Thesis for doctoral degree (Ph.D.)
2022

B Cell Transcriptomics and Immunoglobulin Genetics in Rheumatoid Arthritis

Uta Hardt

Karolinska Institutet

From DEPARTMENT OF MEDICINE, SOLNA, and
DEPARTMENT OF MICROBIOLOGY, TUMOR AND CELL BIOLOGY
Karolinska Institutet, Stockholm, Sweden

B CELL TRANSCRIPTOMICS AND IMMUNOGLOBULIN GENETICS IN RHEUMATOID ARTHRITIS

Uta Hardt

Karolinska Institutet

Stockholm 2022

All previously published papers were reproduced with permission from the publisher.
Published by Karolinska Institutet.
Printed by Universitetsservice US-AB, 2022

ISBN 978-91-8016-731-4
Cover illustration: Inflamed joint by Uta Hardt

B CELL TRANSCRIPTOMICS AND IMMUNOGLOBULIN GENETICS IN RHEUMATOID ARTHRITIS
THESIS FOR DOCTORAL DEGREE (Ph.D.)

By

Uta Hardt

The thesis will be defended in public at the CMM Lecture Hall, Visionsgatan 18, L8:00, Center for Molecular Medicine, Solna, Stockholm, Friday, November 25th, 2022 at 10:00

Principal Supervisor:
Prof. Vivianne Malmström
Karolinska Institutet
Department of Medicine, Solna
Division of Rheumatology

Co-supervisor(s):
Prof. Gunilla B. Karlsson Hedestam
Karolinska Institutet
Department of Microbiology, Tumor and Cell Biology

Assoc. Prof. Leonid Padyukov
Karolinska Institutet
Department of Medicine, Solna
Division of Rheumatology

Dr. Martin M. Corcoran
Karolinska Institutet
Department of Microbiology, Tumor and Cell Biology

Opponent:
Prof. Andreas Radbruch
Humboldt University in Berlin (Charité)
Medical Faculty

Examination Board:
Prof. Richard Rosenquist Brandell
Karolinska Institutet
Department of Molecular Medicine and Surgery

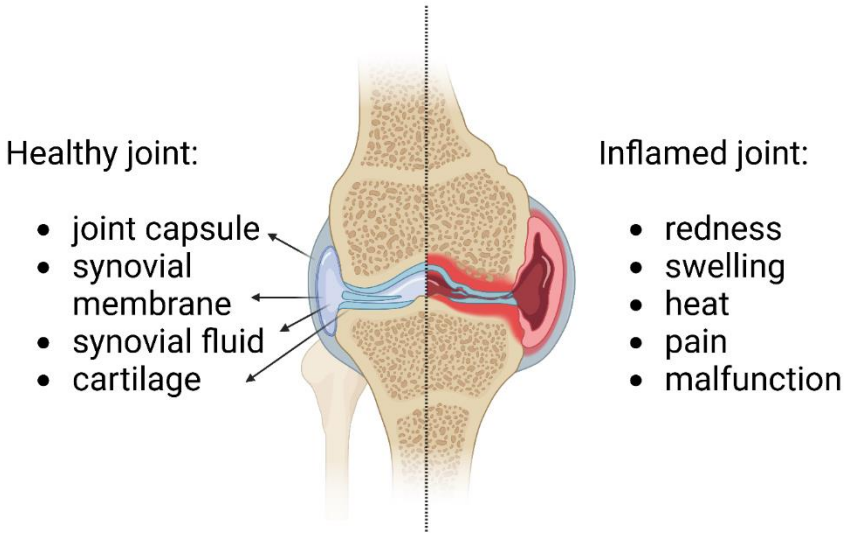
Prof. Inger Gjertsson
University of Gothenburg
Department of Rheumatology and Inflammation Research

Assoc. Prof. Lisa Westerberg
Karolinska Institutet
Department of Microbiology, Tumor and Cell Biology

To my friends and family

POPULAR SCIENCE SUMMARY (ENGLISH)

Rheumatoid arthritis is an autoimmune disease that is characterized by inflammation of the joints. Essentially, the body is attacked by its own immune system. This occurs, for example, when B cells release antibodies, which target the body's own proteins, impairing their natural function.



Antibodies have two structural parts: first, the variable fragment, which recognizes and binds antigens. Antigens are structures that can be bound by antibodies. Depending on the specificity of the antibody, the variable fragment recognizes one single or multiple antigens. Taken as an ensemble, antibodies can recognize a multitude of antigens with their variable fragments. Second, the constant fragment, which mediates effector functions of the antibody. Antibodies can be compared to screwdrivers: the variable fragment corresponds to the tip of the screwdriver that fits and touches the screw, the specific antigen, while the constant fragment corresponds to the handle of the screwdriver that mediates the function of the screwdriver by human movement.

In autoimmune diseases, recognized antigen structures are, for example, the body's own proteins. Proteins are molecules that are composed of a unique combination of twenty different kinds of amino acids linked together as a chain that adopts different three-dimensional forms. The amino acids have different properties that distinguish them from each other. In rheumatoid arthritis, antibodies recognize citrullinated proteins, which are proteins with a citrulline tag. Antibodies that recognize citrullinated proteins are also termed "anti-citrullinated protein antibodies", short ACPAs. Other antibodies recognize the constant part of other antibodies. These antibodies are called "rheumatoid factor", short RF. RF tends to form complexes, where one RF always binds the opposite end of the next RF like a bundle of magnets, of which the opposite poles always stick to one another.

About two thirds of the patients suffering from rheumatoid arthritis develop ACPAs and/or RF and are referred to as seropositive. Seropositive and seronegative patients can also produce

antibodies that target other protein structures. However, these are not crucial for today's classification criteria of rheumatoid arthritis. To date, it is unknown which roles different antibodies and B cells play that populate the joint during rheumatoid arthritis disease. It is possible that different antibodies disrupt healthy processes, foster pathogenic processes or are just bystanders of an immune reaction. However, the ACPA levels rise shortly before the onset and remain constant, regardless of any clinical remission. This has also led to the hypothesis that B cells play a role in the disease pathogenesis early on and that certain variants of antibody genes increase a person's risk of developing rheumatoid arthritis. This knowledge gap formed the basis of the studies in this thesis. Specifically, the following subjects were investigated:

The focus of the first published study was on B cells that were present in the joints at the time of rheumatoid arthritis diagnosis. I found that matured memory cells and antibody-producing cells were present in both seropositive and seronegative patients. Based on previous studies, it is expected that such cells are rather present in seropositive patients. Finding those cells in both patient groups suggests that also seronegative patients could benefit from medication, which disrupts the maturation and the surrounding environment of those cells, at the onset of the disease. Furthermore, I demonstrated that ACPA, normally found in blood, also derive from B cells in affected joints, where they are constantly triggered in presence of their antigen. How to target ACPA-producing B cells specifically and locally with medication, without interfering with other parts of the immune system, is still unknown.

For the second published study, I compared two methods that characterize antibody genes i.e. that determine the sequence of the genes encoding antibodies. I compared two methods that start from the antibody RNA to create libraries of antibody genes, which are then sequenced and analyzed with an algorithm. The RACE method was beneficial to generate unknown initial sequence information, while the MTPX method was better at representing the antibody genes in the experiments. