3) 88 patients (28.9%) and Cluster 4 (C4) 38 patients (12.5%). The supervised step allowed to select a subset of 257 top genes discriminating the 4 clusters of patients (Supplementary Fig. 2) and divided into 3 modules: M.a (105 genes), M.b (20 genes) and M.c (132 genes). An enrichment analysis was used to annotate each gene module, showing that M.a was enriched in IFN signaling, M.b in lymphoid lineage pathways and M.c in inflammatory and myeloid lineage transcripts (Supplementary Fig. 3). C1, and to a lesser extent C3, presented overexpression of gene module M.a, whereas C3 showed overexpression of M.b as well and C4 strong overexpression of M.c (Fig. 1a). Because C2 had no obvious discernible pattern, healthy volunteers (HV) were assigned to the 4 molecular clusters distance to centroids (Fig. 1b). When projected into the patient population, HV did not constitute a separate cluster but mainly matched with C2 (0.5%, 93%, 4% and 2.5% of HV merged with C1, C2, C3, and C4, respectively). This means that the C2 transcriptional signature is not different from HV, at least at the blood level. Interestingly, our data are consistent with the previous observation of a healthy-like patient group detected in a pooled population of 7 different autoimmune diseases[5].

We then assessed whether covariates like systemic treatments could drive the transcriptome-based clustering. Indeed, half of the pSS patients were treated with either anti-malarials, immunosuppressants, or steroids at the time of the visit with a statistically significant difference in the distribution among the four clusters (p-values were respectively 0.002 for anti-malarials, <0.001 for immunosuppressants and steroids) (Table 2). When compared to the 3 other clusters, a higher proportion of patients treated with anti-malarials in C2 and a higher proportion of patients receiving immunosuppressants or steroids in C4 were observed. Importantly, sensitivity analyses of treated versus untreated patients in each cluster showed no impact of treatments on cluster distribution (Supplementary Fig. 4).

In depth functional pathway analysis of individual pSS clusters. To investigate molecular processes and their biological function underlying each of the pSS patients' clusters, specific differentially expressed genes (DEG) signatures compared to HV were assessed using Limma in the 4 clusters. Ingenuity Pathway Analysis (IPA)

Table 1 Healthy volunteers (HV) and Primary Sjögren's syndrome (pSS) patient characteristics.

			HV (*N* = 330)	pSS Discovery (*N* = 227)	pSS Validation (*N* = 77)	pSS All (*N* = 304)
Demography						
Age		*n*	330	227	77	304
		Mean ± SD	53.294 ± 10.998	58.524 ± 13.440	58.039 ± 13.554	58.401 ± 13.448
Gender		*n*	330	227	77	304
	Female	*n* (%)	302 (91.52)	211 (92.95)	71 (92.21)	282 (92.76)
Obesity (BMI > = 30)		*n*	328	218	74	292
	Yes	*n* (%)	24 (7.27)	30 (13.76)	3 (4.05)	33 (11.30)
Race		*n*	330	227	77	304
	Asian	*n* (%)	2 (0.61)	1 (0.44)	1 (1.30)	2 (0.66)
	Black/African American	*n* (%)	—	—	1 (1.30)	1 (0.33)
	Caucasian/White	*n* (%)	328 (99.39)	224 (98.68)	74 (96.10)	298 (98.03)
	Other	*n* (%)	—	2 (0.88)	1 (1.30)	3 (0.99)
Diagnostic criteria						
Focus score > 1		*n*	—	82	27	109
	Yes	*n* (%)	—	73 (89.02)	24 (88.89)	97 (88.99)
Anti-SSA positivity		*n*	—	227	77	304
	Yes	*n* (%)	—	205 (90.30)	69 (89.61)	274 (90.13)
Disease activity						
Disease duration, years		*n*	—	225	77	302
		Mean ± SD	—	10.788 ± 7.535	11.094 ± 9.620	10.866 ± 8.101
Disease activity (PGA*)		*n*	—	211	75	286
		Mean ± SD	—	25.687 ± 18.976	24.840 ± 20.984	25.465 ± 19.488
ESSDAI (**)		n	—	133	60	193
		Mean ± SD	—	4.609 ± 5.358	4.850 ± 5.495	4.684 ± 5.388
ESSPRI (**)		*n*	—	106	44	150
		Mean ± SD	—	5.176 ± 2.286	4.568 ± 2.648	4.998 ± 2.405

n: Number of patients with available information.
(*) PGA: Physician Global Assessment.
(**) collected in a substudy.

was subsequently applied to determine the most significantly dysregulated canonical pathways with Benjamini–Hochberg false discovery rate (FDR) adjusted *p*-value ≤ 0.05 and absolute fold change (FC) ≥ 1.5. As a result, 284 DEG were found significant in C1, 301 DEG in C3 and 1686 DEG in C4 (Supplementary Data 1).

Since no DEG were noticed in C2 when compared to HV, only C1, C3, and C4 were functionally annotated. Top 20 significant canonical pathways within each DEG signature are presented in Supplementary Data 2 and pathways related to the most significantly enriched immunological responses are reported as radar plots in Fig. 1c. While all 3 clusters were enriched in genes involved in antiviral and anti-bacterial responses indicative of an innate-mediated activation profile, C1 was mainly enriched with IFN-related pathways including IFN signaling, role of pattern recognition receptors for bacteria and viruses and Interferon Regulatory Factor (IRF) activation. Notably, C3 and C4 were further characterized by alterations in biological networks linked to adaptive immunity. Specifically, significant activation of canonical pathways related to B cell activation such as B cell receptor signaling, and B cell development were observed in C3. In addition, comparative analyses provided evidence for IL7-signaling up-regulation and LXR/RXR activation in C3 compared to C1.

Interestingly, C4 was the endotype with the highest number of DEG compared to HV with highly heterogeneous dysregulated canonical pathways. Ingenuity pathway analysis confirmed the activation of T and B lymphocyte related pathways reflecting Th1 and Th2 activation, B cell receptor signaling, together with prominent inflammatory signatures most particularly linked to cytokine signaling (IL-6 and IL-10 signaling, IL-15 production, STAT-3 pathway).

Further upstream regulator analysis predicted significant activation of IFN-α in all three clusters, as well as CpG ODN in C3 and LPS, IFNγ, TNF-α, and IL-4 in C4, further highlighting B cell activity and inflammatory responses in C3 and C4, respectively.

Noteworthy, while C2 displayed no DEG compared to HV, 14 genes were differentially expressed in C2 patients positive for SSA antibodies compared to HV whereas only 2 DEG were found in SSA-negative C2 patients. These SSA-positive C2 patients were characterized by significant enrichment in IFN-related genes compared to HV including *IFI44, IFI44L, IFI6, IFIFT1, IFIT3, ISG15, MX1, OAS3, SERPING1,* and *SIGLEC1* (Supplementary data 1).

To further characterize patient cluster variability at a molecular level, we then used the blood transcriptome modular repertoire recently established on an expended range of disease and pathological states. The latter includes 382 transcriptome modules based on genes co-expression patterns across 16 diseases and 985 unique transcriptome profiles[9]. Again, no aggregate was found differentially expressed in C2 confirming the healthy-like profile of these patients, whereas an up-regulated IFN signature dominated in C1, C3, and C4 (Fig. 2). In C4, the most induced modules include genes associated with inflammation and neutrophils. As the highest inflammatory phenotype, C4 is associated with a hypercytokinemia/hyperchemokinemia

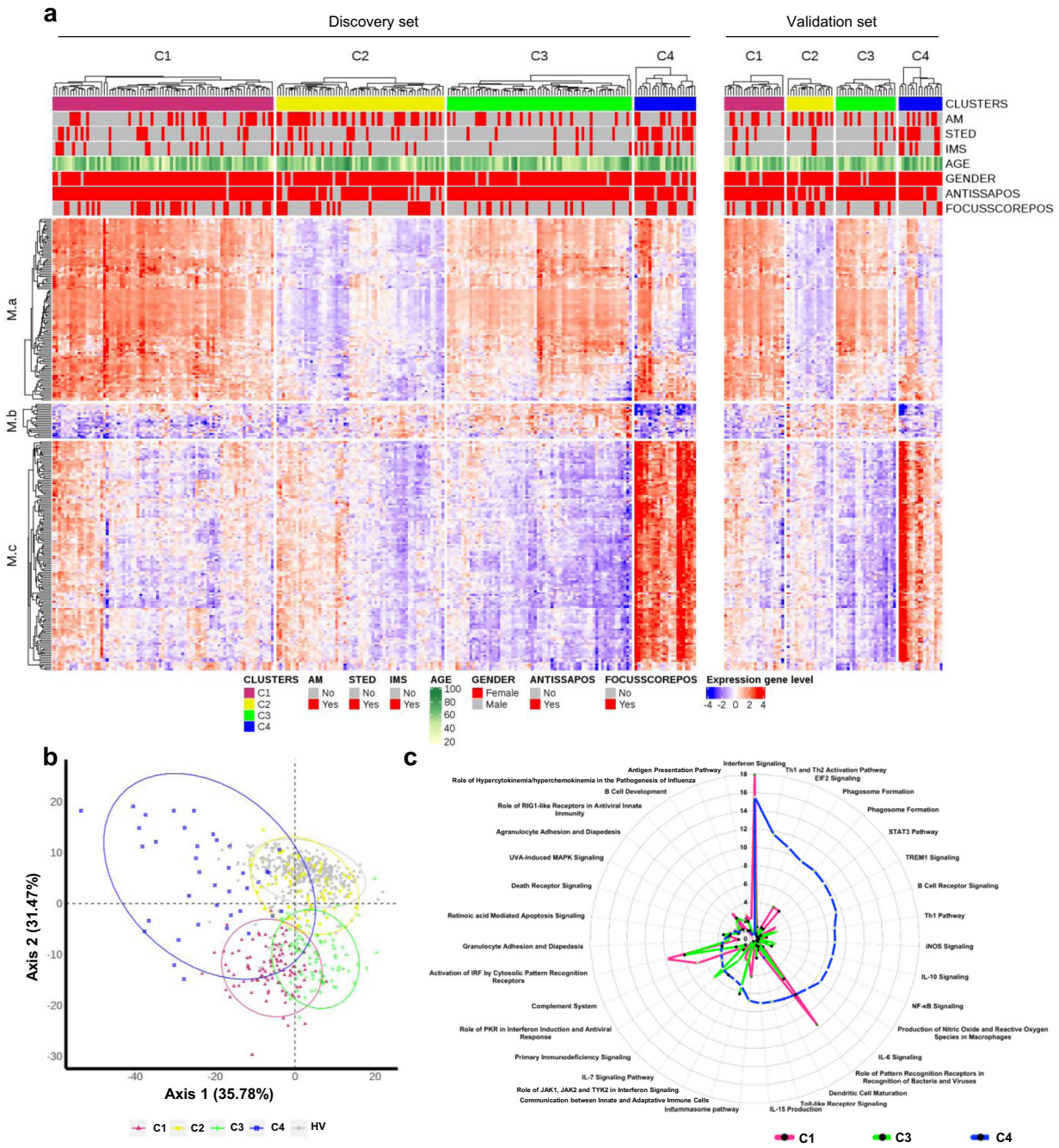


Fig. 1 Molecular pattern distribution is represented by 4 clusters of pSS patients with different canonical pathways. a Heatmap performed for 304 pSS patients (Discovery set: 227, Validation set: 77) showing the distribution of gene transcripts across the 4 clusters. In columns patients are grouped by cluster assignment and in rows genes are grouped by functional modules. Each subset of patients (discovery set on the left and validation set on the right) is presented separately. Red represents overexpression and blue represents under-expression. At the top of the figure annotations show: each of the treatment groups for each individual (AM: antimalarials, STED: steroids and IMS: immunosuppressors, red represents patients with treatment and gray represents patients without treatment), age (levels of yellow to green with yellow for younger patients and dark green for older patients), gender (red represents woman and gray represents man), ANTISSAPOS: anti-SSA/Ro antibody positivity, FOCUSSCOREPOS: focus score of ≥1 foci/mm^2 (red represents focus score of ≥1 foci/mm^2 and gray represents focus score of <1 foci/mm^2). **b** Scatterplot of the first two components PCA (performed for 304 pSS patient and 330 HV) model showing clearly defined clusters in signature gene. HV (gray dot) are confused with C2 cluster (yellow dot). **c** Top 20 most significant canonical pathways for each cluster. Radar plots are represented according to –log (*p*-value) (Fisher's exact test) associated to the most significant pathways of each cluster; C1 (pink), C3 (green), C4 (blue).

observed in modules (M13.16, M15.84, M16.80) consistent with an upregulation of the TNF-associated module (M16.47) and a downregulation of the TGFβ-associated module (M16.65) (Fig. 2). Some modules were under-expressed, such as those associated with both protein synthesis (M12.7, M11.1, M13.28, M14.80), B cells (M13.27, M12.8) and T cells (M15.38, M14.42, M12.6). Genes mainly overexpressed in C1 were also implicated in inflammatory responses and neutrophils (A33, A35), in parallel with down-regulated B and T cell signatures (Supplementary Fig. 5). Moreover, distinct sub-modules expressed in opposite

Table 2 Descriptive analysis of the clinical parameters by primary Sjögren's syndrome cluster.

			C1 ($n = 101$)	C2 ($n = 77$)	C3 ($n = 88$)	C4 ($n = 38$)	*p*-value
Age, years		*n*	101	77	88	38	
		Mean ± SD	57.327 ± 13.705	58.805 ± 13.688	57.250 ± 12.032	63.105 ± 14.790	0.10
Gender		*n*	101	77	88	38	
	Female	*n* (%)	96 (95.05)	71 (92.21)	81 (92.05)	34 (89.47)	0.70
Age at onset, years		*n*	101	76	88	37	
		Mean ± SD	45.663 ± 14.475	50.428 ± 14.532	47.606 ± 12.687	51.739 ± 16.053	0.071
Disease duration, years		*n*	101	76	88	37	
		Mean ± SD	12.247 ± 8.921	8.965 ± 7.336	10.183 ± 7.210	12.625 ± 8.524	0.029
Disease activity (PGA*)		n	94	71	85	36	
		Mean ± SD	27.245 ± 20.535	22.718 ± 17.698	23.212 ± 18.766	31.556 ± 20.646	0.092
ESSDAI		n	70	52	44	27	
		Mean ± SD	5.029 ± 5.959	3.731 ± 4.594	4.227 ± 4.017	6.370 ± 6.828	0.10
ESSPRI		n	56	43	30	21	
		Mean ± SD	4.833 ± 2.460	5.031 ± 2.429	5.300 ± 2.703	4.937 ± 1.803	0.87
Arthritis		*n*	98	77	86	38	
	Past	*n* (%)	39 (39.80)	18 (23.38)	20 (23.26)	12 (31.58)	0.016
	Present	*n* (%)	2 (2.04)	3 (3.90)	4 (4.65)	5 (13.16)	
Focus score > 1		*n*	96	29	21	14	
	Yes	*n* (%)	39 (40.63)	28 (96.55)	17 (80.95)	12 (85.71)	0.4
Anti-SSA positivity		*n*	101	77	88	38	
	Yes	*n* (%)	99 (99.00)	56 (72.72)	87 (98.86)	31 (81.57)	<0.001
Anti-SSB positivity		*n*	100	77	86	38	
	Yes	*n* (%)	61 (61.00)	12 (15.58)	39 (45.35)	11 (28.95)	<0.001
Hypergammabulinemia		*n*	97	73	86	38	
	Past	*n* (%)	23 (23.71)	8 (10.96)	9 (10.47)	3 (7.89)	<0.001
	Present	*n* (%)	44 (45.36)	10 (13.70)	41 (47.67)	7 (18.42)	
Abnormal inflammatory indexes		*n*	100	77	87	38	
	Past	*n* (%)	28 (28.00)	13 (16.88)	20 (22.99)	12 (31.58)	0.003
	Present	*n* (%)	35 (35.00)	11 (14.29)	22 (25.29)	10 (26.32)	
Reduced C3 levels		*n*	93	74	82	35	
	Past	*n* (%)	13 (13.98)	5 (6.76)	11 (13.41)	4 (11.43)	0.8
	Present	*n* (%)	7 (7.53)	4 (5.41)	5 (6.10)	3 (8.57)	
Reduced C4 levels		*n*	93	74	82	35	
	Past	*n* (%)	13 (13.98)	3 (4.05)	9 (10.98)	4 (11.43)	0.10
	Present	*n* (%)	10 (10.75)	3 (4.05)	3 (3.66)	4 (11.43)	
Abnormal Creatinine		*n*	98	77	88	38	
	Past	*n* (%)	10 (10.20)	4 (5.19)	-	2 (5.26)	0.009
	Present	*n* (%)	5 (5.10)	2 (2.60)	7 (7.95)	6 (15.79)	
Proteinuria		*n*	65	58	56	25	
	Moderate	*n* (%)	5 (7.69)	2 (3.45)	1 (1.79)	3 (12.00)	0.093
	Past	*n* (%)	5 (7.69)	—	3 (5.36)	—	
Current use of antimalarials		*n*	101	77	88	38	
	Yes	*n* (%)	33 (32.67)	42 (54.55)	24 (27.27)	15 (39.47)	0.002
Current use of Immunosuppressants		*n*	101	77	88	38	
	Yes	*n* (%)	17 (16.83)	14 (18.18)	7 (7.95)	15 (39.47)	<0.001
Current use of steroids		*n*	101	77	88	38	
	Yes	*n* (%)	23 (22.77)	14 (18.18)	10 (11.36)	23 (60.53)	<0.001

n: Number of patients with available information, (*) PGA: Physician Global Assessment.
Statistical tests performed: chi-square test of independence for categorial variable and Kruskal–Wallis test for continue variable.

directions allows to functionally discriminate C1 and C3. Patients from C3 demonstrated a significant under-expression of modules related to erythrocytes (A37; M9.2, M11.3) and cytokines/chemokines (A35; M15.84, M13.16) and an increased expression in some of the B cell modules (A1; M12.8) (Supplementary Fig. 5 and Fig. 2).

IFN signatures. Consistent with the literature, the most significantly enriched pathway confirmed to be up-regulated in all three clusters was the IFN signaling pathway (Fig. 2, Supplementary Fig. 5). In SLE, Chiche et al. have previously identified three strongly up-regulated IFN-annotated modules (M1.2, M3.4, and M5.12) from peripheral blood transcriptomic data, with for each module a distinct activation threshold[10]. Genes within the M1.2 module are induced by IFNα, while other genes from both M1.2 and M3.4 are up-regulated by IFNβ, corresponding to a type I IFN signature. The M5.12 genes are poorly induced by IFNα and IFNβ alone but are rather up-regulated by IFNγ characterizing a type II IFN signature[11]. Moreover, transcripts belonging to M3.4 and M5.12 were only fully induced by a combination of Type I and Type II IFNs. Kirou et al. made similar observations and identified genes preferentially induced by IFNα or IFNγ[12]. The different z-scores were then calculated accordingly to characterize further the IFN signature observed in the various clusters (Fig. 3). All IFN z-scores were increased to some extent in C2 when compared to HV. In line with the strong signal observed, C1 patients had the highest Type I and type II scores. Interestingly, C3 had higher Type I IFN score than C4 but these 2 clusters were not different for Type II IFN score.

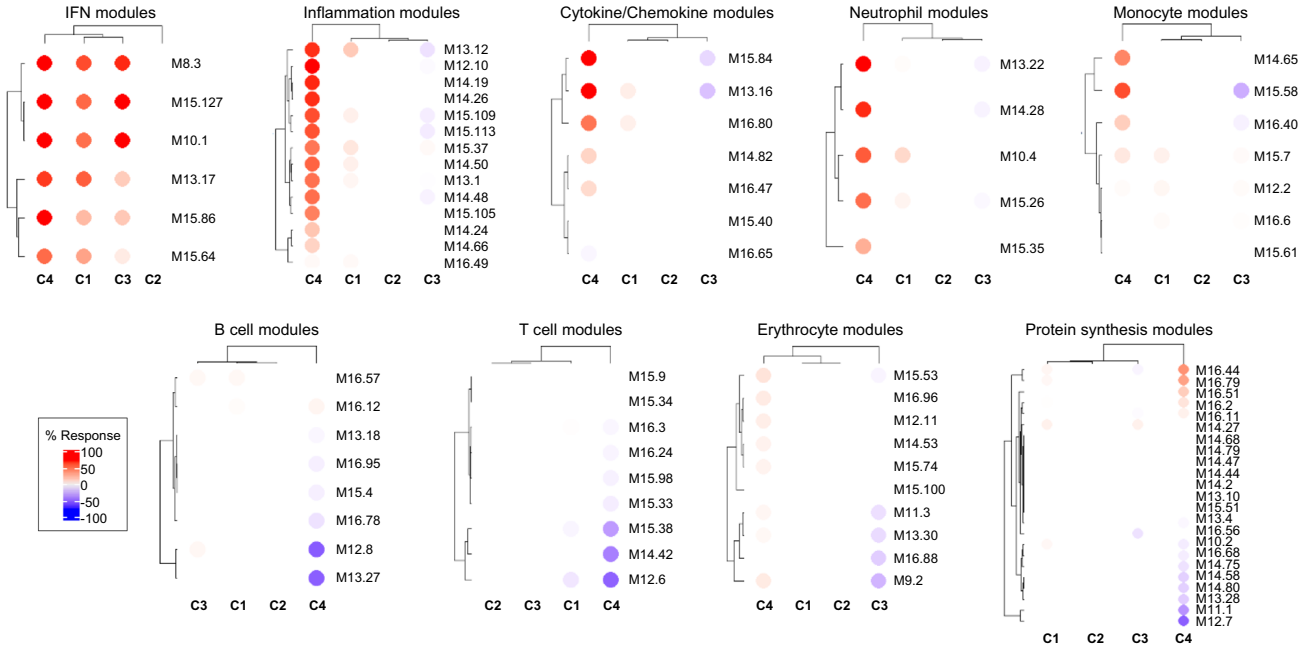


Fig. 2 Patterns of abundance of the different modules distinguish the four pSS clusters. Each heatmap, achieved with BloodGen3Module R package[9], represents one of the most significant patterns differentiating the four clusters of 304 pSS patients (C1: 101, C2: 77, C3: 88, and C4: 38) compared to 330 healthy volunteers (HV). These patterns correspond to modules associated with IFN, neutrophils, inflammation, cytokines/chemokines, protein synthesis, erythrocytes, monocytes, B cells and T cells. Columns on this heatmap corresponds to clusters. Each row corresponds to one of the modules associated with the pattern. For each module, the percentage of increased genes (from 0 to 100) and decreased genes (from 0 to 100) were calculated. A red spot on the heatmap indicates an increase in abundance of transcripts comprising a given module for a given cluster. A blue spot indicates a decrease in abundance of transcripts. The absence of color indicates no changes.

Upstream analysis of C4 DEG predicted IFNγ as an important regulator suggesting that Type II IFN activation was prominent in C4.

Genome-wide association study analysis. We investigated whether clusters showed any differences in the genetic contribution of risk alleles known to be associated with pSS[13–15]. Even in the mid-size cohort of patients analyzed (304 pSS and 330 HV), we unambiguously detected (with signals genome wide significance level $<5 \times 10^{-8}$) 35 single nucleotide polymorphisms (SNPs) in C1 compared to only six in C3 and one in C4 (Fig. 4a, Supplementary Data 3). Interestingly, no significant enrichment was found in C2. The 35 SNPs assessed in C1 are found within genes associated with either the immune system (*HLA-DQB1, HLA-DQA1, HLA-DRA, HLA-C, HLA-G*), signal transduction (*NOTCH4*), developmental biology (*POU5F1*), gene expression (*DDX39B*) or cell cycle (*TUBB*). The presence of such significant genetic associations was already found in clusters of systemic autoimmune disease patients whose molecular disease pathway is the Type I IFN pathway[5]. Moreover, a strong association of SNPs with HLA class II genes was reported in SLE patients with a high level of autoantibodies[16]. One SNP (rs2734583) was common to C1 and C3 and is associated to the *DDX39* gene. Of note, DDX39B, the protein encoded by this gene, is required for the prevention of dsRNA formation during influenza A virus infection, thereby preventing the activation of the Type I IFN system[17]. The five others SNPs in C3 are nearby *HLA-DQA*, *HLA-DRA* (2 SNPs), *BTNL2* and *HCG23*. The only SNP (rs2247056) found in C4, also common with C1, is located in intron 1 of the *LINC02571* gene and was previously associated with a risk for developing SLE.

Linkage disequilibrium is a non-random association of alleles at different loci in a given population. When analyzing linkage disequilibrium (Fig. 4b) in the loci of the 35 SNPs detected in C1 and located on chromosome 6 (from base 29809362 to 32681631), three SNPs were strongly associated in *HLA-DQA1* locus (rs9272219, rs9271588, rs642093), five SNPs in *HLA-DRA | HLA-DQA1* locus (rs7195, rs1041885, rs3129890, rs9269043, rs7749057) and three SNPs in *HCG27 | HLA-C* locus (rs3130473, rs2394895 and rs3130467). Two other regions contain strongly associated SNPs. The *NOTCH4 | C6orf10* locus presented 5 associated SNPs (rs3130347, rs204991, rs3132935, rs7751896, rs9268220) as well as the *IER3 | DDR1* locus (rs3094122, rs6911628, rs3094112, rs2517576, rs3095151).

Methylation analysis. The methylation analysis was performed with a Benjamini Hochberg FDR < 0.1 and absolute ΔBeta > 0.075. Only two differentially methylated positions (DMPs) corresponding to two genes were found in C2. Those DMPs were common with the 3 other clusters (Fig. 5a) and were located in the TSS1500 shore of the *NLRC5* gene and in the 5'UTR of the gene encoding *MX1*, two genes involved in the IFN signature. NLRC5 plays a role in cytokine response and antiviral immunity through inhibition of NF-kappa-B activation and negative regulation of Type I IFN signaling pathways[18]. *MX1* encodes an IFN induced dynamic-like GTPase with antiviral activity which was proposed as a clinically applicable biomarker for identifying systemic Type I IFN in pSS[19].

145 DMPs corresponding to 87 genes and 96 DMPs corresponding to 56 genes were found in C1 and C3 respectively, whereas an aberrant methylation status with 8,445 DMPs corresponding to 3,636 genes characterized C4 (Fig. 5a). In order to test whether the methylation defect in C4 was associated with steroids treatment, we compared the 9 untreated to the 17 treated patients. No CpG with a Benjamini-Hochberg FDR adjusted *p*-value < 0.1 was found to be differentially methylated in treated versus untreated patients. A global hypomethylation of CpG was observed for all clusters (89.6% in C1, 100% in C2, 67.7% in C3

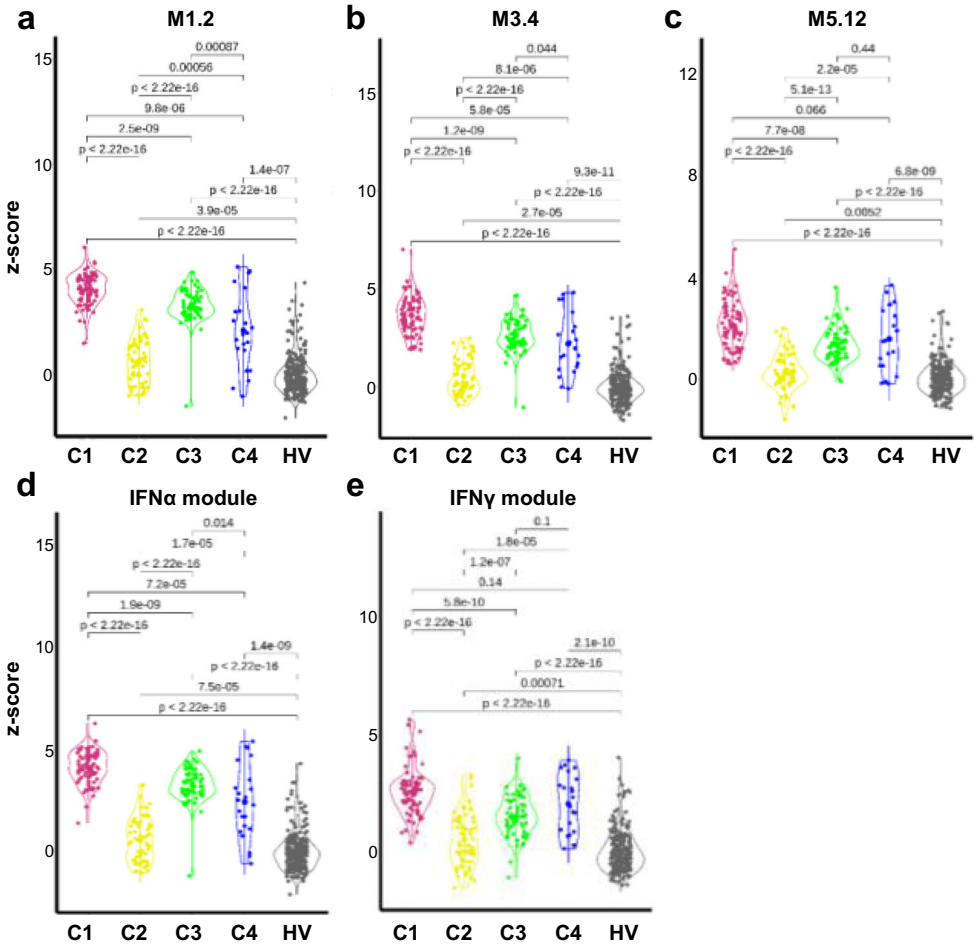


Fig. 3 The 4 pSS clusters show typical IFN signature according to modular IFN z-scores. IFN score analyses were performed for 304 pSS patients and 330 healthy volunteers (HV). Repartition of samples from the 4 pSS clusters are shown according to the most characterized IFN module z-scores. The genes (*IFI44*, *IFI44L*, *IFIT1* and *MX1*) of the M1.2 module (**a**) are induced by IFNα, while genes from both M1.2 and M3.4 (**b**) (*ZBP1*, *IFIH1*, *EIF2AK2*, *PARP9* and *GBP4*) are up-regulated by IFNβ. **c** The genes (*PSMB9*, *NCOA7*, *TAP1*, *ISG20* and *SP140*) from the M5.12 module are poorly induced by IFNα and IFNβ alone while they are up-regulated by IFNγ. Moreover, transcripts belonging to M3.4 and M5.12 were only fully induced by a combination of Type I and Type II IFNs[10]. Other modules identified genes preferentially induced by IFNα (*IFIT1*, *IFI44* and *EIF2AK2*) (**d**) or IFNγ (*IRF1*, *GBP1* and *SERPING1*) (**e**)[12]. Two-tailed pairwise Wilcoxon-rank sum test results are shown. Plots show median with error bars indicating ± interquartile range.

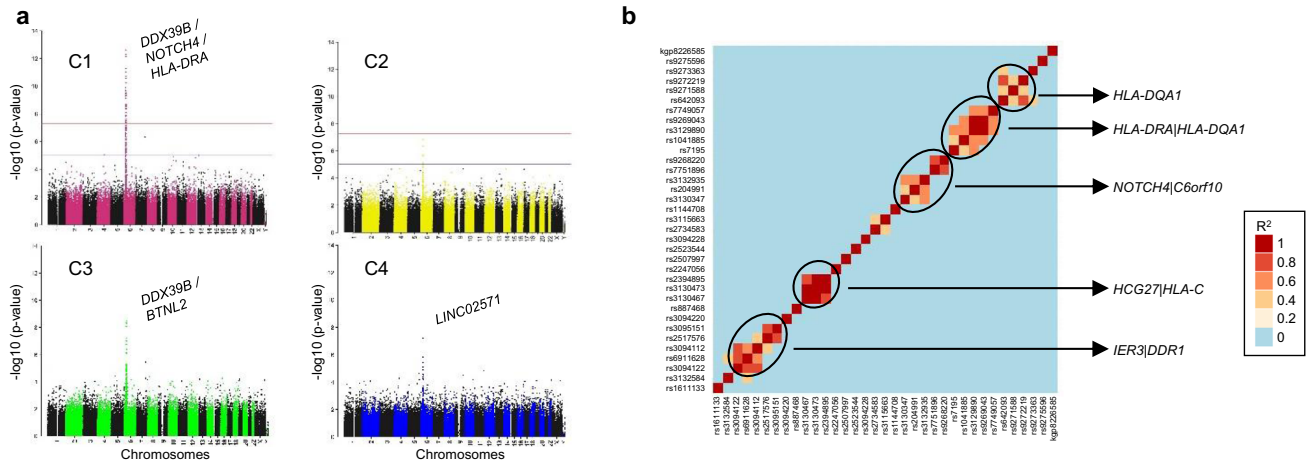


Fig. 4 Cluster genome-wide association analyses (GWAS). GWAS analysis was performed using Plink, an open-source whole genome association analysis toolset, using a logistical regression for 304 pSS (C1: 101, C2: 77, C3: 88 and C4: 38) patients and 330 healthy volunteers (HV) and each cluster was compared to HV. **a** Manhattan plots for each cluster are shown. **b** Linkage disequilibrium analysis in the loci of the 35 SNPs detected in C1 and located on chromosome 6 from base 29809362 to base 32681631. The R^2 correlation coefficient and linkage disequilibrium heatmap were obtained with Plink, and oncofunco R package, respectively. Strongest associations between SNPs are annotated.

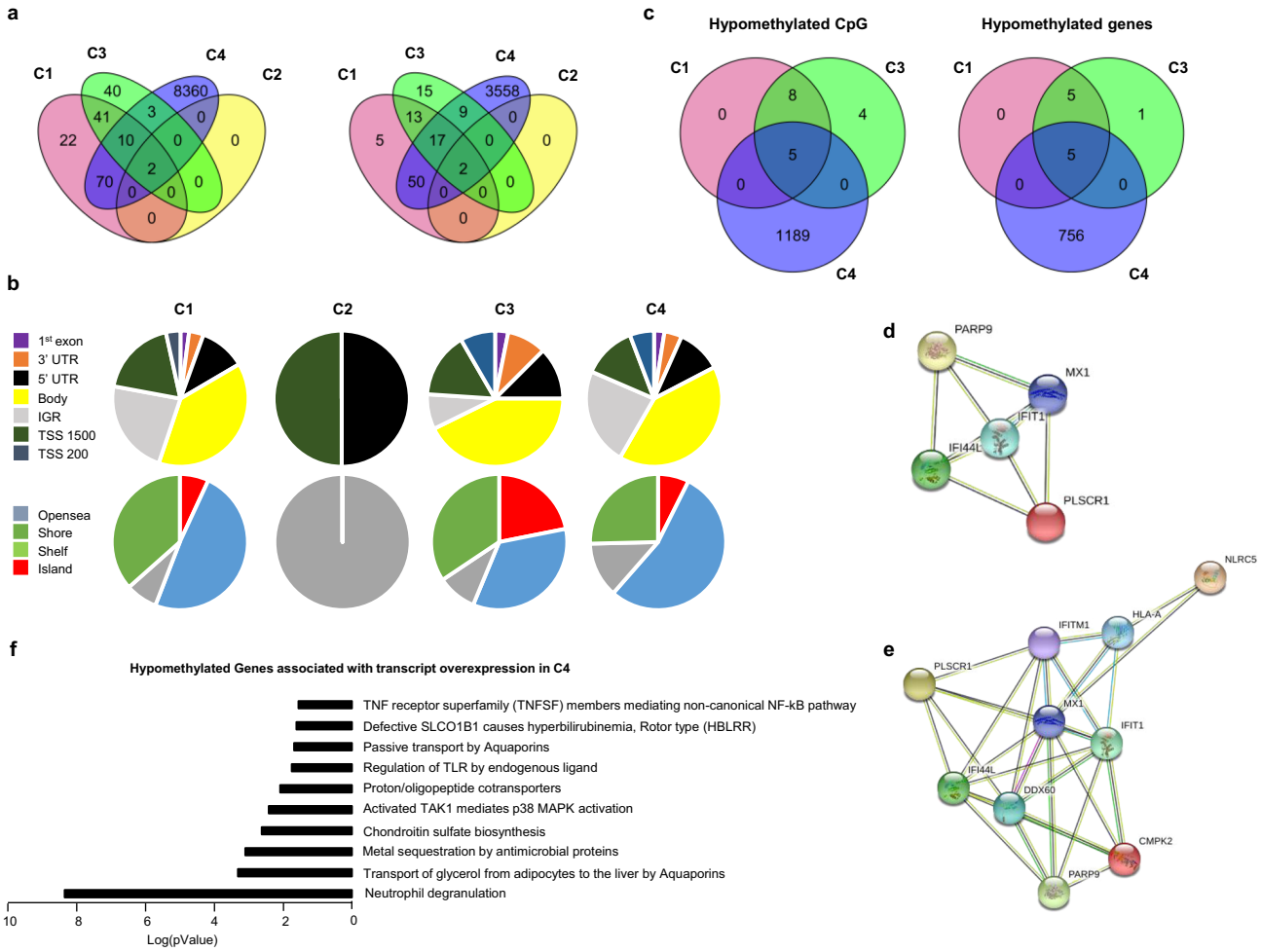


Fig. 5 Methylation analysis confirms the strong IFN signature in C1 and C3 and reveals an aberrant methylation status in C4. Whole blood methylation analysis was performed for 226 pSS patients (C1: 81, C2: 57, C3: 62, and C4: 26) and 175 healthy volunteers (HV) doing pairwise comparisons between each cluster and HV. **a** Venn diagram showing the overlap of differentially methylated CpG sites and genes between the 4 clusters with absolute ΔBeta > 0.075. **b** DMP distribution across the different genomic regions (gene body, 3'UTR, intergenic (IGR), 5'UTR, Exon 1, TSS 1500 and TSS 200; and according to the CpG density to CpG island, shelf, shore, and open sea. **c** Venn diagram showing the overlap of hypomethylated CpG and genes with absolute ΔBeta > 0.15 between the three IFN clusters. **d** Interaction network of these 5 genes common to the three clusters by STRING analysis with a confidence cut-off of 0.4 reveals a common IFN signature. **e** Interaction network of the 10 genes hypomethylated common to C1 and C3 by STRING analysis with a confidence cut-off of 0.4. **f** Reactome analysis[22] of the functional pathways enriched for the 126 genes hypomethylated and over expressed in C4 (absolute ΔBeta > 0.15, FC ≥ 1.5).

and 80.4% in C4). Because functionally important DNA methylation occurs in promoter regions and in CpG islands[20], DMP distribution across the different genomic regions was investigated (Fig. 5b). A higher representation of DMPs in the promoter region was found in C3 (36.4%) and C1 (33.1%) when compared to C4 (29.1%). The consequence was a lower representation of DMPs in intergenic regions for C3 (8.8%) compared to C1 (22.8%) and C4 (23.1%). To gain insight on this pattern, we divided the probes according to CpG islands; shores (regions up to 2 kb from CpG island), shelves (regions from 2 to 4 kb from CpG island) and open sea (the rest of the genome). Interestingly, 21.8% of the DMPs for C3 were located in CpG islands versus 6.9 and 7.4% for C1 and C4, respectively.

To identify the most robust and significant signature of hypo- and hyper-methylated genes, we fixed the ΔBeta cut-off at 0.15. Regarding hypomethylated CpGs, 13 DMPs were found in C1, 17 in C3 and 1,194 in C4, corresponding to 10, 11 and 761 hypomethylated genes, respectively. Five genes with hypomethylated DMPs were common to these 3 clusters (*IFI44L*, *IFIT1*, *MX1*, *PARP9* and *PLSCR1*) (Fig. 5c), corresponding to genes reported to present strong interactions (Fig. 5d). Interestingly, these genes were also significantly hypomethylated in C2 when compared to HV (Supplementary Fig. 6). Of note, 5 additional genes (*HLA-A*, *DDX60*, *CMPK2*, *IFITM1* and *NLRC5*) were common to C1 and C3 and were also strongly associated with the previous ones, reinforcing the IFN signature in these two clusters (Fig. 5e). These common 10 hypomethylated genes are implicated in defense responses to virus and are induced by IFN[21].

The remaining 756 hypomethylated genes in C4 were mainly associated with the neutrophil degranulation pathway. Regarding hypermethylated CpGs, 41 DMPs corresponding to 25 genes were only found in C4. Those genes are mainly implicated in translocation of ZAP-70 to the immunological synapse, phosphorylation of CD3 chains including zeta, platelet activation, signaling and aggregation, homeostasis and PD-1 signaling.

Combining transcriptomic (FC ≥ 1.5) and methylomic (absolute ΔBeta > 0.15) analyses, the transcripts of 8, 8 and 126 genes were found to be increased in association with a decreased methylation status in C1, C3 and C4, respectively. Interestingly, the previously isolated 5 common hypomethylated genes

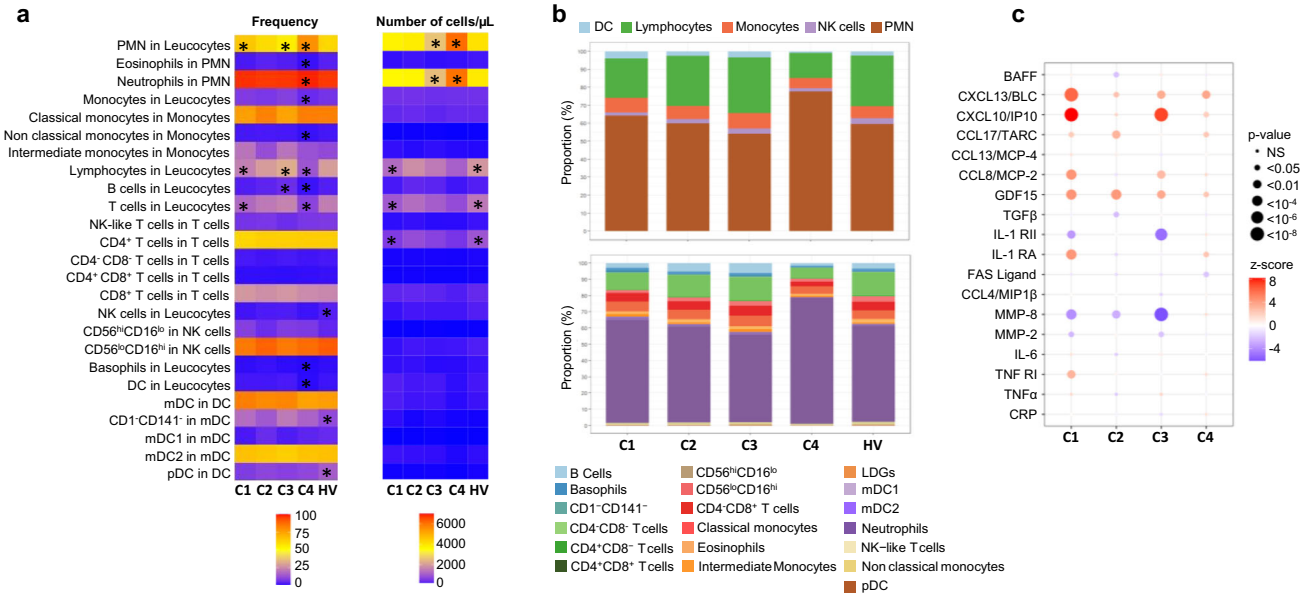


Fig. 6 Cell subset distribution in blood and cytokines, chemokines and inflammatory mediators in serum in the 4 clusters and healthy volunteers (HV). **a** Flow cytometry analysis was performed for 283 patients (C1: 96, C2: 71, C3: 80, and C4: 36) and 309 HV. The 2 heatmaps show the mean distribution of blood cell subsets in frequency (0–100%) and in absolute numbers (per µL of blood) across the 4 clusters and HV assessed by flow cytometry. Columns represent clusters and HV and rows the different cell subsets. The asterisk means that the cluster (or HV) is statistically different from all the others. **b** Flow cytometry data represented by bar charts cell types proportion per cluster. **c** Serum mediators were analyzed for 192 pSS patients (C1: 67, C2:48, C3: 61, C4:16) and 171 HV. Patient and HV distribution according to each analyzed variable is described in Methods. CXCL13/BLC, FAS Ligand, GDF-15, CXCL10/IP-10, CCL8/MCP-2, CCL13/MCP-4, CCL4/MIP-1β, MMP-8, CCL17/TARC, IL-1 RII, TNF-RI, and IL1-RA were measured using the Luminex system and expressed as pg/ml. Soluble MMP-2, CRP, TNFα, IL-6, BAFF, and TGFβ were measured by the quantitative sandwich enzyme immunoassay technique and expressed as pg/ml. Cytokine or chemokine concentration levels for each cluster were compared to HV. Statistical significance is determined using a one-way ANOVA followed by post-hoc Tukey's test. The significance between the cluster and HV is represented as bullet ranging from small (non-significant) to big (significant). The direction of the association is shown as the z-score where red bullet is up-regulated, and blue bullet is down-regulated.

implicated in IFN signaling were also overexpressed at the transcriptional levels in the 3 clusters. Transcript overexpression was strongly associated with hypomethylation in C1 (8/10) and C3 (8/11) and to a less extend in C4 (126/761). Among the 126 genes from C4, 21 were implicated in neutrophil degranulation which constitutes the most relevant pathways according to Reactome Pathway Database[22] (Fig. 5f). Only 6/25 transcripts were repressed in association with an increased methylation status of their genes in this cluster (*CD247*, *CD3G*, *CDC25B*, *CXCR6*, *TBC1D4*, *UBASH3A*).

Flow cytometry analysis. As significant alterations in patterns of peripheral blood leukocytes have been previously described[23,24], we then investigated the composition of leukocyte subsets in the various clusters. (Fig. 6a, b, Supplementary Fig. 7). In C2, the frequency and absolute numbers were similar to HV in all the different subsets analyzed. An increase in the frequency of monocytes and lymphocytes characterized C3, in association with a marked increase in the frequency of B cells. At the same time, a lymphopenia affecting mainly T cells was found in C1. Finally, the most distinguishable cluster in terms of distribution and absolute number of cells is C4. Specifically, C4 was characterized by higher percentages and absolute numbers of PMN (especially neutrophils) in peripheral blood in comparison with those in other clusters and HV. Conversely, the percentages of lymphocytes (B and T cells) and monocytes were markedly decreased in C4 compared to either the controls or the other clusters. Finally, lower frequencies and absolute numbers of basophils and DCs were also found in this cluster.

An in-depth analysis of the different cell subpopulations was then conducted. First, monocytes represent a heterogeneous cell population in terms of both phenotype and function. Based on the expression of CD14 and CD16, 3 monocyte subsets can be defined, including classical ($CD14^{++}CD16^{-}$), intermediate ($CD14^{++}CD16^{+}$) and non-classical ($CD14^{+}CD16^{++}$). Classical monocytes are critical for the initial inflammatory response, can differentiate into macrophages in tissue and contribute to chronic disease. Intermediate monocytes are highly phagocytic cells that produce high levels of ROS and inflammatory mediators. Non classical monocytes have been widely viewed as anti-inflammatory, as they maintain vascular homeostasis and constitute a first line of defense in recognition and clearance of pathogens[25]. Interestingly, the frequency and absolute number of intermediate monocytes were increased in C1 and C3 whereas the frequency of classical monocytes was decreased when compared to the 2 others and the nonclassical subset was markedly decreased in C4, in line with the inflammatory response observed in these different clusters.

Second, NK cells are defined by the expression of CD56 and the lack of CD3-TCR complex. Moreover, based on CD16 and CD56 expression levels, they are classified in two subsets: $CD56^{hi}CD16^{lo}$ and $CD56^{lo}CD16^{hi}$. The latter NK cell subset mediates natural and antibody-dependent cellular cytotoxicity, exhibiting high levels of perforin and enhanced killing. In contrast, $CD56^{hi}CD16^{lo}$ NK cells are characterized by low levels of perforin, and are primarily specialized for cytokine production including IFN[26,27]. Accordingly, the frequency of $CD56^{hi}CD16^{lo}$ NK cells subset over $CD56^{lo}CD16^{hi}$ was increased in C4, C1, C3 and to a lower extent in C2. This may partly explain the up-regulation of cytokines and interferon pathways in disease clusters. Although plasmacytoid dendritic cells (pDCs) are thought to represent the main IFNα producing cells, no differences were observed between clusters and their reduction was confirmed in peripheral blood of pSS patients when compared to HV[28].

Cytokine analysis. We subsequently assessed whether pSS clusters also showed differences in systemic parameters of inflammation, such as cytokines, chemokines and other soluble factors (Fig. 6c and Supplementary Fig. 8). The IFNγ-induced protein (CXCL10/IP-10) as well as CCL8/MCP-2 and TNFα were increased in C1 and C3, i.e. the two main clusters associated with a strong IFN signature. At the same time, IL-1 RII, the decoy receptor for cytokine belonging to the IL-1 family, was down regulated in C1 and C3. Overall, C1 was largely enriched in CXCL13/BLC, IL-6, and IL-1RA. Levels of MMP-8, a protease mainly expressed by neutrophils, were not different from HV in C4 but lower in the other clusters. Of note, many cytokines such as CXCL10/IP-10, CXCL13/BLC, BAFF, and GDF15 were increased in all clusters including C2 when compared to HV. However, no differences between clusters were found for CRP, Fas Ligand, CCL13/MCP-4, CCL4/MIP-1β, CCL17/TARC and TGFβ.

To confirm that patients with an active IFN signature have elevated circulating Type I IFN, we measured levels of IFNα in plasma using Simoa Single Molecule Array Technology in pSS patients and HV. Median levels of IFNα in plasma were 807 (177–1744) fg/ml and 530 (106–1033) fg/ml in C1 and C3, respectively, while circulating levels in the other clusters and HV were close to the lower limit of quantification (Supplementary Fig. 9a). Interestingly, IFNα in serum was positively correlated with the two IFN transcriptomic modules (M1.2 and IFNα module) described in Fig. 3, especially in C1 and to a lesser extent in C3, confirming the Type I IFN signature observed in these patients (Supplementary Fig. 9b). Of note, half of the patients in C2 received antimalarials and previous studies have also shown that hydroxychloroquine use can reduce the levels of circulating Type I[29,30] and Type II[31,32]; IFN z-scores. IFNα in serum was not associated with ESSDAI (Supplementary Fig. 9b) but higher levels of serum IFNα were associated with hematological and biological domains of ESSDAI (Supplementary Data 4).

Clinical symptoms and serological characteristics. Patient medical history and disease characteristics including clinical and serological parameters were collected for the 304 pSS patients. Details are displayed in Table 2 and Supplementary Data 5. Patients from C2 had a lower disease duration when compared to patients from other clusters.

Although the Physician Global Assessment (PGA) was collected for the whole population, ESSDAI and ESSPRI were only assessed in expert centers (Barcelona, Brest, Cordoba, Geneva, Hannover, Leuven, Milano, Porto and Szeged) in a subset of 193 and 150 respectively of the 304 pSS studied patients (70/101 and 56/101 from C1, 52/77 and 43/77 from C2, 44/88 and 30/88 from C3 and 27/38 and 21/38 from C4, Supplementary Data 5).

The lowest mean ESSDAI score was observed in C2 and the highest ESSDAI and PGA mean scores in C4 (Table 2, Fig. 7a) but there were no statistically significant differences between the 4 clusters. No clear difference in the ESSDAI components nor in the objective measures of ocular and salivary dryness was observed between the 4 clusters. Moreover, there was no significant difference for the global ESSPRI score and its 3 components (i.e. dryness, pain and fatigue) except between SSA-positive C2 patients who reported lower ESSPRI scores (p-value < 0.001) compared to the SSA-negative patients (Supplementary Data 6).

Statistically significant differences in the distribution of reported arthritis (p-value = 0.016), rate of cancer history (p-value = 0.028), coronary artery disease (p-value = 0.002) and chronic obstructive pulmonary disease (p-value = 0.016) were observed between the four clusters. (Supplementary Data 7). Interestingly, patients from C4 reported more severe clinical symptoms compared to the 3 other clusters.

Some serological characteristics were significantly different across the 4 clusters, hypergammaglobulinemia (p-value < 0.001) (Table 2), extractable nuclear antigen (ENA) antibodies (p-value <0.001), the presence of serum anti-SSA52/anti-SSA60 autoantibodies (p-value < 0.001) and higher circulating kappa and lambda free light chains (cFLC) (p-value < 0.001) (Fig. 7b, and Supplementary Data 8). C1 and C3 were associated with higher levels of these parameters when compared to C2 and C4. Moreover, C2 and C4 were enriched in patients with glandular manifestations of the disease assessed by a positive focus score in the absence of anti-SSA antibodies (Table 2).

In addition, the levels of rheumatoid factor (p-value < 0.001) and complement C4 fraction levels (p-value = 0.003) were statistically different between the four clusters. C1 was characterized by a higher rheumatoid factor and by a reduced complement C4 fraction levels compared to the other clusters. While some patients presented anti-dsDNA antibodies in C1 and C3 and anti-CCP antibodies in C4, almost none of these autoantibodies were present in the other clusters (Supplementary Data 8).

Prediction of patient membership to each of the four clusters. We then developed through machine learning approaches a composite model able to predict, according to a small number of variables, to which of the 4 clusters each patient belongs (see Methods). The proposed composite model was built with a 2-step approach to allocate patient to the right cluster (Supplementary Fig. 10). The final sets of selected features were composed of 10 genes for the C4 prediction model (first step) and 31 genes for the C1, C2, and C3 classification model (second step). The distribution among clusters of the variance stabilizing transformation (vst) normalized expression for all these transcripts is shown in Supplementary Fig. 11. The validation set (Fig. 1 and Table 1) was used for training, due to the heterogeneity of C4 pSS patients in this set, and the composite model was then run on the discovery set. The accuracy of the model was 95.15%, with 99.12% and 95.57%, for the first and the second steps respectively. The confusion matrix, the corresponding discriminant function analysis, and the probabilities to belong to one of the 4 clusters are shown in Fig. 8a, b, and Supplementary Data 9, respectively.

To generalize the composite model, we used an independent inception cohort of 37 pSS patients. After prediction, C1 contained 16 patients (43.2%), C2 6 patients (16.2%), C3 7 patients (18.9%) and C4 8 patients (21.6%). The corresponding discriminant function analysis and the probabilities for a patient to belong to one of the 4 clusters are shown in Fig. 8c and Supplementary Data 10, respectively. We then used the minimal list of 257 discriminative genes signature previously selected in Fig. 1a to generate a heat map with the prediction established by the composite model (Supplementary Fig. 12a). The clusters observed had the same profile than those identified in the discovery set and observed again in the validation set (Fig. 1a), confirming once more the clustering model. Furthermore, the predicted patients showed a distribution of the IFN signatures (Supplementary Fig. 12b) consistent with the one characterizing the identified clusters (Fig. 3). Altogether, these observations strengthen the validation of our composite model.

Finally, in order to allow our model to process other cohorts of patients, we implement an interpolation function based on 6 genes presenting a constant expression across all 4 clusters and HV (Supplementary Fig. 13). The composite model is integrated into an analysis tool available on the laboratory's github repository[33].

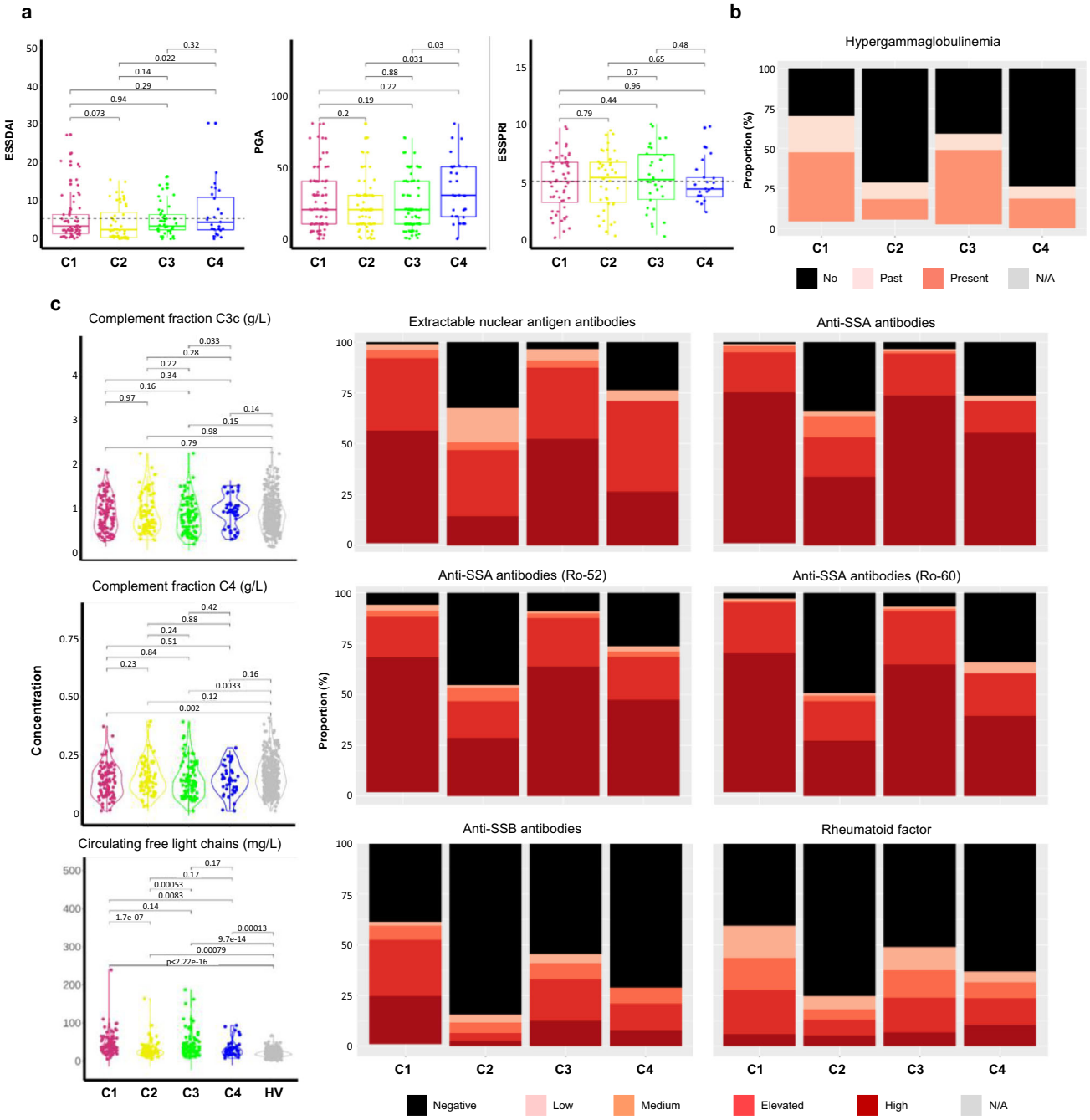


Fig. 7 Disease activity and serological distributions in the 4 clusters. a ESSDAI collected for 193 pSS patients (C1: 70, C2: 52, C3: 44, C4: 27), PGA collected for 286 pSS patients (C1: 94, C2: 71, C3: 85, C4: 36,) and ESSPRI collected for 150 pSS patients (C1: 56, C2: 43, C3: 30, C4: 21) distributions are shown in the 4 clusters. Two-tailed pairwise Wilcoxon-rank sum test results are shown. **b** The barplot shows the proportion of past (light orange) or present (orange) hypergammaglobulinemia (C1: 97, C2: 73, C3: 86, C4: 38) in each cluster. **c** Extractable nuclear antigen antibodies, anti-SSA antibodies, anti-SSA antibodies (Ro-52), anti-SSA antibodies (Ro-60), anti-SSB antibodies, rheumatoid factor were performed for 304 pSS patients (C1: 101, C2:77, C3:88, C4:38) and 330 HV and measured in serum, at the same center, using an automated chemiluminescent immunoanalyzer (IDS-iSYS). Barplots show the proportion of concentration level in each cluster (black: negative, light pink: low, orange: medium, red: elevated and dark red: high). Turbidimetry was used for rheumatoid factor (RF), complement fractions C3c and C4 determination and circulating free light chains. Statistical significance is determined by two-tailed pairwise Wilcoxon-rank sum test. Plots show median with error bars indicating ± interquartile range. Patient and HV distribution according to PGA and biological parameters analyzed variable is described in Methods.

Discussion

Over the last decade, numerous targeted immunomodulatory therapies for pSS have failed to show a benefit in clinical trials, hence no disease-modifying therapy has yet been approved for this disease[34–39]. The heterogeneous nature of pSS and its non-linear development, with flares of activity and subsequent remission associated to a very heterogeneous clinical presentation may explain clinical trial failures[40]. In this context, there is growing interest in the identification of well-characterized subgroups of patients, a prerequisite to the identification of molecular biomarkers predictive of treatment response[41].

We report herein on a large molecular profiling study carried out in pSS patients, a comprehensive molecular profiling of these patients irrespective of their clinical phenotypes. Previous studies

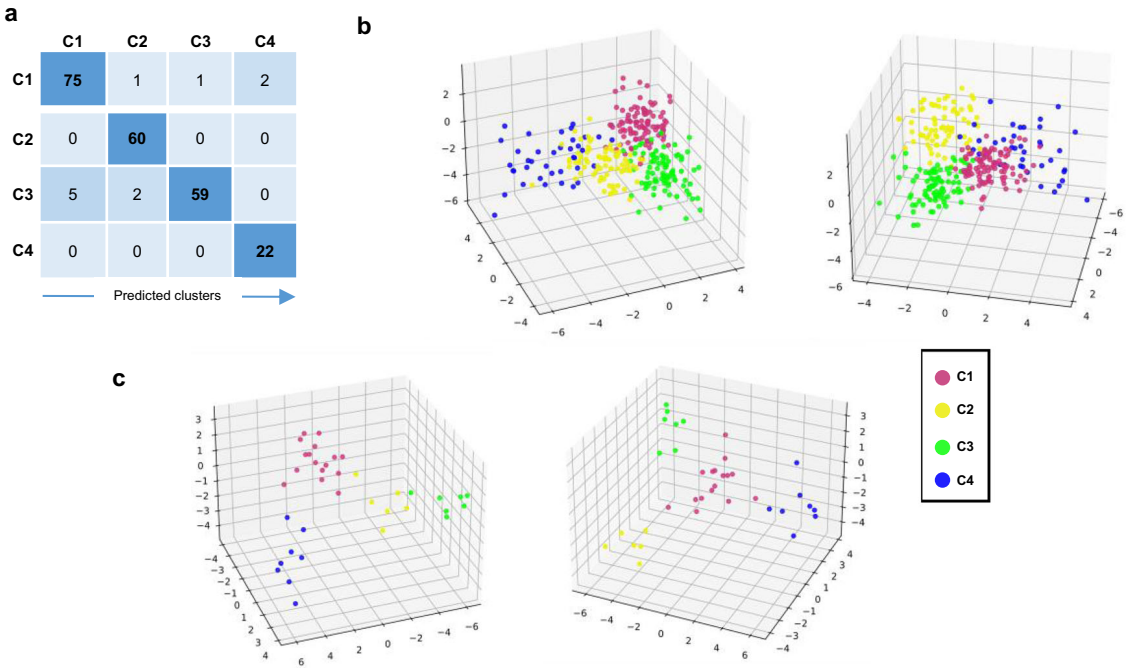


Fig. 8 Development of a composite model to predict the belonging of a patient to one of the 4 clusters. a Confusion matrix of the composite model in the discovery cohort performed for 227 pSS patients (C1: 79, C2: 60, C3: 66, and C4: 22) is shown. **b** Discriminant function analysis (DFA) of the predicted patients from the discovery cohort shows clearly separated clusters. Two different views of the same DFA are shown. **c** DFA of the predicted patients from the inception cohort shows clearly separated clusters. Two different views of the same DFA are shown. Thirty-seven pSS patients from the inception cohort were analyzed and predicted as C1: 16, C2: 6, C3: 7, and C4: 8.

in pSS focus particularly on the IFN signaling involvement[11]. Thereby, pSS patients could be stratified in interferon negative, Type I or Type I + II positive subgroups with higher prevalence of anti-SSA and anti-SSB among those with IFN activation without relation with systemic activity. Another group[42] performed a clustering analysis of blood gene expression microarray which classified the 47 pSS patients in three clusters characterized by IFN and inflammation with no discriminant clinical features Moreover, four subgroups of patients with similar patients' clinical characteristics were identified based on absolute cell counts per µL of blood[23]. Lastly, a stratification based on patient clinical phenotypes characterized a posteriori at the molecular level was proposed[43]. These works provide good basis for building a molecular taxonomy of pSS. Our integrative approach using multi-omics and patient clinical characteristics allows going further in understanding pSS heterogeneity.

We identified transcriptional modules allowing to separate pSS patients into four distinct clusters, irrespective of their treatment, reflecting specific patterns of immune dysregulation, with disease activity and patient reported symptom mean scores similar to naturalistic cohorts like ASSESS[44] and UKPSSR [45].

Patients from C2 displayed a healthy-like profile which nonetheless encompasses bona fide pSS patients reporting a similar level of objective symptoms of dryness, pain and fatigue, albeit a lower ESSDAI compared to the 3 other clusters. C2 was also enriched in patients with glandular manifestations of the disease assessed by a positive focus score and no anti-SSA antibodies. A similar cluster was recently described[42] with no increase in the IFN modules and minimal activity of inflammation-related gene modules. Noteworthy, all molecular profiling data reported here were obtained from blood samples which could affect interpretation of some of the results. For example, the reduction of peripheral blood pDCs of pSS patients when compared to HV already reported[28] does not consider that pDC are enriched in the salivary glands and the possibility that tissue sites may be the major source of IFNα in these individuals[46]. Extending in the future those analyses to the salivary gland will provide a more complete picture of the pathophysiology of the disease, especially in C2.

The three other clusters exhibited significant differences with HV and in particular a prominent IFN gene signature. These findings add to the growing evidence towards a significant role of the IFN pathways in the pathogenesis of systemic and organ-specific disorders including pSS. Whereas Type I IFN were proposed as predominant contributors in the pathogenesis of pSS, a role of Type II IFN in disease pathogenesis has also been highlighted[6,47]. Interestingly, our results show that the IFN signature in the 3 IFN-driven clusters is different. C1 patients had the highest Type I and Type II IFN scores, C3 a higher Type I IFN score than C4, these 2 clusters having similar Type II IFN score. Thus, C4 IFN score was mainly driven by IFN Type II activation. Consequently, C1 and C3 were similar to the IFN cluster recently described by James et al.[42] also associated with high blood protein levels of CXCL10/IP-10.

In line with observed IFN scores, circulating serum levels of IFNα were positively correlated with Type I IFN signature (Supplementary Fig. 9 and Fig. 3) especially in C1 and to a lesser extent in C3. However, levels of IFNα in serum were not correlated with ESSDAI global score, but higher levels of serum IFNα were associated with hematological and biological domains of ESSDAI.

While C1 was mainly driven by IFN, an increase in frequency of B lymphocytes in the blood associated with a significant activation of canonical pathways related to B cell activation such as B cell receptor signaling, and B cell development were observed in C3. Main biological features associated with C3 but also C1 were hypergammaglobulinemia, anti-nuclear antibodies, the presence of serum anti-SSA52/anti-SSA60 autoantibodies and higher cFLC confirming what was already reported in autoantibody-positive pSS patients[21]. Finally, SNPs associated with HLA class II genes

were mainly reported in patients from C1 and C3 presenting a positive IFN signature and high levels of autoantibodies as already shown in SLE[16].

Patients from C4 exhibited a more severe clinical phenotype compared to the others with an inflammatory transcriptomic signature particularly linked to cytokine signaling from the acute phase response. C4 was also characterized by a massive lymphopenia and high levels of neutrophils. The neutrophil-to-lymphocyte ratio (NLR) has been previously shown to correlate with disease activity in systemic autoimmunity[48,49] and elevated NLR are thought to represent a pro-inflammatory state. Indeed, in a study of 483 adult patients with multiple sclerosis, NLR could differentiate between relapsing-remitting and primary progressive multiple sclerosis and predict worsening disability[50]. Further studies are required in pSS to evaluate the importance of this ratio.

Because the main current challenge in clinical trials of new therapies for pSS is the selection of the appropriate patients, we propose here a combination of molecular parameters allowing patient classification by endotypes (Supplementary Fig. 14). We then developed a composite model derived from machine learning, based on the use of a limited number of transcripts from whole blood RNASeq and validated in an independent data set from a pSS inception study, to allow a reanalysis of the previous and ongoing clinical trials to depict predictors of treatment response.

These findings have major implications for the treatment of pSS patients, providing a rationale for both optimal drug positioning and combinations of drugs with complementary mechanisms of action. Specifically, our findings provide a strong rationale for treating patients with either a C1, C3, or C4 profile with inhibitors of type I IFN responses alone or in combination as they support the relevance of B cells as potential therapeutic targets in C3 patients. Trials with B cell depleting antibodies (rituximab) have shown promising results primarily in reducing systemic activity in pSS[51].

Areas requiring further investigation have been identified. First, although our identified cluster gene signatures are strong enough to overcome the disequilibrium in blood cell counts and are not associated with disease duration, except for C2, RNA-Seq analysis is oblivious to sample cell-type composition[52]. Further analyses are on-going, using deconvolution approaches. Second, as hypotheses were derived from a cross-sectional study and a small inception cohort, findings need to be confirmed in longitudinal cohorts to clarify whether patients will stay longitudinally in their initial cluster whatever the disease activity level and the treatments received, or whether treatments decrease disease activity by modifying the extent and scope of gene signaling dysregulations.

Altogether, our results can improve pSS treatment strategies allowing a patient centric approach. This paradigm already implemented in the oncology field will increase the probability of trial successes and boost the development of new efficient drugs against pSS.

Methods

Computational tools. Except when indicated, data analyses were carried out using either an assortment of R system software (http://www.R-project.org, V2.10.1) packages including those of Bioconductor or original R code. R packages are indicated when appropriate. For GWAS analysis, we used Plink, an open-source whole genome association analysis toolset. Machine learning approaches were carried out using python programs (v3.8.5) based on the following modules: scikit-learn, numpy and xgboost.

Patient population. The present study was conducted in patients with pSS and HV included in the European multi-center cross-sectional study of the PRECISESADS IMI consortium which involved patients from seven systemic autoimmune diseases. This study was a pre-planned substudy to be specifically conducted in the pSS population and fulfill the STROBE statements (Supplementary note). Diagnosis of pSS was made according to the 2002 American-European Consensus Group classification criteria, with at least the presence of anti-SSA and/or a positive focus on a minor salivary gland biopsy. Choice of the patient analysis set is detailed in Supplementary Fig. 1a. Recruitment was performed between December 2014 and October 2017 involving 19 institutions in 9 countries (Austria, Belgium, France, Germany, Hungary, Italy, Portugal, Spain and Switzerland). The composite model was validated using transcriptome of 37 pSS newly diagnosed patients recruited in the inception study also obtained from the PRECISESADS consortium. Inception patients were recruited by 10 institutions in Spain, Belgium, France, Italy, Germany and Switzerland. Eligible patients were diagnosed within less than a year since pSS diagnosis.

The two studies (cross-sectional and inception) adhered to the standards set by International Conference on Harmonization and Good Clinical Practice (ICH-GCP), and to the ethical principles that have their origin in the Declaration of Helsinki (2013). Each patient signed an informed consent prior to study inclusion. The Ethical Review Boards of the 19 participating institutions approved the protocol of the cross-sectional study. Moreover, the protocol of the inception study was approved by the ethical committees of the 10 participating institutions. These 10 sites were also participating to the cross-sectional study, therefore these ethical committees reviewed both protocols. The ethical committees involved were: Comitato Etico Milano, Italy; Comité de Protection des Personnes Ouest VI Brest, France; Louvain, Comité d'Ethique Hospitalo-Facultaire, Belgium; Comissao de ética para a Saude—CES do CHP Porto, Portugal; Comité Ética de Investigación Clínica del Hospital Clínic de Barcelona, Spain; Commissie Medische Ethiek UZ KU Leuven/Onderzoek, Belgium; Geschaftsstelle Ethikkommission, Cologne, Germany; Ethikkommission Hannover, Germany; Ethik Kommission. Borschkegasse, Vienna, Austria; Comité de Ética e la Investigación de Centro de Granada, Spain; Commission Cantonale d'éthique de la recherche Hopitaux universitaires de Genève, Switzerland; Csongrad Megyei Kormanyhivatal, Szeged, Hungary; Ethikkommission, Berlin, Germany; Andalusian Public Health System Biobank, Granada, Spain.

The protection of the confidentiality of records that could identify the included subjects is ensured as defined by the EU Directive 2001/20/EC and the applicable national and international requirements relating to data protection in each participating country. The cross-sectional and inception studies are registered in ClinicalTrials.com with respectively number NCT02890121 and number NCT02890134.

For each individual, blood samples as well as biological and clinical information were collected as described in the next Methods sections. For more technical details on sample and data collection, please refer to the main PRECISESADS paper [5].

After quality control on transcriptomics RNAseq data (described below), verification of the ARC/EULAR classification criteria (focus score ≥ 1 foci/mm²) and anti-SSA/Ro antibody positivity), and match of the HV to the patients based on age and gender, our final study cohort comprises 304 patients with pSS and 330 HV. This selection is detailed in Supplementary Fig. 1. Among the 304 pSS, 227 (75%) were used for the discovery step and 77 (25%) were kept for validation (Table 1).

Available data. High-dimensional omics genotype, transcriptome, DNA methylome and proportions of relevant cell types using flow cytometry custom marker panels were analyzed from whole blood samples. Low dimensional information was obtained from serum samples, including selected serology information such as autoantibodies, cytokines, chemokines and inflammatory mediators. Of note, except for samples collected for flow cytometry analysis, all samples were shipped by the clinical sites to a Central Biobank (Granada) for processing, storage, and onward shipment to the analysis sites, where the various determinations were performed. Flow cytometry was managed at each center on fresh blood after a multi-center harmonization of flow cytometers to ensure mirroring of all instruments[53,54], thereby allowing subsequent integration of all the data obtained across the different sites and instruments. Consequently, all the different omics samples were processed with the same protocols at the same site (RNA-Seq at Bayer, cytokines at UNIMI, autoantibodies and integrated analyses of flow cytometry at UBO, methylome at IDIBELL, GWAS at CSIC which guarantees the high quality of the data generated.

Methods used for RNA sequencing, quality control, data processing, and expression profiling are detailed below and in Supplementary Fig. 1c.

RNA-Seq. Methods used for RNA sequencing, quality control, data processing, and expression profiling are detailed below and in Supplementary Fig. 1c. Total RNA was extracted from whole blood samples collected in Tempus tubes using Tempus Spin technology (Applied Biosystems). 1857 samples were processed in batches of 384, randomized to four 96-well plates with respect to patient diagnosis, recruitment center and RNA extraction date. The samples were depleted in alpha- and beta-globin mRNAs using globinCLEAR protocol (Ambion) and 1 µg of total RNA was used as input. Subsequently, 400 ng of globin-depleted total RNA was used for library synthesis with TruSeq Stranded mRNA HT kit (Illumina). The libraries were quantified using qPCR with PerfeCTa NGS kit (Quanta Biosciences), and equimolar amounts of samples from the same 96-well plate were pooled. Four

pools were clustered on a high output flow cell (two lanes per pool) using HiSeq SR Cluster kit v4 and the cBot instrument (Illumina). Subsequently, 50 cycles of single-read sequencing were performed on a HiSeq2500 instrument using and HiSeq SBS kit v4 (Illumina). The clustering and sequencing steps were repeated for a total of three runs in order to generate sufficient number of reads per sample. The raw sequencing data for each run were preprocessed using bcl2fastq software and the quality was assessed using FastQC tools. Cutadapt[55] was used to remove 3′ end nucleotides below 20 Phred quality score and extraneous adapters, additionally reads below 25 nucleotides after trimming were discarded. Reads were then processed and aligned to the UCSC Homo sapiens reference genome (Build hg19) using STAR v2.5.2b[56]. 2-pass mapping with default alignment parameters were used. To produce the quantification data, we used RSEM v1.2.31[57] resulting in gene level expression estimates (Transcripts Per Million, TPM and read counts).

For sample filtering, samples were filtered in at least one of the following situations: (i) the total sum of count is too low (<5000,000), (ii) they were extracted with another method than Tempus Spin, and (iii) the RIN (RNA Integrated Number) value of the sample is below 6.5, (iv) samples with RNAseq inferred gender inconsistent with clinical data, and (v) there was a disagreement between genotypes inferred from RNA-Seq and those obtained from GWAS genotyping.

For normalizations and batch correction, read counts were normalized by the variance stabilizing transformation vst function from DESeq2 (v1.30.0) R package[58]. To reduce the effect of the RIN, a correction was applied using the ComBat function from sva (v3.38.0) R package[59], after categorization of RIN values into 7 classes: (7.5, 8], (8.5, 9], (9.5, 10], (8, 8.5], (7, 7.5], (9, 9.5], (6.5, 7].

For Gene filtering, among the 55,771 genes detected in the data, those with 0 count over all the samples or having an expression level below 1 in more than 95% were filtered. At the end, our final RNA-Seq data comprises 16,876 genes. This selection is detailed in Supplementary Fig. 1.

Molecular subgroups discovery. Our rational was to produce a robust classification scheme and to ensure the greatest possible homogeneity within identified subgroups. To this aim, subgroup discovery was based on the pre-processed RNA-seq data of the discovery set (after vst transformation). We implemented a strategy already applied in breast cancer that iterates unsupervised and supervised steps, which was, therefore, designated as "semi-supervised" approach[8]. It is described hereafter and summarized in Supplementary Fig. 2.

Step 1: Unsupervised gene selection

The coefficient of variation ($CV_g = \frac{\sigma_g}{\mu_g}$, with σ_g is the standard deviation of the gene g and μ_g the mean of the gene g estimated on discovery population) and its robust version ($rCV_g = \frac{\gamma_g}{\mu_g}$, with γ_g is the median absolute deviation) were calculated for each gene. Both were highly concordant. The top 25% most variants were selected to perform the subsequent clustering analysis.

Step 2: Robust consensus clustering

To determine the number of clusters, a consensus clustering between three methods was performed: (i) Agglomerative Hierarchical Clustering (hclust function from stats v4.0.2 R package) with Pearson correlation as a similarity measure and the Ward's linkage method, (ii) K-means clustering (kmeans function from stats R package) with 4 groups and (iii) Gaussian mixture clustering (mclust function from mclust v5.4.6 R package).

Step 3: Identification of molecular signature

A supervised analysis was performed on the 149 patients with consistent cluster assignments between the three clustering methods (considered as "core" molecular profils), in order to identify the most discriminating signature of the 4 clusters. The first signature of set of 3577 genes was selected from a classical one-way ANOVA (FDR < 1e-10), and then reduced by Random Forest to 257 top discriminating genes (randomForest function from randomForest v4.6-14 R package[60]).

Step 4: Robustness classification

To validate the robustness of our clustering, we re-applied Step 2 on our discovery set and on the final signature.

Step 5: Classification of discordant patients

Patients assigned to different groups with the 3 clustering methods were assigned to one of the 4 clusters by applying a distance-to-centroid method based on Pearson correlation.

Molecular subgroup validation. Validation datasets were independently classified in the pSS molecular subgroups by applying a classical distance-to-centroid approach based on correlation. Following the same approach, HV did not constitute a separate cluster but mainly matched with C2 (0.5% in C1, 93% in C2, 4% in C3, and 2.5% in C4) pSS molecular subgroups by applying a classical distance-to-centroid approach based on correlation. The final clustering (without HV) is represented with heatmap using the Heatmap function from ComplexHeatmap (v2.6.2) R package. Clusters are separately constrained for better visualization. This method allows to spotlight heterogeneous intra-clusters. The principal component analysis (PCA) representation will explore the clearly defined clusters and the matching between C2 and HV.

Half of the pSS patients was treated with either anti-malarial, immunosuppressant, or steroids at the time of the visit. When compared to the 3 other clusters, we observed higher proportion of treated patients in C4. To investigate the impact of the treatment on the clustering, we compared treated patients and untreated patients. For this, we apply a hierarchical clustering on treated patients and untreated patients and compare the cluster distribution. The heatmap (Supplementary Fig. 4) of treated vs untreated patients were highly similar which shows that the final clustering is not driven by treatments.

Enrichment analysis. Enrichment analysis was performed by applying a two-tailed Fisher-exact test[61] against different sources of gene modules or pathways: (i) 3 strongly upregulated IFN-annotated modules from[10] (M1.2, M3.4, and M5.12) determined from peripheral blood transcriptomic data with for each a distinct activation threshold, (ii) genes preferentially induced by IFNα or IFNγ identified by[10], (iii) canonical pathway from Ingenuity Pathway Analysis (IPA, Release Date: 2020-06-01), (iv) repertoire recently established on an expended range of disease and pathological states (382 transcriptome modules based on genes co-expression patterns across 16 diseases and 985 unique transcriptome profiles) by [9].

Differential gene expression analysis. To identify genes differentially expressed between pSS subgroups and HV, we performed a linear model (lmFit function from limma v3.46.0 R package[62]) on vst transformation gene expression dataset. Resulting p-values were adjusted for multiple hypothesis testing and filtered to retain DE genes with FDR adjusted p-value ≤ 0.05 and a |Fold-Change (FC)| ≥ 1.5.

Genome-wide association study. Genome-wide association studies (GWAS) were performed for each pSS subgroups (C1: 101, C2: 77, C3: 88, and C4: 38) versus 330 HV. After DNA extraction, the samples were genotyped using HumanCore-24 v1.0 and Infinium CoreExome-24 v1.2 genome-wide SNP genotyping platform (Illumina Inc., San Diego, CA, USA). Individuals were excluded on the basis of incorrect gender assignment, high missingness (>10%), non- European ancestry (<55% using Frappe15 and REAP), and high relatedness (PLINK v1.9[45], pi_hat >0.5)[63]. Genotypes were filtered before imputation due to high missingness (>2%), Hardy–Weinberg equilibrium (HWE) < 0.001, minor allele frequency (MAF) <1%, and AT/CG changes with MAF >40%. PLINK v1.9[45] was used to carry out quality control (QC) measures, genotype data filtering. The basic association for a cluster trait locus, based on comparing allele frequency between patients from each cluster vs HV, was also obtained with this toolset thanks to computational resources from the Roscoff Bioinformatics platform ABiMS. Genotypes were phased using Eagle v2.3 and imputed using Minimac3 against the HRC v1.1 Genomes reference panel from the Michigan Imputation Server platform. Genotypes were filtered after imputation to have HWE p-value > 0.001, MAF > 1% and imputation info score > 0.7 and resulted in 6,664,685 imputed genotypes. Statistical analysis of association for each cluster versus HV was performed by logistic regression under the additive allelic model. The GWAS significant level was fixed at p-value < 5×10^{-8}. SNP annotations and Manhattan plot were obtained using the web-based tool SNP snap from the Broad Institute[64] and qqman (v0.1.8)[65] R packages respectively.

Methylation. Whole blood methylation analysis was performed for 226 pSS patients (C1: 81, C2: 57, C3: 62, and C4: 26) and 175 healthy volunteers (HV). DNA was extracted using a magnetic-bead nucleic acid isolation protocol (Chemagic DNA Blood Kit special, CHEMAGEN) automated with chemagic Magnetic Separation Module I (PerkinElmer) from K2EDTA blood tube (lavender cap, BD Vacutainer) of 10 ml (extractions were performed on 3 ml). 2 µg of DNA were sent for DNA methylation assay. The samples were analyzed using Infinium Human Methylation 450 K BeadChip (Illumina, Inc., San Diego, CA, USA) which covers more than 400,000 CpG sites. DNA samples were bisulfite-converted using the EZ DNA methylation kit (Zymo Research, Orange, CA, USA). After bisulfite conversion, the remaining assay steps were performed following the specifications recommended by the manufacturer. The array was hybridized using a temperature gradient program, and arrays were imaged using a BeadArray Reader (Illumina Inc., San Diego, CA, USA). Sample QC and functional normalization were completed using minfi (v3.3) R package[66]. Briefly, during QC steps, subjects were removed based on outliers for methylated vs unmethylated signals, deviation from mean values at control probes, and high proportion of undetected probes (using minfi default parameters). DNA methylation probes that overlapped with SNPs (dbSNPs v147), located in sexual chromosomes or considered cross-reactive were removed. Additionally, only probes quality controlled and shared between both arrays were used in the subsequent analysis (368,607 probes). Measure of methylation level (B values) were produced for each CpG probe and ranged from 0 (0% molecules methylated at a particular sites) to 1 (100% molecules methylated).

To identify differentially methylated positions (DMPs) between HV and each pSS subgroups (C1 to C4), the champ.DMP function of ChAMP (v2.18.3) R package[67] was implemented doing pairwise comparison between each cluster and HV. Many Δ-beta thresholds were described in the literature and the most frequently used for whole blood studies in autoimmune diseases were 0.05 (5% difference) and 0.1 (10% difference). In order to fix the best threshold for our study, we tested the values of 0.05, 0.075, 0.1, and 0.15 for the absolute ΔBeta. Supplementary Data 11 presents the numbers of DMPs and genes obtained with these different thresholds.

Then, we decided to analyze the data in two steps: the first step with a significant adjusted p-value (Benjamini Hochberg) at 0.1 and an absolute ΔBeta > 0.075. We assumed that a threshold of 0.05 was too low and it would have been

very difficult to interpret the signification of these defects in methylation for C4. If we had applied a ΔBeta threshold of 0.1 in the first intention, we could have missed DMPs. In the second step in order to identify the most robust and significant signature of hypo and hyper methylated genes, a significant adjusted *p*-value (Benjamini Hochberg) at 0.1 and an absolute ΔBeta > 0.15 were applied.

For network viewing, we tested gene lists onto the STRING 9.1 Network of Known and Predicted Protein–Protein Interactions (http://string-db.org/)[68].

Flow cytometry. Multi-parameter flow cytometry analyses have been performed in eleven different centers from the PRECISESADS consortium. Therefore, the integration of all data in common bioinformatical and biostatistical investigations has required a fine mirroring of all instruments[54]. The calibration procedure elaborated to achieve this prerequisite and the antibody panels used have been previously described[53].

The antibody panels, specificities, and clones used are shown in Supplementary Fig. 15a.

The strategy developed to avoid any redundancy in the different cell subsets and to increase the accuracy of the phenotypes has been automated by AltraBio (Lyon, France). The generated automatons have been validated in a preliminary study on 300 patients comparing data from automated gating to data manually gated by the same operator (coefficient of correlation 0.9996). The gating strategy was as follows: after exclusion of debris, dead cells and doublets, frequencies and absolute numbers of $CD15^{hi}CD16^{hi}$ neutrophils, $CD15^{hi}CD16^{+}$ eosinophils, $CD14^{+}CD15^{hi}$ LDGs, $CD14^{++}CD16^{-}$ classical monocytes, $CD14^{++}CD16^{+}$ intermediate monocytes, $CD14^{+}CD16^{++}$ non classical monocytes, $CD3^{+}$ T cells (with $CD4^{+}CD8^{-}$, $CD4^{+}CD8^{+}$, $CD4^{-}CD8^{+}$, $CD4^{-}CD8^{-}$ T cell subsets), $CD19^{+}$B cells, $CD3^{-}CD56^{+}$ NK cells (with $CD16^{lo}CD56^{hi}$ and $CD16^{hi}CD56^{lo}$ NK cell subsets), $CD3^{+}CD56^{+}$ NK-like cells, Lin-HLA-DR^{+} DCs (with $CD11c^{-}CD123^{+}$ pDCs, $CD11c^{+}CD123^{-}$ mDCs (with $CD141^{-}CD1c^{+}$ mDC1, $CD141^{+}CD11c^{-}$ mDC2 and $CD141^{-}CD1c^{-}$ mDC subsets)) and $CD123^{+}$HLA-DR^{-} basophils were automatically extracted from FCS and LMD files of 283 patients and 309 HV and sent in an Excel flow cytometry workflow. The mean distribution of blood cell subsets in frequency (0–100%) and absolute numbers by clusters are compared using a Kruskal–Wallis test.

Gating strategies of the automatons are shown in Supplementary Fig. 15b. For all instruments, the data from the flow cytometry files are analyzed with a similar strategy by one automaton for panel 1 and another automaton for panel 2, and then specifically for each instrument from the gate [S4] to account for the variability of FSC and SSC signals. The desired cell populations are identified by gating strategies identical for all instruments for panel 1 and panel 2 stainings. The mean distribution of blood cell subsets in frequency and absolute numbers are shown in Supplementary Data 12 and 13, respectively.

Cytokines. Cytokines were measured on serum samples. CXCL13/BLC, FAS Ligand, GDF15, CXCL10/IP-10, CCL8/MCP-2, CCL13/MCP-4, CCL4/MIP-1β, MMP-8, CCL17/TARC, IL-1 RII, TNF RI, and IL1-Ra were measured using the Luminex system. The 12-analyte customized panel was built using human pre-mixed multi-analyte Luminex assay (R&D Systems). Samples were thawed on the day of analysis and tested in batches. Soluble MMP-2, CRP, TNFα, IL-6, BAFF, and TGFβ were measured using ELISA assay. Descriptive statistics are shown in Supplementary Data 14. We measured levels of IFNα in plasma using Simoa Single Molecule Array Technology. Results were calculated referring to a standard curve created using a four parameters logistic curve fit and were expressed as pg/ml. For more technical details on sample and data collection, please refer to the main PRECISESADS study[5]. The differential cytokine concentration between subgroups vs HV was performed using a one-way ANOVA followed by post-hoc Tukey's test (function ghlt from multicomp multcomp v1.4-13 R package[69]). The z-score indicate the direction of the concentration between the cluster and the HV. A z-score > 0 means that the cluster has an overexpression compare to HV. A z-score < 0 means that the cluster has a lower expression compare to HV (Fig. 6). Concentration distribution by subgroup is represented in Supplementary Fig. 8. Two-tailed pairwise Wilcoxon-rank sum tests have been computed.

Autoantibodies. Autoantibodies (Extractable nuclear antigen antibodies, anti-SSA antibodies, anti-SSA antibodies (Ro-52), anti-SSA antibodies (Ro-60), Anti-SSB antibodies), were measured in serum using an automated chemiluminescent immunoanalyzer (IDS-iSYS). After processing, the final result is indicative of the concentration of the specific autoantibody present in the sample. Rheumatoid factor (RF), complement C3c, C4, and individualized (kappa, lambda) free light chains (Combilite and freelight, respectively) were measured in serum using a turbidimetric immunoassay method according to manufacturer's recommendations (SPAPLUS analyser). For more technical details on sample and data collection, please refer to the main PRECISESADS study[5]. Autoantibodies and RF distribution have been described by concentration level (Negative/Low/Medium/Elevated/High) and a Fisher's exact test was applied to compare the proportion and the concentration across the 4 clusters. Complements C3 and C4 and circulating free light chains have been described in continued concentration expressed in g/L and mg/L respectively and a Krukal–Wallis test was applied to compare the concentration level across the 4 clusters. Descriptive statistics are described in Supplementary Data 8.

Clinical data. Clinical data on 304 patients with pSS and 330 HV describing the disease phenotype was collected using an electronic case report form (eCRF). A working group of experts on systemic autoimmune diseases was established and the desired items were selected via a Delphi technique. A final set of items was created, digitalized and pilot tested divided into 8 domains (constitutional symptoms, gastrointestinal, vascular, heart and lung, nervous system, skin and glands, musculoskeletal, therapy). After the confirmation of patient inclusion, clinical data were collected including patient's age, sex, ethnicity, dates of first disease manifestation (disease onset), clinical and biological characteristics at baseline, the physician global assessment of disease activity, comorbidity, and current use of treatments.

Another working group of pSS pathology experts was established to select pSS disease-specific items, mainly pSS disease activity scales like ESSDAI and its components, and ESSPRI and its components. These items were collected on a pSS sub-population ($n = 193$).

To characterize pSS subgroups, association test was performed with clinical data. A two-tailed Fisher's exact test (fisher.test function from stats R package) or chi-square test (chisq.test function from stats R package) as appropriate was applied to evaluate the association between the pSS supbgroups and a qualitative clinical factor. A Kruskal–Wallis test (Kruskal.Wallis function from stats R package) was used to evaluate the association between pSS subgroups and quantitative clinical variables.

Development of the composite model for cluster prediction. This feature selection process is composed of two distinct parts: (i) identify a subset of genes potentially interesting to predict the 4 clusters, (ii) use these previously identified subsets to actually craft a prediction model and extract the features used by the model to increase its precision. In the first part, with FC ≥1.5 and FDR ≤0.05, we selected the DEGs according to the following 7 combinations: C2 vs C1, C3 vs C1, C4 vs C1, C4 vs C2, C3 vs C2, C4 vs C2, C4 vs C3. We identified 14,240 and selected those common to all combinations representing 1154 DEGs.

We used the Boruta algorithm[70] on all dataset (discovery and validation sets) to extract features that significantly contributed to predict the patient's cluster.

The algorithm started to extend the dataset by adding copies of each feature in the original dataset. These features were called "shadow features" and consisted in random permutation of the modality of the original feature, in order to remove any correlation with the target variable, in our case, the cluster assignment. Once shadow features were crafted, a random forest classifier was run on the whole dataset and z-scores were computed for all features (real and "shadow"). Shadow features were then sorted according to their z-score and the maximum score was kept in memory as a threshold. The algorithm assigned a hit to each real feature that had a z-score above this threshold. Finally, Boruta marked the features which had a z-score significantly lower than the shadow with maximum z-score as "unimportant" and removed them from the dataset, before removing all shadow features and returning a clean dataset.

This process allowed us to identify variables in the dataset that were significantly more contributing to the classification problems than noisy variables and random artefacts emulated by the original variable modality permutation, ensuring the use of robust features for the second step of our feature selection strategy.

The relatively small size and heterogeneity of C4 in comparison to the other clusters can impact the feature selection process, therefore we chose to solve two classification problems: (i) identify C4 versus all clusters, (ii) discriminate between C1, C2, and C3.

The operation was performed twice: one to predict C4 cluster versus all other clusters and one to discriminate between C1, C2, and C3. In both cases, the algorithm ran over 100 iterations with a max depth of 5 and balanced classes for initializations of random forests.

The two sets of selected features were respectively composed of 255 genes for the C4 prediction dataset and 597 genes for the C1, C2, and C3 prediction dataset.

We then used xgboost-tree[71] approach, to train a model on the dataset with a binary logistic objective function to predict C4 vs all (using the 255 genes previously identified by the Boruta algorithm) and to extract features that have been used by the algorithm to craft the decision tree of the model.

The model can be summarized by $\hat{y}_i = \sum_{k=1}^{K} f_k(x_i), f_k \in F$ where $\hat{y}_i$ is the cluster prediction for the patient i, x_i the vector describing the patient i (composed of the selected features), F the set of estimators for the model (4 in our case, one for each cluster) and K the number of trees by estimator which is 3 for C4 and 4 for C1, C2, and C3. In this context, f_k refers to the tree number k of the estimator f where $f \in F$. K has been manually refined in order to find a compromise between good predictive performance and a low complexity model.

We performed the same approach with a softmax objective function in a multi-classification context to predict the C1, C2, and C3 cluster based on the 597 features previously highlighted by Boruta for this specific classification problem.

The final sets of selected features were composed of 10 genes for the C4 prediction model and 31 genes for the multi-classification (C1, C2, or C3) model (Supplementary Fig. 10). The accuracies of the models, during the training phase perform on the validation set (Table 1) were 94.81% for the C4 prediction model and 96.72% for the multi-classification model.

We then created a composite model, using the combinatorial results of the C4 predictor model and the multi-classification model to predict all 4 clusters on the patients of the discovery set.

Patients were first evaluated by the C4 predictor model. If C4 was not assigned, the patients were evaluated by the multi-classification model.

In order to allow our model to process other cohorts of patients we implemented an interpolation function described by (2). We selected 6 genes with FC ≤ 1.1 and FDR ≥ 0.05 based on their constant expression across all 4 clusters and HV. Their expressions were between 4 and 14 vst normalized counts [*SPIRE* (4), *NUP210L* (6), *GATAD1* (8), *HVCN1* (10), *ENO* (12), and *FLNA* (14)] (Supplementary Fig. 13). This set of genes was denoted *G*. The interpolated value of a gene *x*, $I(x)$ was computed as $I(x) = I(a) + (I(b) - I(a)) \times \frac{x-a}{b-a}$ with a and b representing the vst normalized expression value of two genes such as genes $a, b \in G$, $a < x < b$ and $b \neq a$.

The composite model is integrated into an analysis tool available[33] and the pseudocode description is reported in Supplementary Fig. 16.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

All data included in our study is available upon request at ELIXIR Luxemburg, except the GWAS data that cannot be anonymized, with the permanent link: https://doi.org/10.17881/th9v-xt85 and access procedure is described on the ELIXIR data landing page. The PRECISESADS Consortium committed to secure patient data access through the ELIXIR platform. This commitment was formerly given by written to all patients at the end of the project and to the involved Ethical Committees. The future use of the Project database was framed according to the scope of the patient information and consent forms, where the use of patient data is limited to scientific research in autoimmune diseases. ELIXIR reviews applicants requests and prepares Data Access Committee's decisions on access to Data, communicates such decisions to the Data Providers, who have 10 days to exercise their right to veto; otherwise access is granted to the User.

Code availability

Except when indicated, data analyses were carried out using either an assortment of R system software (http://www.R-project.org, V4.0.1) packages including those of Bioconductor or original R code. R packages are indicated when appropriate. For GWAS analysis, we used Plink, an open-source whole genome association analysis toolset. Machine learning approaches were carried out using python programs (v3.8.5) based on the following modules: scikit-learn, numpy, and xgboost. The composite model designed to predict the patient's cluster is integrated into an analysis tool available on the laboratory's github repository at the following address: https://lbai-infolab.github.io/SjTree/(33).

Received: 2 November 2020; Accepted: 30 April 2021;
Published online: 10 June 2021

References

1. Brito-Zerón, P. et al. Sjögren syndrome. *Nat. Rev. Dis. Primers* **2**, 16047 (2016).
2. Baldini, C. et al. Primary Sjogren's syndrome as a multi-organ disease: impact of the serological profile on the clinical presentation of the disease in a large cohort of Italian patients. *Rheumatology (Oxford)* **53**, 839–844 (2014).
3. Qin, B. et al. Epidemiology of primary Sjögren's syndrome: a systematic review and meta-analysis. *Ann Rheum. Dis.* **74**, 1983–1989 (2015).
4. Goules, A. V. & Tzioufas, A. G. Primary Sjögren's syndrome: clinical phenotypes, outcome and the development of biomarkers. *Autoimmun. Rev.* **15**, 695–703 (2016).
5. Barturen, G. et al. Integrative analysis reveals a molecular stratification of systemic autoimmune diseases. *Arthritis Rheumatol.* (2020) https://doi.org/10.1002/art.41610 https://doi:10.17881/th9v-xt85.
6. Li, H., Ice, J. A., Lessard, C. J. & Sivils, K. L. Interferons in Sjogren's syndrome: genes, mechanisms and effects. *Front Immunol.* **4**, 290 (2013).
7. Bennett, L. et al. Interferon and granulopoiesis signatures in systemic lupus erythematosus blood. *J. Exp. Med.* **197**, 711–723 (2003).
8. Guedj, M. et al. A refined molecular taxonomy of breast cancer. *Oncogene* **31**, 1196–1206 (2012).
9. Rinchai, D. et al. BloodGen3Module: blood transcriptional module repertoire analysis and visualization using R. *Bioinformatics* **btab121**, 1–8, https://doi.org/10.1093/bioinformatics/btab121 (2021).
10. Chiche, L. et al. Modular transcriptional repertoire analyses of adults with systemic lupus erythematosus reveal distinct type I and type II interferon signatures. *Arthritis Rheumatol.* **66**, 1583–1595 (2014).
11. Bodewes, I. L. A. et al. Systemic interferon type I and type II signatures in primary Sjögren's syndrome reveal differences in biological disease activity. *Rheumatology (Oxford)* **57**, 921–930 (2018).
12. Kirou, K. A. et al. Coordinate overexpression of interferon-alpha-induced genes in systemic lupus erythematosus. *Arthritis Rheum.* **50**, 3958–3967 (2004).
13. Lessard, C. J. et al. Variants at multiple loci implicated in both innate and adaptive immune responses are associated with Sjögren's syndrome. *Nat. Genet.* **45**, 1284–1292 (2013).
14. Li, Y. et al. A genome-wide association study in Han Chinese identifies a susceptibility locus for primary Sjögren's syndrome at 7q11.23. *Nat. Genet.* **45**, 1361–1365 (2013).
15. Le Pottier, L., Amrouche, K., Charras, A., Bordron, A., Pers, J-O. Sjögren's syndrome. In: Martín, J., Carmona, F. (eds) Genetics of Rare Autoimmune Diseases. Rare Diseases of the Immune System. (Springer, Cham, 2019), https://doi.org/10.1007/978-3-030-03934-9_4.
16. Morris, D. L. et al. MHC associations with clinical and autoantibody manifestations in European SLE. *Genes Immun.* **15**, 210–217 (2014).
17. Wisskirchen, C., Ludersdorfer, T. H., Müller, D. A., Moritz, E. & Pavlovic, J. The cellular RNA helicase UAP56 is required for prevention of double-stranded RNA formation during influenza A virus infection. *J. Virol.* **85**, 8646–8655 (2011).
18. Tong, Y. et al. Enhanced TLR-induced NF-κB signaling and type I interferon responses in NLRC5 deficient mice. *Cell Res.* **22**, 822–835 (2012).
19. Naomi, M. et al. MXA as a clinically applicable biomarker for identifying Type 1 interferon in primary Sjogren's syndrome. *Ann. Rheum. Dis.* **73**, 1052–1059 (2014).
20. Irizarry, R. A. et al. The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores. *Nat. Genet.* **41**, 178–186 (2009).
21. Imgenberg-Kreuz, J. et al. Genome-wide DNA methylation analysis in multiple tissues in primary Sjögren's syndrome reveals regulatory effects at interferon-induced genes. *Ann. Rheum. Dis.* **75**, 2029–2036 (2016).
22. Fabregat, A. et al. Reactome pathway analysis: a high-performance in-memory approach. *BMC Bioinf.* **18**, 142 (2017).
23. Davies, R. et al. Patients with primary Sjögren's syndrome have alterations in absolute quantities of specific peripheral leucocyte populations. *Scand. J. Immunol.* **86**, 491–502 (2017).
24. d'Arbonneau, F. et al. BAFF-induced changes in B cell antigen receptor-containing lipid rafts in Sjögren's syndrome. *Arthritis Rheum.* **54**, 115–126 (2006).
25. Mukherjee, R. et al. Non-Classical monocytes display inflammatory features: validation in sepsis and systemic lupus erythematous. *Sci. Rep.* **5**, 13886 (2015).
26. Schleinitz, N., Vély, F., Harlé, J. R. & Vivier, E. Natural killer cells in human autoimmune diseases. *Immunology* **131**, 451–458 (2010).
27. Aramaki, T. et al. A significantly impaired natural killer cell activity due to a low activity on a per-cell basis in rheumatoid arthritis. *Mod. Rheumatol.* **19**, 245–252 (2009).
28. Wildenberg, M. E., van Helden-Meeuwsen, C. G., van de Merwe, J. P., Drexhage, H. A. & Versnel, M. A. Systemic increase in type I interferon activity in Sjögren's syndrome: a putative role for plasmacytoid dendritic cells. *Eur. J. Immunol.* **38**, 2024–2033 (2008).
29. van den Hoogen, L. L. et al. Monocyte type I interferon signature in antiphospholipid syndrome is related to proinflammatory monocyte subsets, hydroxychloroquine and statin use. *Ann. Rheum. Dis.* **75**, e81 (2016).
30. Xourgia, E. & Tektonidou, M. G. Type I interferon gene expression in antiphospholipid syndrome: pathogenetic, clinical and therapeutic implications. *J. Autoimmun.* **104**, 102311 (2019).
31. Wallace, D. J., Gudsoorkar, V. S., Weisman, M. H. & Venuturupalli, S. R. New insights into mechanisms of therapeutic effects of antimalarial agents in SLE. *Nat. Rev. Rheumatol.* **8**, 522–533 (2012).
32. van den Borne, B. E., Dijkmans, B. A., de Rooij, H. H., le Cessie, S. & Verweij, C. L. Chloroquine and hydroxychloroquine equally affect tumor necrosis factor-alpha, interleukin 6, and interferon-gamma production by peripheral blood mononuclear cells. *J. Rheumatol.* **24**, 55–60 (1997).
33. Foulquier N. A new molecular classification to drive precision treatment strategies in primary Sjögren's syndrome, SjTree, https://doi.org/10.5281/zenodo.4643639 (2020).
34. Gottenberg, J. E. et al. Effects of hydroxychloroquine on symptomatic improvement in primary Sjögren syndrome: the JOQUER randomized clinical trial. *JAMA* **312**, 249–258 (2014).
35. Mariette, X. et al. Efficacy and safety of belimumab in primary Sjögren's syndrome: results of the BELISS open-label phase II study. *Ann. Rheum. Dis.* **74**, 526–531 (2015).
36. Devauchelle-Pensec, V. et al. Treatment of primary Sjögren syndrome with rituximab: a randomized trial. *Ann. Intern. Med.* **160**, 233–242 (2014).
37. Bowman, S. J. et al. Randomized controlled trial of rituximab and cost-effectiveness analysis in treating fatigue and oral dryness in primary Sjögren's syndrome. *Arthritis Rheumatol.* **69**, 1440–1450 (2017).
38. Meiners, P. M. et al. Abatacept treatment reduces disease activity in early primary Sjögren's syndrome (open-label proof of concept ASAP study). *Ann. Rheum. Dis.***73**, 1393–1396 (2014).
39. St Clair, E. W. et al. Clinical efficacy and safety of baminercept, a lymphotoxin β receptor fusion protein, in primary Sjögren's syndrome: results from a phase

II randomized, double-blind, placebo-controlled trial. *Arthritis Rheumatol.* **70**, 1470–1480 (2018).
40. Gandolfo, S. & De Vita, S. Emerging drugs for primary Sjögren's syndrome. *Expert Opin. Emerg. Drugs* **24**, 121–132 (2019).
41. Barturen, G., Beretta, L., Cervera, R., Van Vollenhoven, R. & Alarcón-Riquelme, M. E. Moving towards a molecular taxonomy of autoimmune rheumatic diseases. *Nat. Rev. Rheumatol.* **14**, 180 (2018).
42. James, J. A. et al. Unique Sjögren's syndrome patient subsets defined by molecular features. *Rheumatology (Oxford)* **59**, 860–868 (2020).
43. Tarn, J. R. et al. Symptom-based stratification of patients with primary Sjögren's syndrome: multi-dimensional characterisation of international observational cohorts and reanalyses of randomised clinical trials. *Lancet Rheumatol.* **1**, e85–e94 (2019).
44. Carvajal Alegria, G. et al. Epidemiology of neurological manifestations in Sjögren's syndrome: data from the French ASSESS Cohort. *RMD Open* **2**, e000179 (2016).
45. Lewis, I., Hackett, K. L., Ng, W. F., Ellis, J. & Newton, J. L. A two-phase cohort study of the sleep phenotype within primary Sjögren's syndrome and its clinical correlates. *Clin. Exp. Rheumatol.* **37**, S78–S82 (2019).
46. Hillen, M. R. et al. Plasmacytoid DCs from Patients with Sjögren's syndrome are transcriptionally primed for enhanced pro-inflammatory cytokine production. *Front. Immunol.* **10**, 2096 (2019).
47. Nezos, A. et al. Type I and II interferon signatures in Sjogren's syndrome pathogenesis: contributions in distinct clinical phenotypes and Sjogren's related lymphomagenesis. *J. Autoimmun.* **63**, 47–58 (2015).
48. Toro-Domínguez, D. et al. Differential treatments based on drug-induced gene expression signatures and longitudinal systemic lupus erythematosus stratification. *Sci. Rep.* **9**, 15502 (2019).
49. Han, B. K. et al. Neutrophil and lymphocyte counts are associated with different immunopathological mechanisms in systemic lupus erythematosus. *Lupus Sci. Med.* **7**, e000382 (2020).
50. Hemond, C. C., Glanz, B. I., Bakshi, R., Chitnis, T. & Healy, B. C. The neutrophil-to-lymphocyte and monocyte-to-lymphocyte ratios are independently associated with neurological disability and brain atrophy in multiple sclerosis. *BMC Neurol.* **19**, 23 (2019).
51. Devauchelle-Pensec, V. et al. Gene expression profile in the salivary glands of primary Sjögren's syndrome patients before and after treatment with rituximab. *Arthritis Rheum.* **62**, 2262–2271 (2010).
52. Shen-Orr, S. S. et al. Cell type-specific gene expression differences in complex tissues. *Nat. Methods* **7**, 287–289 (2010).
53. Jamin, C. et al. Multi-center harmonization of flow cytometers in the context of the European "PRECISESADS" project. *Autoimmun. Rev.* **15**, 1038–1045 (2016).
54. Le Lann, L. et al. Standardization procedure for flow cytometry data harmonization in prospective multicenter studies. *Sci. Rep* **10**, 11567 (2020).
55. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. J.* **17**, 10–12 (2011).
56. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
57. Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12**, 323 (2011).
58. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biol.* **11**, R106 (2010).
59. Leek, J. T. et al. sva: Surrogate variable analysis. *R package version* **3**, 882–883 (2017).
60. Liaw, A. & Wiener, M. Classification and regression by randomForest. *R News* **2**, 18–22 (2002).
61. Gold, D. L., Coombes, K. R., Wang, J. & Mallick, B. Enrichment analysis in high-throughput genomics—accounting for dependency in the NULL. *Brief Bioinformatics* **8**, 71–77 (2007).
62. Ritchie, M. E. et al. Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47 (2015).
63. Purcell, S. et al. PLINK: a toolset for whole-genome association and population-based linkage analysis. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
64. Johnson, A. et al. SNAP: a web-based tool for identification and annotation of proxy SNPs using HapMap. *Bioinformatics* **15**, 2938–2939 (2008).
65. Turner, S. D. qqman: a R package for visualizing GWAS results using Q-Q and manhattan plots. *J. Open Source Softw.* **3**, 731 (2018).
66. Aryee, M. J. et al. Minfi: A flexible and comprehensive Bioconductor package for the analysis of Infinium DNA Methylation microarrays. *Bioinformatics* **30**, 1363–1369 (2014).
67. Morris, T. J. et al. ChAMP: 450k chip analysis methylation pipeline. *Bioinformatics* **30**, 428–430 (2014).
68. Franceschini, A. et al. STRING v9.1: protein–protein interaction networks, with increased coverage and integration. *Nucleic Acids Res.* **41**, D808–D815 (2013).
69. Hothorn, T., Bretz, F. & Westfall, P. Simultaneous inference in general parametric models. *Biom J.* **50**, 346–363 (2008).
70. Kursa, M. B. & Rudnicki, W. R. Feature selection with the boruta package. *J. Stat. Softw.* **36**, 11 (2010).
71. Chen, T. & Guestrin, C. XGBoost: A Scalable Tree Boosting System. https://doi.org/10.1145/2939672.2939785 (2016).

Acknowledgements

The research leading to these results has received support from the Innovative Medicines Initiative Joint Undertaking under the Grant Agreement Number 115565 (PRECISESADS project), resources of which are composed of financial contribution from the European Union's Seventh Framework Program (FP7/2007–2013) and EFPIA companies' in-kind contribution. LBAI was supported by the Agence Nationale de la Recherche under the "Investissement d'Avenir" program with the Reference ANR-11-LABX-0016-001 (Labex IGO) and the Région Bretagne. The authors would like to particularly express their gratitude to the patients, nurses, technicians and many others who helped directly or indirectly in the consecution of this study. They are grateful to the Institut Français de Bioinformatique (ANR-11-INBS-0013), the Roscoff Bioinformatics platform ABiMS (http://abims.sb-roscoff.fr) for providing computing and storage resources and the Hypérion platform at LBAI (Brest, France) for flow cytometry facilities. Finally, this work is now supported by ELIXIR Luxembourg via its data hosting service.

Author contributions

P.S., C.L.D., E.D., B.C., S.H., and C.B. performed the computational studies and carried out the analysis, N.F. performed the computational studies and developed the composite model, C.J., G.B., G.D., PRECISESADS Flow Cytometry Consortium, E.B., J.M., A.B., Z. M. R.L., M.O.B. performed the experimental studies. V.D.P., D.C., A.S., S.J.J., N.B.P., I.R. P., E.D.L., L.B., C.C., L.K., T.W., and PRECISESADS Clinical Consortium contributed to the recruitment of patients. S.C.G., L.X., M.G., P.M. contributed to the edition of the manuscript, MEAR supervised the PRECISESADS consortium, L.L. and J.O.P. supervised the work and wrote the manuscript. All the authors have approved the content of this paper and its related supplementary files and have agreed to the Nature Communications submission policies.

Competing interests

While engaged in the research project, R.L., F.M., and Z.M. were regular employees of Bayer A.G. At present, R.L. and Z.M. are regular employees of Nuvisan ICB GmbH, a company providing contract research services. P.S., S.H., S.C.G., L.X., M.G., P.M., and L. L. were regular employees of Institut de Recherches Internationales Servier at the time of the research project. B.C., C.B., and E.D. were phD students financed by Institut de Recherches Internationales Servier when they contributed to the research project. All other authors confirmed signing the ICMJE form for Disclosure of Potential Conflicts of Interest and none of them have any conflict of interest related to this work.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-021-23472-7.

Correspondence and requests for materials should be addressed to J.-O.P.

Peer review information *Nature Communications* thanks A. Darise Farris, Zhan-Guo Li, and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

Reprints and permission 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.

PRECISESADS Clinical Consortium

Lorenzo Beretta[8], Barbara Vigone[8], Jacques-Olivier Pers[2,3], Alain Saraux[2,3], Valérie Devauchelle-Pensec[2,3], Divi Cornec[2,3], Sandrine Jousse-Joulin[2,3], Bernard Lauwerys[15], Julie Ducreux[15], Anne-Lise Maudoux[15], Carlos Vasconcelos[16], Ana Tavares[16], Esmeralda Neves[16], Raquel Faria[16], Mariana Brandão[16], Ana Campar[16], António Marinho[16], Fátima Farinha[16], Isabel Almeida[16], Miguel Angel Gonzalez-Gay Mantecón[17], Ricardo Blanco Alonso[17], Alfonso Corrales Martínez[17], Ricard Cervera[6], Ignasi Rodríguez-Pintó[6], Gerard Espinosa[6], Rik Lories[7], Ellen De Langhe[7], Nicolas Hunzelmann[18], Doreen Belz[18], Torsten Witte[11], Niklas Baerlecken[11], Georg Stummvoll[19], Michael Zauner[19], Michaela Lehner[19], Eduardo Collantes[5], Rafaela Ortega-Castro[5], Ma Angeles Aguirre-Zamorano[5], Alejandro Escudero-Contreras[5], Ma Carmen Castro-Villegas[5], Yolanda Jiménez Gómez[5], Norberto Ortego[20], María Concepción Fernández Roldán[20], Enrique Raya[21], Inmaculada Jiménez Moleón[21], Enrique de Ramon[22], Isabel Díaz Quintero[22], Pier Luigi Meroni[13], Maria Gerosa[13], Tommaso Schioppo[13], Carolina Artusi[13], Carlo Chizzolini[9], Aleksandra Zuber[9], Donatienne Wynar[9], Laszló Kovács[10], Attila Balog[10], Magdolna Deák[10], Márta Bocskai[10], Sonja Dulic[10], Gabriella Kádár[10], Falk Hiepe[23], Velia Gerl[23], Silvia Thiel[23], Manuel Rodriguez Maresca[24], Antonio López-Berrio[24], Rocío Aguilar-Quesada[24], Héctor Navarro-Linares[24], Yiannis Ioannou[25], Chris Chamberlain[26], Jacqueline Marovac[26], Marta Alarcón Riquelme[4] & Tania Gomes Anjos[4]

[15]Pôle de pathologies rhumatismales systémiques et inflammatoires, Institut de Recherche Expérimentale et Clinique, Université catholique de Louvain, Brussels, Belgium. [16]Centro Hospitalar do Porto, Port, Portugal. [17]Servicio Cantabro de Salud, Hospital Universitario Marqués de Valdecilla, Santander, Spain. [18]Klinikum der Universitaet zu Koeln, Cologne, Germany. [19]Medical University Vienna, Vienna, Austria. [20]Complejo hospitalario Universitario de Granada (Hospital Universitario San Cecilio), Granada, Spain. [21]Complejo hospitalario Universitario de Granada (Hospital Virgen de las Nieves), Granada, Spain. [22]Hospital Regional Universitario de Málaga, Málaga, Spain. [23]Charite, Berlin, Germany. [24]Andalusian Public Health System Biobank, Granada, Spain. [25]UCB Pharma (PRECISESADS Project office), Slough, UK. [26]Chromatin and Disease Group, Bellvitge Biomedical Research Institute (IDIBELL), Barcelona, Spain.

PRECISESADS Flow Cytometry Consortium

Christophe Jamin[2,3], Concepción Marañón[4], Lucas Le Lann[2], Quentin Simon[2], Bénédicte Rouvière[2,3], Nieves Varela[4], Brian Muchmore[4], Aleksandra Dufour[9], Montserrat Alvarez[9], Carlo Chizzolini[9], Jonathan Cremer[7], Ellen De Langhe[7], Nuria Barbarroja[5], Chary Lopez-Pedrera[5], Velia Gerl[23], Laleh Khodadadi[23], Qingyu Cheng[23], Anne Buttgereit[12], Zuzanna Makowska[12], Aurélie De Groof[14], Julie Ducreux[14], Elena Trombetta[8], Tianlu Li[26], Damiana Alvarez-Errico[26], Torsten Witte[11], Katja Kniesch[11], Nancy Azevedo[15], Esmeralda Neves[15], Sambasiva Rao[27], Pierre-Emmanuel Jouve[28] & Jacques-Olivier Pers[2,3]

[27]Sanofi Genzyme, Framingham, MA, USA. [28]AltraBio SAS, Lyon, France.

7.2 Annex 2

Network-based repurposing identifies anti-alarmins as drug candidates to control severe lung inflammation in COVID-19

Emiko Desvaux, Antoine Hamon, Sandra Hubert, **Cheïma Boudjeniba**, Bastien Chassagnol, Jack Swindle, Audrey Aussy, Laurence Laigle, Jessica Laplume, Perrine Soret, Pierre Jean-François, Isabelle Dupin-Roger, Mickaël Guedj, Philippe Moingeon.

PLoS One. 2021 Jul 22;16(7):e0254374. doi: 10.1371/journal.pone.0254374.

Summary

COVID-19 remains a significant public health concern with substantial economic implications. During this transition, repurposing existing drugs stands out as a rapid, cost-effective approach to alleviate the strain on healthcare systems, notably by reducing the incidence of severe COVID-19-associated acute respiratory distress syndrome.

In our research, I took an active role in implementing a computational repurposing strategy to identify potential therapeutic drugs capable of mitigating the progression of severe airway inflammation in COVID-19. We utilized molecular profiling data from various sources, encompassing SARS-CoV-2-infected epithelial and endothelial cells, immune dysregulation associated with severe COVID-19, and inflammation induced by other respiratory viruses. This comprehensive dataset allowed us to construct a protein-protein interactome model, tracing the evolution of lung inflammation from the onset to a fully developed cytokine release syndrome in COVID-19.

Our predictive model, incorporating proteins closely linked to severe COVID -19, highlighted familiar contributors to the cytokine storm, such as IL1-β, IL6, TNF-α, and JAK2, alongside less recognized participants like IL17, IL23, and C5a. Notably, our analysis emphasized the therapeutic potential of alarmins, including TSLP, IL33, members of the S100 family, and their receptors (ST2, RAGE).

By assessing network-based distances between severe COVID-19-related proteins and established drug targets, our computational approach identified drug candidates that could be repurposed to prevent or slow down the progression of severe airway inflammation. This analysis confirmed the efficacy of drugs like dexamethasone and JAK2 inhibitors while also uncovering various available or in-development drugs interacting with these targets.

PLOS ONE

RESEARCH ARTICLE

Network-based repurposing identifies anti-alarmins as drug candidates to control severe lung inflammation in COVID-19

Emiko Desvaux[1☯], Antoine Hamon[2☯], Sandra Hubert[1☯], Cheïma Boudjeniba[1], Bastien Chassagnol[1], Jack Swindle[2], Audrey Aussy[1], Laurence Laigle[1], Jessica Laplume[1], Perrine Soret[1], Pierre Jean-François[1], Isabelle Dupin-Roger[1], Mickaël Guedj[1☯], Philippe Moingeon[1]*

1 Servier, Research and Development, Suresnes Cedex, France, **2** Lincoln, Research and Development, Boulogne-Billancourt Cedex, France

☯ These authors contributed equally to this work.
* philippe.moingeon@servier.com

Citation: Desvaux E, Hamon A, Hubert S, Boudjeniba C, Chassagnol B, Swindle J, et al. (2021) Network-based repurposing identifies anti-alarmins as drug candidates to control severe lung inflammation in COVID-19. PLoS ONE 16(7): e0254374. https://doi.org/10.1371/journal.pone.0254374

Editor: Svetlana P. Chapoval, University of Maryland School of Medicine, UNITED STATES

Received: March 5, 2021

Accepted: June 24, 2021

Published: July 22, 2021

Data Availability Statement: The RNA Seq data on genes differentially expressed in SARS-CoV-2 infected NHBE or Calu-3 human lung epithelial cells are publicly available from repository Gene Expression Omnibus (GEO, accession number GSE147507). Drug-target links can be retrieved from the Therapeutic Target Database (version 7.1.01) and from Drugbank. All other sources of data used in the present study related to various aspects of COVID-19 pathophysiology were obtained from the scientific literature. A

Abstract

While establishing worldwide collective immunity with anti SARS-CoV-2 vaccines, COVID-19 remains a major health issue with dramatic ensuing economic consequences. In the transition, repurposing existing drugs remains the fastest cost-effective approach to alleviate the burden on health services, most particularly by reducing the incidence of the acute respiratory distress syndrome associated with severe COVID-19. We undertook a computational repurposing approach to identify candidate therapeutic drugs to control progression towards severe airways inflammation during COVID-19. Molecular profiling data were obtained from public sources regarding SARS-CoV-2 infected epithelial or endothelial cells, immune dysregulations associated with severe COVID-19 and lung inflammation induced by other respiratory viruses. From these data, we generated a protein-protein interactome modeling the evolution of lung inflammation during COVID-19 from inception to an established cytokine release syndrome. This predictive model assembling severe COVID-19-related proteins supports a role for known contributors to the cytokine storm such as IL1β, IL6, TNFα, JAK2, but also less prominent actors such as IL17, IL23 and C5a. Importantly our analysis points out to alarmins such as TSLP, IL33, members of the S100 family and their receptors (ST2, RAGE) as targets of major therapeutic interest. By evaluating the network-based distances between severe COVID-19-related proteins and known drug targets, network computing identified drugs which could be repurposed to prevent or slow down progression towards severe airways inflammation. This analysis confirmed the interest of dexamethasone, JAK2 inhibitors, estrogens and further identified various drugs either available or in development interacting with the aforementioned targets. We most particularly recommend considering various inhibitors of alarmins or their receptors, currently receiving little attention in this indication, as candidate treatments for severe COVID-19.

comprehensive list of those publications and detailed references are provided in Supporting information file S1 Table.

Funding: We confirm that Servier and Lincoln only provided financial support in the form of authors' salaries. These companies did not play a role in the study design, data collection and analysis, decision to publish, nor in the preparation of the manuscript. The specific roles of these authors are articulated in the 'author contributions' section.

Competing interests: We confirm that our commercial affiliations do not alter our adherence to PLOS ONE policies on sharing data and materials.

Introduction

Since the emergence of the new strain of Coronavirus SARS-CoV-2 in December 2019, the ongoing crisis associated with the COVID-19 disease has affected more than 170 million individuals worldwide, causing over 3.5 million deaths (World Health Organization Dashboard, June 1st, 2021), mainly as the consequence of an Acute Respiratory Distress Syndrome (ARDS). The pandemic is still progressing actively despite lockdown measures throughout the world, with the recent emergence of highly transmissible viral strains [1]. To date, the only proven medications for reducing either viral loads, hospitalization rates, invasive mechanical ventilation or patient mortality include corticosteroids such as dexamethasone, the antiviral remdesivir, the anti-IL6R tocilizumab as well as neutralizing monoclonal antibodies directed to the spike protein of the virus [2–5]. Many additional drugs have been tested, including the lopinavir antiviral, the anti-malarial hydroxychloroquine or IFNβ with as of today disappointing efficacy results [6].

Recently, several vaccines have been approved by regulatory authorities based on remarkable efficacy results, with evidence that they can protect against infection by eliciting high titers of neutralizing antibodies against the Spike protein of the SARS-CoV-2 virus [7]. Whereas such vaccines will very positively transform the course and gravity of the COVID-19 pandemic, a recent concern is whether they will be fully effective against emerging new variants of the virus bearing point mutations in the Spike protein [1]. Furthermore, the challenge of manufacturing and administering billions of vaccine doses in order to establish a protective herd immunity at a worldwide population level will not be met in a short time frame.

During the time needed to deploy preventive vaccines at such a scale, the repurposing of existing drugs is a valid solution to better address severe forms of COVID-19 and alleviate the burden on health services in a time and cost-effective manner. Previous repurposing strategies have been undertaken in the context of a limited understanding of COVID-19 pathogenesis, prompting to use related viruses such as SARS-CoV and MERS-CoV as proxies to model SARS-CoV-2 infection [8–13]. Several network computing studies have been successful to predict drug disease associations for repurposing in COVID-19. Many of those initial approaches were aiming to identify existing compounds to prevent viral infection by either targeting mechanisms involving the viral receptor ACE2 (angiotensin converting enzyme 2), the TMPRSS2 transmembrane protease serine 2, or clathrin-mediated endocytosis [14–16]. In the present repurposing study, we rather focused on drugs predicted to interfere with pro-inflammatory mediators identified by modelling immune dysregulations caused in the airways by SARS-CoV-2 infection.

Since a vast majority of patients infected with SARS-CoV-2 develop no or only mild symptoms, we reasoned that ideal candidate drugs to repurpose should rather inhibit severe airways inflammation in the course of the disease. Lung inflammation is the main cause requiring hospitalization in up to 20% of COVID-19 cases, with life threatening ARDS affecting 75% of COVID-19 patients transferred to intensive care units [17]. In this subset of patients with severe lung inflammation, persisting proinflammatory immune responses result in a cytokine release syndrome (CRS) linked to the activation of myeloid cells secreting cytokines such as IL1β, IL6 and TNFα [18–20].

Capitalizing on the most recent scientific insights on the pathophysiology of COVID-19, we undertook computational network analyses to integrate a wide variety of data sources encompassing extensive molecular profiling of SARS-CoV-2 infected epithelial or endothelial cells, genetic susceptibilities and immune dysregulations linked to severe COVID-19 as well as molecular mechanisms elicited during lung infection by other respiratory viruses. From this approach, a short list of COVID-19 disease-related proteins considered as potential

therapeutic targets was established and used to computationally assess a topological proximity with drug targets within the comprehensive human protein-protein interactome [21, 22]. Herein, we report on the identification of candidate therapeutic targets, as well as drugs predicted to interact with some of those targets which could be repurposed to prevent or slow down severe lung inflammation during COVID-19.

Materials and methods

Sources of data on COVID-19 pathophysiology

To identify proteins related to lung inflammation in COVID-19, we selected relevant categories of data from the scientific literature (detailed in S1 Table), such as genes differentially expressed following SARS-CoV-2 infection of (i) primary normal human bronchial epithelial cells (*NHBE*) or of the ACE2-expressing lung-epithelial *Calu-3* cell line, (ii) endothelial cells or cells recovered from bronchoalveolar lavages or lung biopsies of patients with severe COVID-19 [23–25]. We also mined public data regarding immunological signatures obtained in the blood or in tissues of patients, distinguishing those with mild COVID-19 from others rather affected by severe forms of the disease [26–34]. We included as well information from previous studies on lung inflammation caused by other respiratory viruses (including asthma exacerbation), in light of an involvement of monocytes, macrophages, myeloid dendritic cells, innate lymphoid cells in those conditions similarly to COVID-19 [18, 35–38].

Identification of disease-related proteins

COVID-19 disease-related proteins predicted to be involved in early lung inflammation and in the transition to the cytokine storm were identified following data mining from scientific publications listed in S1 Table. To establish molecular pathways dysregulated during lung inflammation due to COVID-19, we first used RNAseq data from *NHBE* (normal human bronchial epithelial) and *Calu-3* (human lung epithelial cancer) cells infected or not with SARS-CoV-2. These data were pre-treated by removing outlier samples whose total sum of counts was below 5 000 000. In order to filter out genes undistinguishable from background noise, we modelled gene expression after applying a log2(x + 1) transformation by a two component Gaussian mixture model, with a first peak corresponding to unexpressed genes, and the second peak to truly expressed genes. Numbers of genes pre and post-filtering were 17557 and 21797, respectively. We retrieved the parameters of the mixture distribution using function normalmixEM from mixtools package and determined that the 0.95 quantile for the noise distribution was 1.6. We subsequently removed all genes whose expression was below that threshold in more than 95% of samples. We performed a differential analysis (COVID versus mock) in each cell line using the *limma R* package and *eBayes* function (with mock group corresponding to healthy & no treatment patients). Disease signatures were then extracted by considering differentially expressed genes (DEG) as those with adjusted p-value below 0.05 with an absolute fold change superior to 1.3 (commonly used as a threshold for biological significance). Canonical pathway enrichment analyses were subsequently performed by using the Ingenuity Pathway Analysis (IPA) software.

Network-based drug repurposing

Network-based drug repurposing relies on the hypothesis that the closer a target is to a group of disease related genes in the PPI network, the higher the chance of having a significant impact on the disease. Many approaches focus on the shortest path to determine proximity, with some variations in order to avoid hub protein bias [15, 39]. The latter bias occurs from

certain proteins that have an extremely high degree in the network and thereby cause a highly dense graph. Other approaches take advantage of the diffusion process to define proximity [40] while considering all the topological features of the graph. Diffusion based metrics have a comparable advantage over shortest path distances when in highly dense graphs such as PPI graphs [41]. Other metrics distinct from shortest path and diffusion can be used such as such as largest connected component -based methods [42].

Our computational repurposing approach (Fig 1A) takes advantage of the proximity between disease-related proteins and drug targets through an established network of protein-protein interactions (PPIs, referred to as an *interactome*). Drug-target links were gathered from the Therapeutic Target Database (TTD, version 7.1.01) and Drugbank [43, 44]. The PPIs network was derived from previous work by Cheng et al [45]. It was built from 15 different databases such as BioGRID and HPRD by compiling binary PPIs tested by high-throughput yeast-two-hybrid (Y2H) systems, kinase-substrate interactions from literature-derived low-throughput and high-throughput experiments, high-quality PPIs from three-dimensional (3D) protein structures, and signaling networks from literature-derived low-throughput experiments.

Relevance of drugs to the disease was assessed based on proximity of their targets to disease-related proteins according to two complementary metrics, namely a simple *topological distance* and a more advanced *diffusion-based distance*.

The *topological distance* (d_{topo}) corresponds to the shortest path length in the PPIs network between the disease-related proteins and the drug targets, computed according to the following formula:

$$d_{\text{topo}}(P,T) = \frac{1}{\|T\|}\sum_{t\in T}\min_{p\in P} SP(p,t)$$

With P the set of nodes corresponding to the disease-related proteins, T the set of nodes corresponding to the drug targets, and $SP(p,t)$ the shortest path length between a node p of P and another node t of T. When calculating a topological distance, we generate a distribution from bootstrapping similar nodes defined by same degree in the graph. From the given distribution, we calculate a z-score (and p-value).

The *diffusion-based distance* (d_{diff}) is computed based on the similarity of the impact on the network of perturbations starting from disease-related proteins on one side and drug targets on the other. The impact of a perturbation starting from a given node n_i on the network is assessed by use of a *diffusion* algorithm. Let (n_i,n_j) being a pair of nodes, then $\mathbb{P}\left(n_i, n_j\right)$ represents the random walk-based probability that a perturbation starting from n_i reaches n_j. It allows us to define a numerical vector $V(n_i)$ representing the impact perturbation of n_i on the whole interactome:

$$V(n_i) = [\mathbb{P}(n_i,\ n_1),\ \mathbb{P}(n_i,\ n_2),\ \ldots,\ \mathbb{P}(n_i, n_n)]$$

The similarity between two perturbations starting from n_i and n_j is then assessed by computing the Manhattan distance between $V(n_i)$ and $V(n_j)$. In order to extend this principle to the distance between sets of nodes, we derived the following formula:

$$d_{\text{diff}}(P,T) = \frac{1}{\|T\|}\sum_{t\in T}\min_{p\in P} MD(p,t)$$

With P the set of nodes corresponding to the disease-related proteins, T the set of nodes corresponding to the drug targets, p one given node of P, t one given node of T, and MD (p,t)

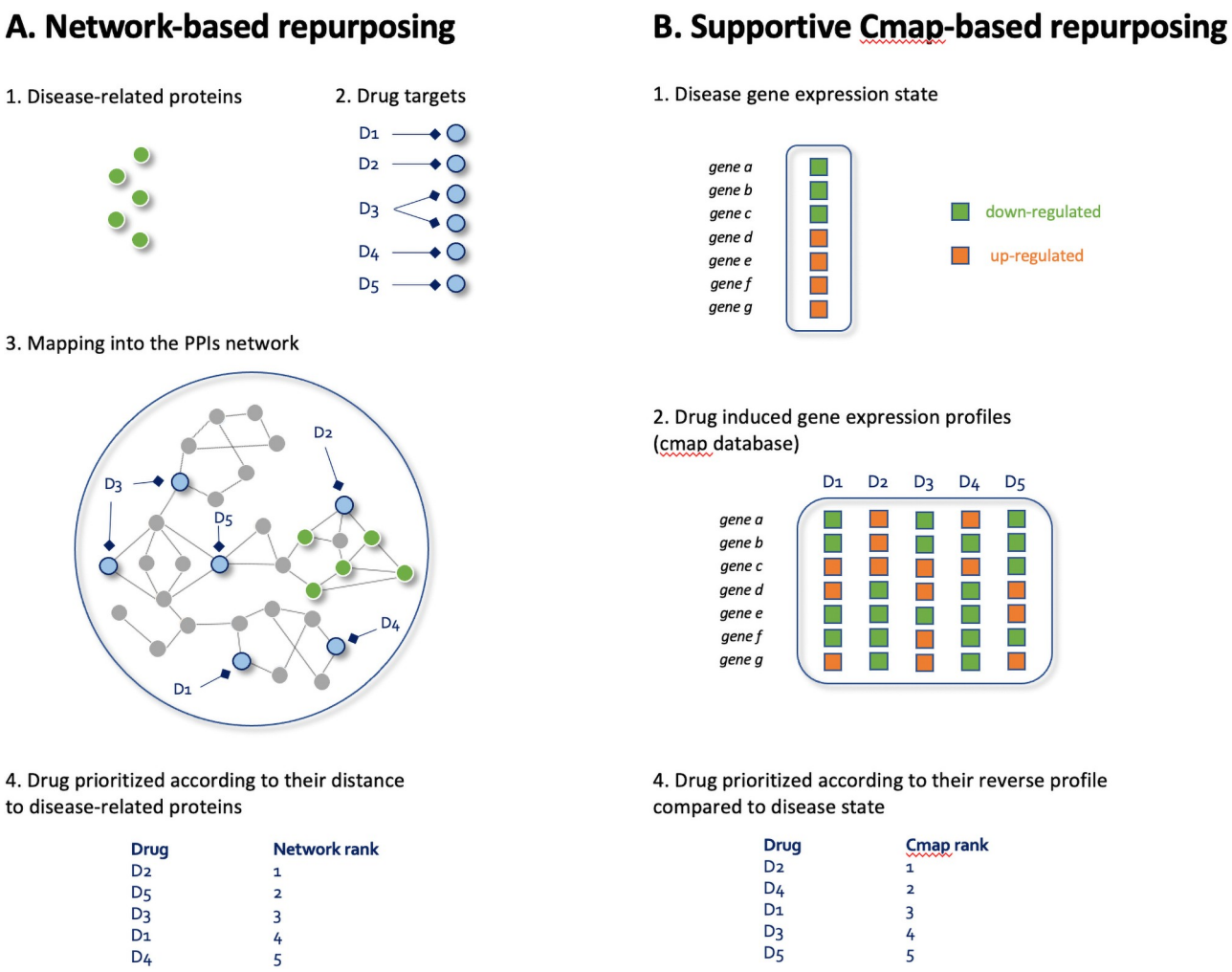


Fig 1. General principles of network and Cmap-based repurposing approaches. A) Network-based repurposing. Disease-related proteins and drug targets are mapped into a network of protein-protein-interactions (PPI). Drugs are prioritized according to their distance to disease-related proteins. B) Supportive Cmap-based repurposing. In those supportive analyses, disease-related as well as drug induced gene expression states are compared in order to identify drugs eliciting reverse profiles compared to those found in the disease.

https://doi.org/10.1371/journal.pone.0254374.g001

the Manhattan distance between $V(p)$ and $V(t)$. This diffusion-based distance was implemented via the DSD algorithm [46]. For each diffusion-based distance, we also calculate associated z-scores (and p-values). Note that DSD is by construction normally distributed. In order to prioritize drugs from this network-based repurposing approach, we defined a network rank resulting from the mean rank aggregation of d_{topo} and d_{diff}. Given that we have p-values for both of our distance measures, we perform a Fisher's combined probability test to obtain a unique combined p-value per drug. Using the DSD algorithm, we generated a computed distance matrix of 15 894 X 15 894 encompassing all proteins in our interactome.

Cmap-based drug repurposing

We complemented the network-based approach by using Cmap as a supportive method (Fig 1B). Cmap identifies drugs inducing a reverse gene expression profile compared to the disease state using a method of similarity [47]. The Cmap database comprises human cancer cell lines either treated or not with chemical drugs, referred to as perturbagens. We used the *R* package

ccdata which encompasses expression profiles for 1309 perturbagens over 13832 genes. Disease state was obtained from gene expression profiles induced in *NHBE* and *Calu-3* cells following infection by SARS-CoV-2. We compare expression profiles induced by disease state with those induced by perturbagens, using mainly the Pearson correlation between transcriptome values of the query signature and the perturbagen signature. A negative correlation score provides a potential therapeutic indication of the perturbagen. Cmap scores (the smaller the better) were first computed on both *NHBE* and C*alu-3* data and then averaged.

Results and discussion

Identification of COVID-19 disease-related proteins

Based on recent scientific advances, the pathophysiology of COVID-19 can be summarized as three sequential steps (Fig 2). We reasoned that treatments suitable to control severe COVID-19 should interfere with molecular pathways involved in the evolution from mild to severe

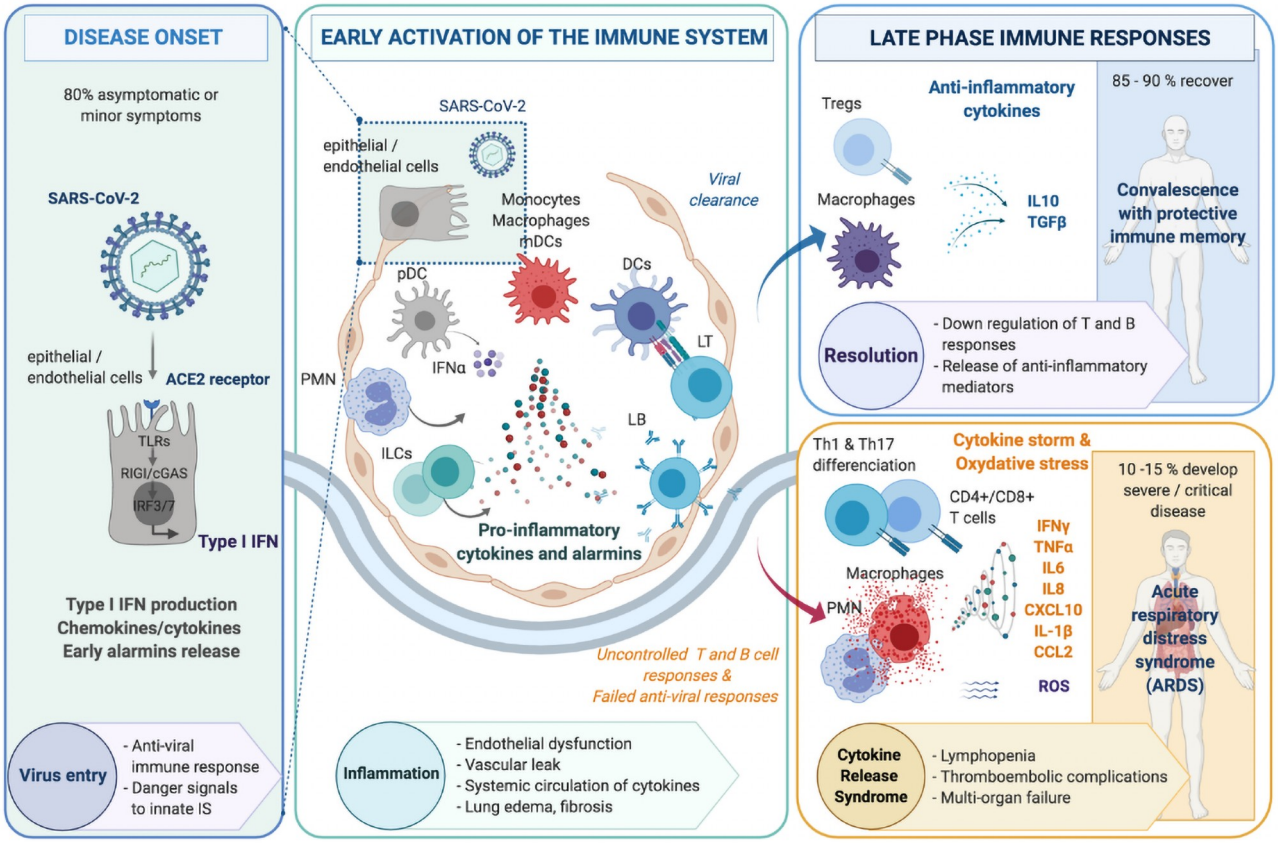


Fig 2. Three step progression towards severe COVID-19. The pathophysiology of COVID-19 in the airways encompasses schematically three successive steps, including **(i)** Disease onset following viral infection of alveolar epithelial or endothelial cells expressing the ACE2 receptor (left panel) leading to the activation of the innate immune system, with IFNα production by plasmacytoid dendritic cells (pDC). **(ii)** An early inflammatory phase within lung tissues where a cross-talk between infected epithelial/endothelial cells and innate immune cells such as monocytes, macrophages, myeloid dendritic cells (mDC) and innate lymphoid cells (ILCs) leads to a release of pro-inflammatory alarmins, cytokines and chemokines (center panel). This results in the activation of adaptive immunity, involving both CD4+ T cell help, CD8+ T cells cytotoxic for virally-infected cells as well as production of neutralizing antibodies against surface viral antigens. **(iii)** A late inflammatory phase with two potential outcomes: 85 to 90% of cases evolve towards resolution of inflammation with downregulation of T and B cell responses concomitant with the release of anti-inflammatory mediators (right upper panel); whereas 10 to 15% patients rather exhibit major tissue damage and severe acute respiratory distress syndrome (ARDS) caused by a deleterious uncontrolled inflammation linked with persisting T cell activation, excessive myeloid cell activation associated with a cytokine storm as well as oxidative stress (right lower panel).

https://doi.org/10.1371/journal.pone.0254374.g002

lung inflammation (Fig 2, central panel), while preserving anti-viral protective immune mechanisms. We thus compiled a comprehensive list of genes differentially upregulated in *NHBE* and *Calu-3* human epithelial cells following SARS-CoV-2 infection, providing important quantitative information [23]. We cross-validated this list in comparison with molecular signatures reported at the level of endothelial cells, bronchoalveolar lavage cells or lung biopsies in other studies to be associated with severe COVID-19 or exposure to other respiratory viruses (S1 Table). The latter was further completed with deep immunophenotyping, RNA seq and cytokine profiling data related to dysregulated innate or adaptive immune responses in the blood or the lungs of patients with severe COVID-19. A compilation of the most relevant COVID-19 disease related-proteins thus obtained, together with data sources supporting their relevance to lung inflammation in COVID-19 are presented in S1 Table.

Ingenuity pathway analyses were then performed on this list, allowing to confirm that genes/proteins upregulated following SARS-CoV-2 infection in the airways belong to multiple well-known pro-inflammatory pathways (Fig 3, S2 Table). Further data interpretation led us to classify disease-related proteins in two distinct sets of highly represented proinflammatory mediators and cytokines termed *Alarmins* and *Cytokine storm*, respectively (S1 Table). Alarmins represent a family of immunomodulatory proteins acting as damage-associated molecular patterns provided by injured stromal cells to recruit and activate various innate immune cells such as monocytes, macrophages, innate lymphoid cells as well as myeloid dendritic cells. Multiple proteins belonging to this family (*i.e.* defensins, HMGB1, IL1α, IL25, IL33, TSLP, S100A4, S100A7, S100A8, S100A9, S100A12, S100B, S100P) as well as their receptors such as IL1R1, RAGE, ST2 were predicted by our model to be involved in the evolution towards severe lung inflammation in COVID-19.

Our study also draws attention on disease-related proteins linked to the cytokine storm occurring in severe forms of COVID-19. The latter includes proinflammatory cytokines produced by activated myeloid cells such as IL1β, IL6 and TNFα directly involved as a cause of the CRS observed in COVID-19 [18, 35, 36]. Other potential targets associated with the cytokine storm include various cytokines (*e.g.* IL1β, IFNγ, IL2, IL12, IL15, IL17, IL23, IL32), chemokines (*e.g.* CCL5, CCL20, CXCL5, CXCL10, CXCL11), as well as selected proinflammatory factors (*e.g.* JAK1, JAK2, C5a) (S1 Table) [19, 20, 26–28, 36, 48–50].

Mapping into the interactome and identification of drug candidates for repurposing

COVID-19 disease-related proteins were mapped in parallel with known drug targets into the human complete interactome made of 15894 proteins (including 951 known drug targets) and 213861 interactions (Fig 4). From this, 3092 drugs were ranked according to computational proximity of their targets to each of the alarmins and cytokine storm sets by using a network-based method (S3 Table). Both COVID-19-related proteins as well as some functionally-related proteins in the interactome (such as the NR3C1 glucocorticoid receptor or receptors for reproductive steroids) were identified as candidate therapeutic targets.

Table 1 provides a list of selected targets as well as drugs interacting with those targets predicted to be of interest in severe COVID-19. Specifically, several high-ranking drugs were identified to treat severe COVID-19, such as anti-IL1β, anti-IL6 and IL6R or anti-TNFα antibodies. Our model supports as well the interest of corticosteroids such as dexamethasone, broadly used currently to treat severe COVID-19 [2]. Other high-ranking candidates for repurposing identified in our study are JAK2 inhibitors, with drugs not yet approved such as momelotinib or gandotinib previously shown by structure-based virtual screening to interact with ACE2 and the SARS-CoV-2 main protease, but also baricitinib, as well as other JAK1/

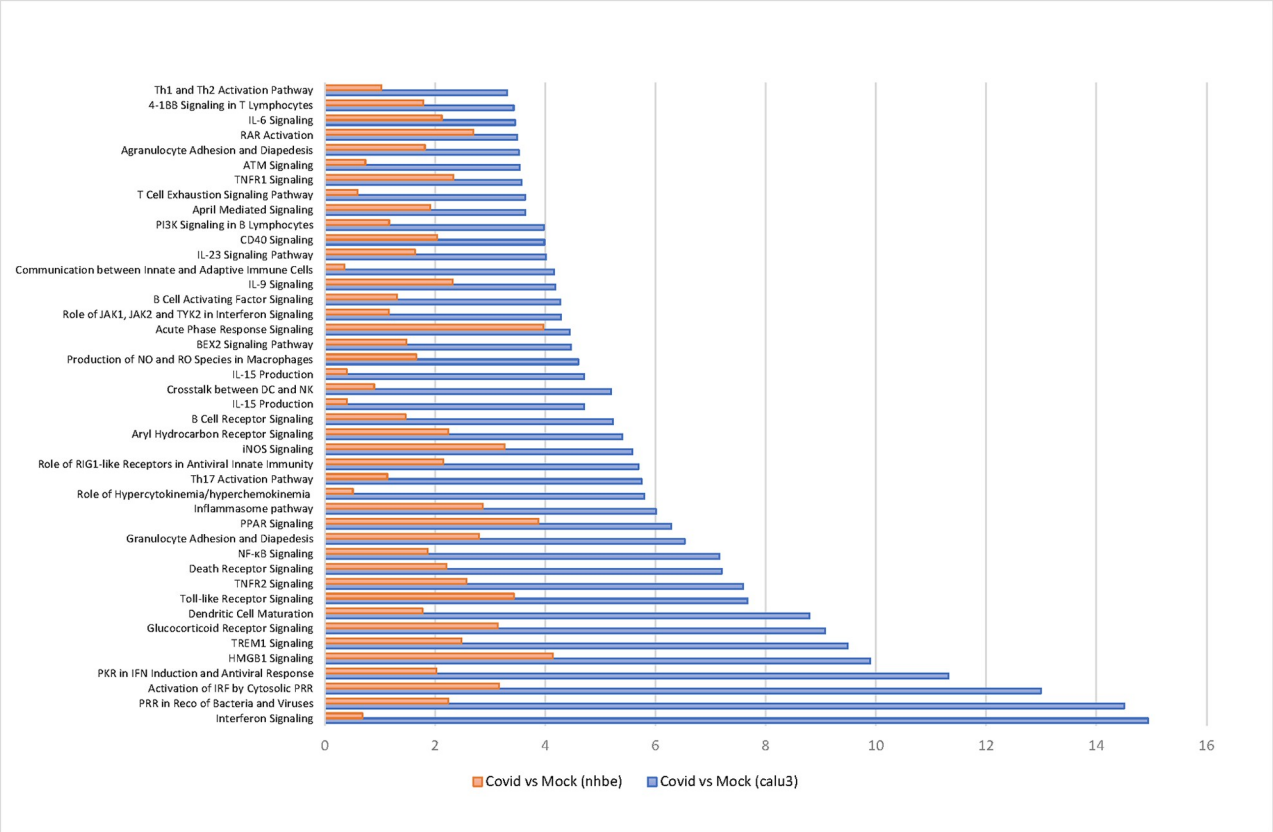


Fig 3. Pathway enrichment analysis from disease signatures (COVID-19 versus mock) in epithelial cell lines infected by SARS-CoV-2. The top 40 most significantly dysregulated immunological canonical pathways in either the Calu-3 (yellow) and NHBE (brown) infected cell lines are represented in a radar plot according to -log (*p*-value). Pathway enrichment *z*-scores, based on fold change direction, represent predicted up-regulation (green dots) or down-regulation (blue dots) for positive or negative values, respectively.

https://doi.org/10.1371/journal.pone.0254374.g003

JAK2 inhibitors currently being evaluated in COVID-19 patients (Table 1). Interestingly, some network computing approaches aiming to repurpose drugs inhibiting cell infection by SARS--CoV-2 also concluded to the interest of blocking antibodies against IL1β, IL6 and TNFα as well as JAK inhibitors in treating COVID-19 patients, in agreement with the present study [15, 16]. In addition, we also identify several reproductive steroids (estrogens and progesterone) as interesting candidates for treating COVID-19 patients.

Whereas the previous targets and some of the drugs directed to them could be expected from the current state of knowledge, our modeling study provided as well interesting hypotheses regarding other therapeutic options receiving less attention as of today. For example, drugs interacting with alarmins were also strongly suggested to be useful in COVID-19. To our knowledge, only three clinical studies have been initiated in COVID-19 with anti-alarmins, despite the availability of multiple additional drug candidates in this class (Table 1). Noteworthy, since Alarmins of the S100 family activate Toll-like receptors such as TLR2 and TLR4, a therapeutic option might be to target specific TLRs downstream of alarmins. Indeed, several TLR-antagonists are currently undergoing clinical evaluation in order to restore immune-homeostasis in patients with COVID-19 [51].

Similarly, anti-IL17 antibodies rank very high in our repurposing analysis, suggesting that inhibitory drugs directed to this well-known pro-inflammatory cytokine as well as the

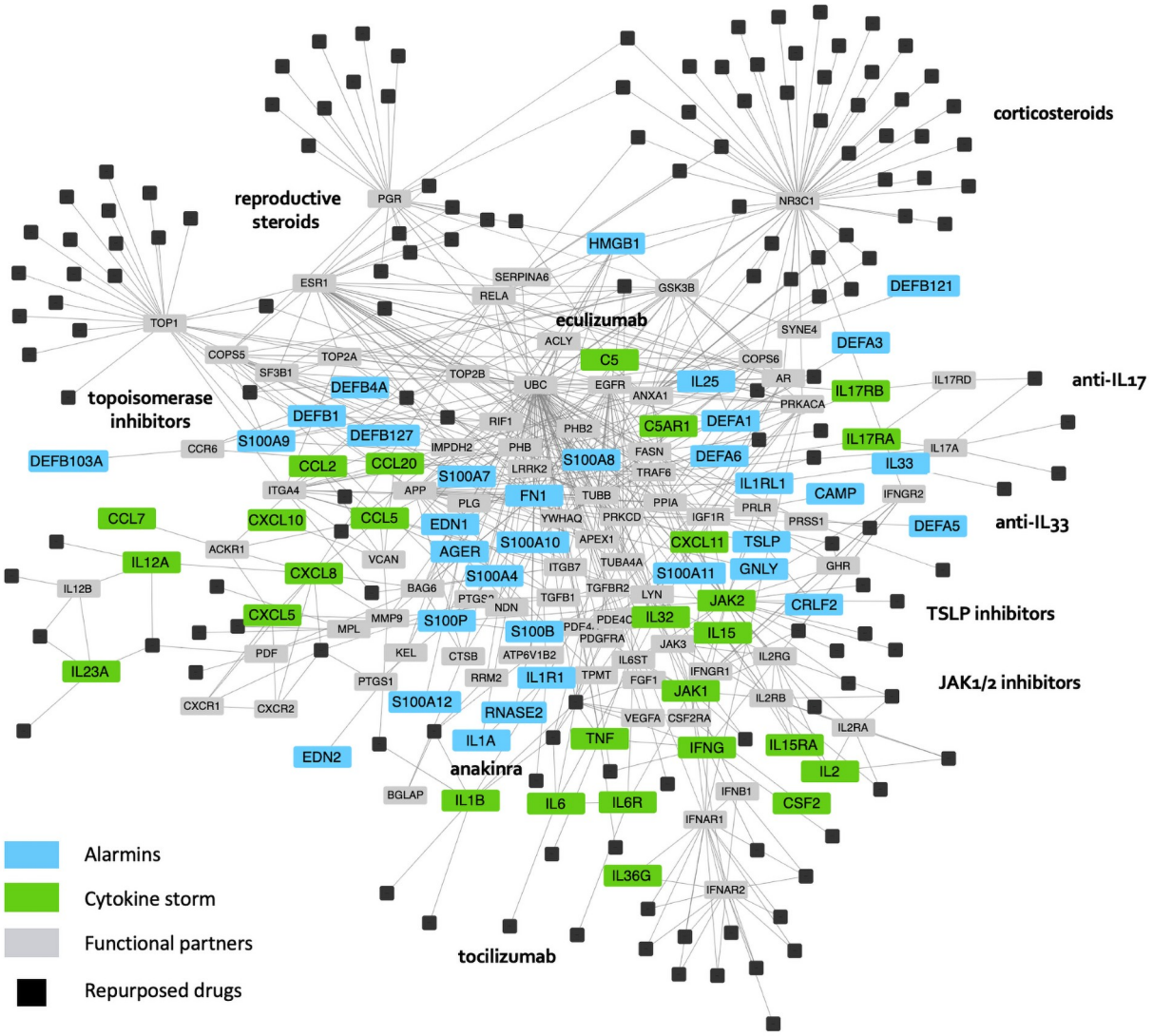


Fig 4. Druggable interactome of proteins contributing to lung inflammation in COVID-19. Extraction of the interactome encompassing proteins predicted to contribute to COVID-19 evolution towards a cytokine storm. Following SARS-CoV-2 infection of lung tissues and ensuing activation of innate and adaptive immune cells, different categories of proteins represent potential therapeutic targets to prevent or slow down lung inflammation associated with severe COVID-19. The latter include *Alarmins*, as well as cytokines, chemokines and selected proinflammatory factors associated with the *Cytokine storm*. For clarity, this figure only displays the disease related proteins (Alarmins & Cytokine storm) identified in our model, our top ranking repurposed drugs as well as some functional partners. The latter represent additional proteins needed in order to form a minimal principal component graph.

https://doi.org/10.1371/journal.pone.0254374.g004

functionally related IL23 cytokine or their receptors should be further investigated in COVID-19, with only one ongoing clinical trial in COVID-19 as of today [52]. In addition, the C5 complement inhibitor eculizumab is also predicted to represent an interesting treatment option, in agreement with recent evidence that the C5a-C5aR axis contributes to severe lung inflammation in COVID-19 patients [53]. As a strong chemoattractant, C5a provides in parallel to alarmins a link between innate and adaptive immune responses during severe COVID-19.

The thrombopoietin receptor appears as well to be a valid therapeutic target for agonists in light of the high incidence of thrombocytopenia associated with COVID-19 infection [54]. Rather unexpectedly, Topoisomerase 1 inhibitors, currently used as cytotoxic drugs in

Table 1. Overview of main therapeutic targets and clinical-stage candidate drugs for repurposing in COVID-19- related lung inflammation.

Therapeutic targets [Disease-related genes]	Candidate drugs for repurposing [Company name]	Modalities	Marketed drugs: Yes/ No	Clinical status in COVID-19 [Clinical trial ref]	Ref.
Cytokine Release Syndrome: IL1β, IL6, TNFα and their receptors	**Anti-IL1 β** Canakinumab [Novartis]	Antibody	Yes	Completed phase 2 in COVID-19 severe pneumonia [NCT04476706]. No impact on survival without the use of an invasive artificial respirator.	[4, 5, 18, 35, 56–58]
	Anti-IL1 β GLS1027 [GeneOne Life Science]	Small molecule	No	Recruitment planned for phase 2 in severe COVID-19 pneumonia [NCT04590547].	
	Anti-IL6 Clazakizumab [CSL Limited]	Antibody	No	Ongoing phase 2 in life-threatening COVID-19 infection [NCT04343989].	
	Anti-IL6 Olokizumab [R-Pharm]	Antibody	Yes	Completed phase 3 in acute respiratory distress syndrome [NCT04380519]. Results not yet available.	
	Anti-IL6 Siltuximab [EUSA Pharma]	Antibody	Yes	Ongoing phase 3 in acute respiratory Distress Syndrome [NCT04616586].	
	Anti-IL6 Sirukumab [Johnson & Johnson]	Antibody	No	Ongoing phase 2 in severe COVID-19 infection [NCT04380961].	
	Anti-IL6R Sarilumab [Sanofi]	Antibody	Yes	Completed phase 3 in severe or critical COVID-19 infection [NCT04327388], which did not meet its primary endpoint. Some improvement in survival when treating critically ill COVID-19 patients in association with dexamethasone.	
	Anti-IL6R Tocilizumab [Roche]	Antibody	Yes	Several trials completed in severe COVID-19 showing only limited efficacy [NCT04381936]. Some improvement in survival when treating critically ill COVID-19 patients in association with dexamethasone.	
	Anti-TNFα Infliximab [Johnson & Johnson]	Antibody	Yes	Ongoing phase 3 in COVID-19 [NCT04593940].	
	Anti-TNFα Adalimumab, [AbbVie]	Antibody	Yes	Ongoing phase 3 in mild to moderate COVID-19 [NCT04705844].	
	TNF-α inhibitor XPro-1595 [INmune Bio]	Peptide	No	Ongoing phase 2 in pulmonary complications of COVID-19 [NCT04370236].	
	Anti-TNFα Etanercept [Amgen]	Fusion protein	Yes	No evaluation yet in COVID-19.	
Glucocorticoid receptor NR3C1	**Corticosteroids** Dexamethasone [Mylan], Hydrocortisone [Sanofi-Aventis], Prednisolone [Mylan]	Small agonist molecules	Yes	Positive results obtained in the RECOVERY phase 3 study [NCT04381936], confirmed by a WHO-sponsored meta-analysis of 7 randomized clinical trials, collectively providing evidence for a reduced mortality of critically ill patients.	[2, 59]
				Dexamethasone is broadly used as a treatment for severe COVID-19.	
JAK1, JAK2	**JAK1/JAK2 inhibitor** Baricitinib [Eli Lily]	Small molecule	Yes	Ongoing phase 2 in moderate pneumonia [NCT04358614]. Recent evidence that Baricitinib can inhibit viral entry by clathrin-mediated endocytosis.	[60, 61]
	JAK/JAK2 inhibitor Ruxolitinib [Novartis]	Small molecule	Yes	Ongoing phase 2 in severe COVID-19 pneumonia [NCT04359290].	
	JAK2 inhibitor Jaktinib [Suzhou Zelgen Biopharmaceutical]	Small molecule	No	Completed phase 2 in severe and acute exacerbation of COVID-19 pneumonia [ChiCTR2000030170].*	
	JAK2 inhibitor Pacritinib [CTI BioPharma]	Small molecule	No	Ongoing phase 3 in severe COVID-19 [NCT04404361].	
	JAK2 inhibitor TD-0903 [Theravance Biopharma]	Small molecule	No	Ongoing phase 2 in symptomatic acute lung injury associated with COVID-19 [NCT04402866].	
Reproductive steroids: Estrogens, progesterone and their receptors	**Receptor agonists** Ethinylestradiol + Norelgestromin [Johnson & Johnson]	Small molecules	Yes	Planned phase 2 in non-severe COVID-19 patients [NCT04539626].	[63]

(*Continued*)

Table 1. (Continued)

Therapeutic targets [Disease-related genes]	Candidate drugs for repurposing [Company name]	Modalities	Marketed drugs: Yes/ No	Clinical status in COVID-19 [Clinical trial ref]	Ref.
Cytokines: IL2, IL15, IL17	**IL2Rβ superagonist** Bempegaldesleukin [Nektar]	Recomb protein	No	Ongoing phase 1b in mild COVID-19 [NCT04646044].	[52]
	IL15 super agonist ALT803 [Altor Biosciences]	Recomb protein	No	Planned phase 1 study in mild to moderate COVID-19.	
	Anti-IL17 Secukinumab [Novartis]	Antibody	Yes	Ongoing phase 2 in mild and severe COVID 19 [NCT04403243].	
	Anti-IL17, -IL17R, -IL23	Antibodies	Yes	No evaluation yet in COVID-19. Anti IL17 [Ixekizumab, Eli Lilly], anti IL17R [Brodalumab, Astra Zeneca/ Amgen], anti IL23 [Ustekinumab, Johnson & Johnson; Tildrakizumab, Merck] antibodies are commercialized as treatments for inflammatory diseases.	
C5, C5aR	**Anti C5** Eculizumab [Alexion]	Antibody	Yes	Proof-of-concept evidence suggesting that eculizumab provides some benefit in severe COVID-19. Ongoing phase 2 in moderate, severe or critical COVID-19 pneumonia [NCT04346797].	[53, 64–66]
	Anti C5aR Avdoralimab [Innate Pharma]	Antibody	No	Ongoing phase 2 in severe COVID-19 pneumonia [NCT04371367].	
Alarmins and their receptors: IL1α, TSLP, IL33	**IL1R1 antagonist** Anakinra [Sobi]	Peptide	Yes	Completed phase 2 in severe COVID-19 [NCT04366232]. Results not yet available.	[71, 75, 77]
	Anti-IL33R [ST2] AMG282-Astegolimab [Genentech]	Antibody	No	Ongoing phase 2 in severe COVID-19 Pneumonia [NCT04386616].	
	TSLP inhibitor HY-209- NuSepin [Shaperon] agonist for G protein-coupled TGR5 receptor	Small molecule	No	Ongoing phase 2 in COVID-19 pneumonia [NCT04565379].	
	Anti IL25, -IL33, -TSLP	Antibodies	No	No evaluation yet in COVID-19. Anti IL25 [ABM-125, Abeome], Anti-IL33 [REGN3500, Regeneron] and anti TSLP [Teepelumab, Amgen] are in clinical evaluation as treatments for asthma or atopic dermatitis.	
	Anti S100A4, -S100A7,—S100P	Antibodies	No	No evaluation yet in COVID-19. Antibodies in preclinical development in cancer or autoimmune diseases by Cancer Res Technol and Lykera Biomed.	
Thrombopoietin receptor	**Receptor agonist** Romiplostim [Amgen]	Peptibody [peptide agonist fused to Fc IgG1]	Yes	Case study documenting platelet recovery following treatment by Romiplostim of a pediatric patient with thrombocytopenia due to COVID-19.	[54]

All clinical trial information are available in Clinical trials gov: https://www.clinicaltrials.gov/ or* in Chinese clinical trial Registry: http://www.chictr.org.cn/.

https://doi.org/10.1371/journal.pone.0254374.t001

oncology, were also identified as of potential interest in COVID-19, with as of today only pre-clinical evidence that they can inhibit SARS-CoV-2 inflammation and death in animal models [55].

Supportive Cmap-based for drug repurposing

Given the rather limited set of transcriptomics data available and the small Cmap coverage for repurposable drugs (*i.e.* only 17% of molecules in our drug database, with none of the biologics), results were taken as supportive in the present study. Among the top network-based drugs proposed for repurposing, only 2 corticosteroids (betamethasone and hydrocortisone) were confirmed to elicit a reversed gene expression profile (Cmap score < -0.3) when compared to the disease gene expression state.

Conclusion

This study was designed to identify existing drugs which could be repurposed in a short time frame as a treatment for severe forms of COVID-19. We reasoned that such drugs should target those molecular pathways involved in the transition from mild lung inflammation caused by viral infection up to the cytokine storm associated with advanced stages of the disease (Fig 2, central and right lower panels). To this aim, using multiple sources of molecular profiling data from the literature relevant to distinguish mild from severe forms of the disease at the level of tissues and immune cells, we established a model of lung inflammation associated with COVID-19 in the form of an interactome of disease-related proteins. Combined with pharmacological knowledge of drug targets, this interactome allowed us to identify existing compounds which could be made available to patients in a short time frame.

Our network computational analyses identified several candidate therapeutic targets and corresponding drugs to repurpose which were confirmatory of existing knowledge (Table 1). This includes for example therapeutic antibodies interfering with either IL1β, IL6, TNFα or their receptors directly contributing to the CRS associated with severe COVID-19. Various inhibitory antibodies directed to these targets have already been evaluated in COVID-19 patients, such as anti-IL1® (canakinumab), anti-IL6R (tocilizumab, sarilumab) or anti-TNFα (infliximab, adalimumab) antibodies [4, 56]. Overall, these drugs yielded conflicting efficacy results, likely explained by evidence that such anti-cytokine treatments are rather effective if administered to patients before they develop advanced COVID-19 [57]. Nonetheless, a recent study evaluating the anti-IL6R antibodies tocilizumab and sarilumab demonstrated some improvement in survival when treating critically ill COVID-19 patients, even more so when these drugs were associated with dexamethasone [4, 5, 58]. Corticosteroids, are also predicted by the present study to be useful in severe COVID-19, in agreement with positive results previously obtained in multiple randomized clinical trials, eventually leading to a broad use of dexamethasone as a treatment for severe COVID-19 [2, 59]. JAK1 and JAK2 inhibitors came out also as interesting candidates for repurposing, with several inhibitors being actively tested in COVID-19 patients [60]. In this therapeutic class, the JAK1/JAK2 inhibitor baricitinib is currently raising most of the interest in light of recent evidence that it interferes with virus entry mediated by clathrin-associated endocytosis (Table 1) [61]. We also identified drugs interfering with reproductive steroids or their receptors as valid candidates for repurposing. This observation makes sense in light of the strong bias towards males among patients with severe COVID-19, perhaps explained in part by the upregulation by androgens of the expression of the SARS CoV-2 receptor [62]. In contrast estrogens and progesterone are rather considered to be protective in light of their anti-inflammatory properties as well as their capacity to promote proliferation and repair of respiratory epithelial cells [63]. On this basis, treatment with estrogens are being considered in patients with mild COVID-19 (Table 1).

Perhaps more interestingly, our repurposing study sheds light on other therapeutic classes which as of today receive insufficient attention as potential treatments for severe COVID-19. We predict that inhibitors of the well-known IL17 and IL23 proinflammatory cytokines (or their receptors) could be useful in COVID-19, with to our knowledge a single clinical trial evaluating as of today the anti-IL17 antibody secukinumab in COVID-19 [52]. Multiple monoclonal antibodies blocking those cytokines have been registered as treatments for other inflammatory diseases, which thus could be promptly repurposed in COVID-19 (Table 1). Similarly, the C5 complement inhibitor eculizumab was also identified to represent a valid therapeutic option, in agreement with recent evidence that the C5a-C5aR axis promotes severe lung inflammation in COVID-19 patients by mediating recruitment and activation of pro-inflammatory myeloid cells [53, 64]. Only proof of concept studies have been conducted so far

in human with eculizumab, suggesting that this antibody may provide some benefit in severe COVID-19 [65, 66], with a confirmatory trial ongoing in a larger cohort of patients. Noteworthy, another clinical study has been recently initiated to evaluate as well in this indication the anti C5a receptor antibody avdoralimab (Table 1). Also, approaches combining JAK1/2 inhibitors with blockade of C5a with eculizumab are being considered as a treatment of severe pulmonary damage in COVID-19 patients [67]. Moreover, drugs such as romiplostim acting as an agonist for the thrombopoietin receptor are also predicted to be useful to treat COVID-19-associated thrombocytopenia, in agreement with a recent case study documenting platelet recovery following treatment with this drug of a COVID-19 pediatric patient [54].

The most significant outcome of our repurposing study is the prediction that several members of the alarmin family such as defensins, HMBG1, IL1α, IL25, IL33, TSLP, S100A4, S100A7, S100A8, S100A9, S100A12, S100B, S100P likely contribute to lung inflammation during COVID-19 (Fig 4) [68–70]. The role of each individual alarmin in this regard remains to be investigated, with presumably some of them (*e.g.* IL25, TSLP) rather contributing to the initial recruitment of myeloid cells and innate lymphoid cells following epithelial or endothelial cell infection, whereas others (IL33, S100 members) are likely being involved in later stages of lung inflammation culminating in the cytokine storm. The later assumption is consistent with recent observations that some alarmins can stimulate the production of both IL1β, IL6 and TNFα as well as multiple other proinflammatory cytokines and chemokines [71]. Furthermore, blood levels of IL1α, calprotectin (a heterodimer made of S100A8 and S100A9), S100A12, S100B and HGBM1 appear to correlate with COVID-19 severity [72–76] (S1 Table). Also, IL33 has been recently proposed to play a broad role in the pathophysiology of COVID-19 pneumonia by dampening both the antiviral interferon response as well as regulatory T cells, while promoting thrombosis and activating pro-inflammatory type 2 innate lymphoid cells and γδ T cells [77]. To our knowledge, only few clinical studies are being conducted as of today in COVID-19 with a TSLP inhibitor or with blocking antibodies directed to receptors for IL1α or IL33 (i.e. ST2), whereas multiple additional blocking monoclonal antibodies directed to IL25, IL33 or TSLP are well under clinical evaluation to treat severe forms of asthma or atopic dermatitis [62, 69]. Furthermore, various inhibitors of the S100 family of proteins currently in preclinical development may represent promising drug candidates for the future (Table 1). We thus recommend considering existing anti-alarmins therapies to treat severe COVID-19, most particularly in the context of the converging rationale from this computational study as well as recent wet-lab evidence that this important class of proteins conveying proinflammatory signals plays a critical role in the pathophysiology of severe COVID-19. Lastly, this first model of severe lung inflammation in COVID-19 should be updated as new data are generated to better distinguish at an early stage patients with a high risk of evolving towards severe lung inflammation from those who will only develop mild forms of the disease.

Supporting information

S1 Table. Candidate COVID-19 related disease genes.
(PDF)

S2 Table. Pathways enrichment analysis.
(PDF)

S3 Table. Drug repurposing.
(XLSX)

Acknowledgments

The authors are thankful to Dorothée Piva for providing excellent secretarial assistance.

Author Contributions

Conceptualization: Emiko Desvaux, Sandra Hubert, Audrey Aussy, Laurence Laigle, Mickaël Guedj, Philippe Moingeon.

Data curation: Emiko Desvaux, Sandra Hubert.

Formal analysis: Antoine Hamon, Cheïma Boudjeniba, Bastien Chassagnol, Jack Swindle, Audrey Aussy, Laurence Laigle, Jessica Laplume, Perrine Soret, Pierre Jean-François, Isabelle Dupin-Roger, Mickaël Guedj, Philippe Moingeon.

Investigation: Antoine Hamon, Cheïma Boudjeniba, Bastien Chassagnol, Jack Swindle, Jessica Laplume, Perrine Soret, Pierre Jean-François, Isabelle Dupin-Roger.

Methodology: Antoine Hamon, Cheïma Boudjeniba, Bastien Chassagnol, Jack Swindle, Jessica Laplume, Perrine Soret, Isabelle Dupin-Roger, Mickaël Guedj, Philippe Moingeon.

Software: Antoine Hamon, Cheïma Boudjeniba, Bastien Chassagnol, Jack Swindle, Perrine Soret.

Supervision: Mickaël Guedj, Philippe Moingeon.

Validation: Audrey Aussy, Laurence Laigle, Jessica Laplume, Pierre Jean-François, Isabelle Dupin-Roger, Mickaël Guedj, Philippe Moingeon.

Visualization: Emiko Desvaux, Sandra Hubert.

Writing – original draft: Mickaël Guedj, Philippe Moingeon.

Writing – review & editing: Emiko Desvaux, Sandra Hubert, Audrey Aussy, Laurence Laigle, Isabelle Dupin-Roger.

References

1. Callaway E. Could new COVID variants undermine vaccines? Labs scramble to find out. Nature. 2021; 589: 177–178. https://doi.org/10.1038/d41586-021-00031-0 PMID: 33432212
2. Group TRC. Dexamethasone in Hospitalized Patients with Covid-19—Preliminary Report. New England Journal of Medicine. 2020 [cited 18 Jan 2021].
3. Beigel JH, Tomashek KM, Dodd LE, Mehta AK, Zingman BS, Kalil AC, et al. Remdesivir for the Treatment of Covid-19—Final Report. N Engl J Med. 2020; 383: 1813–1826. https://doi.org/10.1056/NEJMoa2007764 PMID: 32445440
4. Salvarani C, Dolci G, Massari M, Merlo DF, Cavuto S, Savoldi L, et al. Effect of Tocilizumab vs Standard Care on Clinical Worsening in Patients Hospitalized With COVID-19 Pneumonia: A Randomized Clinical Trial. JAMA Intern Med. 2021; 181: 24–31. https://doi.org/10.1001/jamainternmed.2020.6615 PMID: 33080005
5. Hermine O, Mariette X, Tharaux P-L, Resche-Rigon M, Porcher R, Ravaud P, et al. Effect of Tocilizumab vs Usual Care in Adults Hospitalized With COVID-19 and Moderate or Severe Pneumonia: A Randomized Clinical Trial. JAMA Intern Med. 2021; 181: 32–40. https://doi.org/10.1001/jamainternmed.2020.6820 PMID: 33080017
6. Repurposed Antiviral Drugs for Covid-19—Interim WHO Solidarity Trial Results. New England Journal of Medicine. 2021; 384: 497–511. https://doi.org/10.1056/NEJMoa2023184 PMID: 33264556
7. Krammer F. SARS-CoV-2 vaccines in development. Nature. 2020; 586: 516–527. https://doi.org/10.1038/s41586-020-2798-3 PMID: 32967006
8. Zhou Y, Hou Y, Shen J, Huang Y, Martin W, Cheng F. Network-based drug repurposing for novel coronavirus 2019-nCoV/SARS-CoV-2. Cell Discov. 2020; 6: 14. https://doi.org/10.1038/s41421-020-0153-3 PMID: 32194980

9. Nabirotchkin S, Peluffo AE, Bouaziz J, Cohen D. Focusing on the Unfolded Protein Response and Autophagy Related Pathways to Reposition Common Approved Drugs against COVID-19. 2020 [cited 18 Jan 2021]. https://doi.org/10.20944/preprints202003.0302.v1

10. Li X, Yu J, Zhang Z, Ren J, Peluffo AE, Zhang W, et al. Network Bioinformatics Analysis Provides Insight into Drug Repurposing for COVID-2019. 2020 [cited 18 Jan 2021]. https://doi.org/10.20944/preprints202003.0286.v1

11. Ciliberto G, Cardone L. Boosting the arsenal against COVID-19 through computational drug repurposing. Drug Discovery Today. 2020; 25. https://doi.org/10.1016/j.drudis.2020.04.005 PMID: 32304645

12. Chowdhury KH, Chowdhury MR, Mahmud S, Tareq AM, Hanif NB, Banu N, et al. Drug Repurposing Approach against Novel Coronavirus Disease (COVID-19) through Virtual Screening Targeting SARS-CoV-2 Main Protease. Biology. 2021; 10: 2. https://doi.org/10.3390/biology10010002 PMID: 33374717

13. Stebbing J, Phelan A, Griffin I, Tucker C, Oechsle O, Smith D, et al. COVID-19: combining antiviral and anti-inflammatory treatments. Lancet Infect Dis. 2020; 20: 400–402. https://doi.org/10.1016/S1473-3099(20)30132-8 PMID: 32113509

14. Gysi DM, Valle ÍD, Zitnik M, Ameli A, Gan X, Varol O, et al. Network Medicine Framework for Identifying Drug Repurposing Opportunities for COVID-19. arXiv:200407229 [cs, q-bio, stat]. 2020 [cited 1 Jun 2021]. Available: http://arxiv.org/abs/2004.07229 PMID: 32550253

15. Fiscon G, Conte F, Farina L, Paci P. SAveRUNNER: A network-based algorithm for drug repurposing and its application to COVID-19. PLOS Computational Biology. 2021; 17: e1008686. https://doi.org/10.1371/journal.pcbi.1008686 PMID: 33544720

16. Fiscon G, Paci P. SAveRUNNER: An R-based tool for drug repurposing. BMC Bioinformatics. 2021; 22: 150. https://doi.org/10.1186/s12859-021-04076-w PMID: 33757425

17. Tzotzos SJ, Fischer B, Fischer H, Zeitlinger M. Incidence of ARDS and outcomes in hospitalized patients with COVID-19: a global literature survey. Critical Care. 2020; 24: 516. https://doi.org/10.1186/s13054-020-03240-7 PMID: 32825837

18. Vardhana SA, Wolchok JD. The many faces of the anti-COVID immune response. J Exp Med. 2020; 217. https://doi.org/10.1084/jem.20200678 PMID: 32353870

19. Moore JB, June CH. Cytokine release syndrome in severe COVID-19. Science. 2020; 368: 473–474. https://doi.org/10.1126/science.abb8925 PMID: 32303591

20. de la Rica R, Borges M, Gonzalez-Freire M. COVID-19: In the Eye of the Cytokine Storm. Front Immunol. 2020; 11. https://doi.org/10.3389/fimmu.2020.558898 PMID: 33072097

21. Guney E, Menche J, Vidal M, Barábasi A-L. Network-based in silico drug efficacy screening. Nature Communications. 2016; 7: 10331. https://doi.org/10.1038/ncomms10331 PMID: 26831545

22. Cheng F, Desai RJ, Handy DE, Wang R, Schneeweiss S, Barabási A-L, et al. Network-based approach to prediction and population-based validation of in silico drug repurposing. Nat Commun. 2018; 9: 2691. https://doi.org/10.1038/s41467-018-05116-5 PMID: 30002366

23. Blanco-Melo D, Nilsson-Payant BE, Liu W-C, Møller R, Panis M, Sachs D, et al. SARS-CoV-2 launches a unique transcriptional signature from in vitro, ex vivo, and in vivo systems. bioRxiv. 2020; 2020.03.24.004655. https://doi.org/10.1101/2020.03.24.004655

24. Ackermann M, Verleden SE, Kuehnel M, Haverich A, Welte T, Laenger F, et al. Pulmonary Vascular Endothelialitis, Thrombosis, and Angiogenesis in Covid-19. New England Journal of Medicine. 2020; 383: 120–128. https://doi.org/10.1056/NEJMoa2015432 PMID: 32437596

25. Liao M, Liu Y, Yuan J, Wen Y, Xu G, Zhao J, et al. Single-cell landscape of bronchoalveolar immune cells in patients with COVID-19. Nature Medicine. 2020; 26: 842–844. https://doi.org/10.1038/s41591-020-0901-9 PMID: 32398875

26. Laing AG, Lorenc A, del Molino del Barrio I, Das A, Fish M, Monin L, et al. A dynamic COVID-19 immune signature includes associations with poor prognosis. Nature Medicine. 2020; 26: 1623–1635. https://doi.org/10.1038/s41591-020-1038-6 PMID: 32807934

27. Hadjadj J, Yatim N, Barnabei L, Corneau A, Boussier J, Péré H, et al. Impaired type I interferon activity and exacerbated inflammatory responses in severe Covid-19 patients. medRxiv. 2020; 2020.04.19.20068015. https://doi.org/10.1126/science.abc6027 PMID: 32661059

28. Ng LFP, Hibberd ML, Ooi E-E, Tang K-F, Neo S-Y, Tan J, et al. A human in vitro model system for investigating genome-wide host responses to SARS coronavirus infection. BMC Infect Dis. 2004; 4: 34. https://doi.org/10.1186/1471-2334-4-34 PMID: 15357874

29. Brodin P. Immune determinants of COVID-19 disease presentation and severity. Nature Medicine. 2021; 27: 28–33. https://doi.org/10.1038/s41591-020-01202-8 PMID: 33442016

30. Wen W, Su W, Tang H, Le W, Zhang X, Zheng Y, et al. Immune cell profiling of COVID-19 patients in the recovery stage by single-cell sequencing. Cell Discovery. 2020; 6: 1–18. https://doi.org/10.1038/s41421-020-0168-9 PMID: 32377375

31. Burke H, Freeman A, Cellura DC, Stuart BL, Brendish NJ, Poole S, et al. Inflammatory phenotyping predicts clinical outcome in COVID-19. Respiratory Research. 2020; 21: 245. https://doi.org/10.1186/s12931-020-01511-z PMID: 32962703

32. Wu M, Chen Y, Xia H, Wang C, Tan CY, Cai X, et al. Transcriptional and proteomic insights into the host response in fatal COVID-19 cases. PNAS. 2020; 117: 28336–28343. https://doi.org/10.1073/pnas.2018030117 PMID: 33082228

33. Combes AJ, Courau T, Kuhn NF, Hu KH, Ray A, Chen WS, et al. Global absence and targeting of protective immune states in severe COVID-19. Nature. 2021; 1–10. https://doi.org/10.1038/s41586-021-03234-7 PMID: 33494096

34. Arunachalam PS, Wimmers F, Mok CKP, Perera RAPM, Scott M, Hagan T, et al. Systems biological assessment of immunity to mild versus severe COVID-19 infection in humans. Science. 2020; 369: 1210–1220. https://doi.org/10.1126/science.abc6261 PMID: 32788292

35. Merad M, Martin JC. Pathological inflammation in patients with COVID-19: a key role for monocytes and macrophages. Nature Reviews Immunology. 2020; 20: 355–362. https://doi.org/10.1038/s41577-020-0331-4 PMID: 32376901

36. Vabret N, Britton GJ, Gruber C, Hegde S, Kim J, Kuksin M, et al. Immunology of COVID-19: Current State of the Science. Immunity. 2020; 52: 910–941. https://doi.org/10.1016/j.immuni.2020.05.002 PMID: 32505227

37. Choreño-Parra JA, Jiménez-Álvarez LA, Cruz-Lagunas A, Rodríguez-Reyna TS, Ramírez-Martínez G, Sandoval-Vega M, et al. Clinical and immunological factors that distinguish COVID-19 from pandemic influenza A(H1N1). medRxiv. 2020; 2020.08.10.20170761. https://doi.org/10.1101/2020.08.10.20170761

38. Atamas SP, Chapoval SP, Keegan AD. Cytokines in chronic respiratory diseases. F1000 Biol Rep. 2013; 5. https://doi.org/10.3410/B5-3 PMID: 23413371

39. Wang M, Withers JB, Ricchiuto P, Voitalov I, McAnally M, Sanchez HN, et al. A systems-based method to repurpose marketed therapeutics for antiviral use: a SARS-CoV-2 case study. Life Science Alliance. 2021; 4. https://doi.org/10.26508/lsa.202000904 PMID: 33593923

40. Stolfi P, Manni L, Soligo M, Vergni D, Tieri P. Designing a Network Proximity-Based Drug Repurposing Strategy for COVID-19. Front Cell Dev Biol. 2020; 8. https://doi.org/10.3389/fcell.2020.545089 PMID: 33123533

41. Cowen L, Ideker T, Raphael BJ, Sharan R. Network propagation: a universal amplifier of genetic associations. Nature Reviews Genetics. 2017; 18: 551–562. https://doi.org/10.1038/nrg.2017.38 PMID: 28607512

42. Song J-S, Wang R-S, Leopold JA, Loscalzo J. Network determinants of cardiovascular calcification and repositioned drug treatments. FASEB J. 2020; 34: 11087–11100. https://doi.org/10.1096/fj.202001062R PMID: 32638415

43. Chen X, Ji Z-L, Chen Y. TTD: Therapeutic Target Database. Nucleic acids research. 2002; 30: 412–5. https://doi.org/10.1093/nar/30.1.412 PMID: 11752352

44. Wishart DS, Knox C, Guo AC, Cheng D, Shrivastava S, Tzur D, et al. DrugBank: a knowledgebase for drugs, drug actions and drug targets. Nucleic Acids Res. 2008; 36: D901–906. https://doi.org/10.1093/nar/gkm958 PMID: 18048412

45. Cheng F, Kovács IA, Barabási A-L. Network-based prediction of drug combinations. Nature Communications. 2019; 10: 1197. https://doi.org/10.1038/s41467-019-09186-x PMID: 30867426

46. Cao M, Zhang H, Park J, Daniels NM, Crovella ME, Cowen LJ, et al. Going the Distance for Protein Function Prediction: A New Distance Metric for Protein Interaction Networks. PLOS ONE. 2013; 8: e76339. https://doi.org/10.1371/journal.pone.0076339 PMID: 24194834

47. Lamb J, Crawford ED, Peck D, Modell JW, Blat IC, Wrobel MJ, et al. The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. Science. 2006; 313: 1929–1935. https://doi.org/10.1126/science.1132939 PMID: 17008526

48. Sokulsky LA, Garcia-Netto K, Nguyen TH, Girkin JLN, Collison A, Mattes J, et al. A Critical Role for the CXCL3/CXCL5/CXCR2 Neutrophilic Chemotactic Axis in the Regulation of Type 2 Responses in a Model of Rhinoviral-Induced Asthma Exacerbation. The Journal of Immunology. 2020; 205: 2468–2478. https://doi.org/10.4049/jimmunol.1901350 PMID: 32948685

49. Ye Q, Wang B, Mao J. The pathogenesis and treatment of the 'Cytokine Storm' in COVID-19. J Infect. 2020; 80: 607–613. https://doi.org/10.1016/j.jinf.2020.03.037 PMID: 32283152

50. Leisman DE, Ronner L, Pinotti R, Taylor MD, Sinha P, Calfee CS, et al. Cytokine elevation in severe and critical COVID-19: a rapid systematic review, meta-analysis, and comparison with other inflammatory syndromes. The Lancet Respiratory Medicine. 2020; 8: 1233–1244. https://doi.org/10.1016/S2213-2600(20)30404-5 PMID: 33075298

51. Information NC for B, Pike USNL of M 8600 R, MD B, Usa 20894. National Center for Biotechnology Information. [cited 19 May 2021]. https://www.ncbi.nlm.nih.gov/

52. Pacha O, Sallman MA, Evans SE. COVID-19: a case for inhibiting IL-17? Nature reviews Immunology. 2020; 20: 345–346. https://doi.org/10.1038/s41577-020-0328-z PMID: 32358580

53. Carvelli J, Demaria O, Vély F, Batista L, Chouaki Benmansour N, Fares J, et al. Association of COVID-19 inflammation with activation of the C5a–C5aR1 axis. Nature. 2020; 588: 146–150. https://doi.org/10.1038/s41586-020-2600-6 PMID: 32726800

54. Schneider CW, Penney SW, Helfrich AM, Hartman KR, Lieuw K. A Novel Use of Romiplostim for SARS-CoV-2–induced Thrombocytopenia. Journal of Pediatric Hematology/Oncology. 2021;Publish Ahead of Print. https://doi.org/10.1097/MPH.0000000000001961 PMID: 33003146

55. Ho JSY, Mok BW-Y, Campisi L, Jordan T, Yildiz S, Parameswaran S, et al. Topoisomerase 1 inhibition therapy protects against SARS-CoV-2-induced inflammation and death in animal models. bioRxiv. 2020; 2020.12.01.404483. https://doi.org/10.1101/2020.12.01.404483 PMID: 33299999

56. Robinson PC, Liew DFL, Liew JW, Monaco C, Richards D, Shivakumar S, et al. The Potential for Repurposing Anti-TNF as a Therapy for the Treatment of COVID-19. Med (N Y). 2020; 1: 90–102. https://doi.org/10.1016/j.medj.2020.11.005 PMID: 33294881

57. De Stefano L, Bobbio-Pallavicini F, Manzo A, Montecucco C, Bugatti S. A "Window of Therapeutic Opportunity" for Anti-Cytokine Therapy in Patients With Coronavirus Disease 2019. Front Immunol. 2020; 11. https://doi.org/10.3389/fimmu.2020.572635 PMID: 33123149

58. Della-Torre E, Campochiaro C, Cavalli G, De Luca G, Napolitano A, La Marca S, et al. Interleukin-6 blockade with sarilumab in severe COVID-19 pneumonia with systemic hyperinflammation: an open-label cohort study. Ann Rheum Dis. 2020; 79: 1277–1285. https://doi.org/10.1136/annrheumdis-2020-218122 PMID: 32620597

59. WHO Rapid Evidence Appraisal for COVID-19 Therapies (REACT) Working Group, Sterne JAC, Murthy S, Diaz JV, Slutsky AS, Villar J, et al. Association Between Administration of Systemic Corticosteroids and Mortality Among Critically Ill Patients With COVID-19: A Meta-analysis. JAMA. 2020; 324: 1330–1341. https://doi.org/10.1001/jama.2020.17023 PMID: 32876694

60. Luo W, Li Y-X, Jiang L-J, Chen Q, Wang T, Ye D-W. Targeting JAK-STAT Signaling to Control Cytokine Release Syndrome in COVID-19. Trends in Pharmacological Sciences. 2020; 41: 531–543. https://doi.org/10.1016/j.tips.2020.06.007 PMID: 32580895

61. Seif F, Aazami H, Khoshmirsafa M, Kamali M, Mohsenzadegan M, Pornour M, et al. JAK Inhibition as a New Treatment Strategy for Patients with COVID-19. Int Arch Allergy Immunol. 2020; 181: 467–475. https://doi.org/10.1159/000508247 PMID: 32392562

62. Fagone P, Ciurleo R, Lombardo SD, Iacobello C, Palermo CI, Shoenfeld Y, et al. Transcriptional landscape of SARS-CoV-2 infection dismantles pathogenic pathways activated by the virus, proposes unique sex-specific differences and predicts tailored therapeutic strategies. Autoimmunity Reviews. 2020; 19: 102571. https://doi.org/10.1016/j.autrev.2020.102571 PMID: 32376402

63. Pinna G. Sex and COVID-19: A Protective Role for Reproductive Steroids. Trends in Endocrinology & Metabolism. 2021; 32: 3–6. https://doi.org/10.1016/j.tem.2020.11.004 PMID: 33229187

64. Peffault de Latour R, Bergeron A, Lengline E, Dupont T, Marchal A, Galicier L, et al. Complement C5 inhibition in patients with COVID-19—a promising target? Haematologica. 2020; 105: 2847–2850. https://doi.org/10.3324/haematol.2020.260117 PMID: 33256385

65. Annane D, Heming N, Grimaldi-Bensouda L, Frémeaux-Bacchi V, Vigan M, Roux A-L, et al. Eculizumab as an emergency treatment for adult patients with severe COVID-19 in the intensive care unit: A proof-of-concept study. EClinicalMedicine. 2020; 28. https://doi.org/10.1016/j.eclinm.2020.100590 PMID: 33173853

66. Diurno F, Numis FG, Porta G, Cirillo F, Maddaluno S, Ragozzino A, et al. Eculizumab treatment in patients with COVID-19: preliminary results from real life ASL Napoli 2 Nord experience. Eur Rev Med Pharmacol Sci. 2020; 24: 4040–4047. https://doi.org/10.26355/eurrev_202004_20875 PMID: 32329881

67. Giudice V, Pagliano P, Vatrella A, Masullo A, Poto S, Polverino BM, et al. Combination of Ruxolitinib and Eculizumab for Treatment of Severe SARS-CoV-2-Related Acute Respiratory Distress Syndrome: A Controlled Study. Front Pharmacol. 2020; 11: 857. https://doi.org/10.3389/fphar.2020.00857 PMID: 32581810

68. Yalcin Kehribar D, Cihangiroglu M, Sehmen E, Avci B, Capraz A, Yildirim Bilgin A, et al. The receptor for advanced glycation end product (RAGE) pathway in COVID-19. Biomarkers. 2021; 1–5. https://doi.org/10.1080/1354750X.2020.1861099 PMID: 33284049

69. Roth A, Lütke S, Meinberger D, Hermes G, Sengle G, Koch M, et al. LL-37 fights SARS-CoV-2: The Vitamin D-Inducible Peptide LL-37 Inhibits Binding of SARS-CoV-2 Spike Protein to its Cellular

Receptor Angiotensin Converting Enzyme 2 In Vitro. bioRxiv. 2020; 2020.12.02.408153. https://doi.org/10.1101/2020.12.02.408153

70. Idris MM, Banu S, Siva AB, Nagaraj R. Downregulation of Defensin genes in SARS-CoV-2 infection. medRxiv. 2020; 2020.09.21.20195537. https://doi.org/10.1101/2020.09.21.20195537

71. Yang D, Han Z, Oppenheim JJ. ALARMINS AND IMMUNITY. Immunol Rev. 2017; 280: 41–56. https://doi.org/10.1111/imr.12577 PMID: 29027222

72. Chen L, Long X, Xu Q, Tan J, Wang G, Cao Y, et al. Elevated serum levels of S100A8/A9 and HMGB1 at hospital admission are correlated with inferior clinical outcomes in COVID-19 patients. Cellular & Molecular Immunology. 2020; 17: 992–994. https://doi.org/10.1038/s41423-020-0492-x PMID: 32620787

73. Silvin A, Chapuis N, Dunsmore G, Goubet A-G, Dubuisson A, Derosa L, et al. Elevated Calprotectin and Abnormal Myeloid Cell Subsets Discriminate Severe from Mild COVID-19. Cell. 2020; 182: 1401–1418.e18. https://doi.org/10.1016/j.cell.2020.08.002 PMID: 32810439

74. Zuniga M, Gomes C, Carsons SE, Bender MT, Cotzia P, Miao QR, et al. Autoimmunity to the Lung Protective Phospholipid-Binding Protein Annexin A2 Predicts Mortality Among Hospitalized COVID-19 Patients. medRxiv. 2021; 2020.12.28.20248807. https://doi.org/10.1101/2020.12.28.20248807

75. Aceti A, Margarucci LM, Scaramucci E, Orsini M, Salerno G, Di Sante G, et al. Serum S100B protein as a marker of severity in Covid-19 patients. Scientific Reports. 2020; 10: 18665. https://doi.org/10.1038/s41598-020-75618-0 PMID: 33122776

76. Zeng Z, Hong X-Y, Li Y, Chen W, Ye G, Li Y, et al. Serum-soluble ST2 as a novel biomarker reflecting inflammatory status and illness severity in patients with COVID-19. Biomarkers in Medicine. 2020; 14: 1619–1629. https://doi.org/10.2217/bmm-2020-0410 PMID: 33336592

77. Zizzo G, Cohen PL. Imperfect storm: is interleukin-33 the Achilles heel of COVID-19? The Lancet Rheumatology. 2020; 2: e779–e790. https://doi.org/10.1016/S2665-9913(20)30340-4 PMID: 33073244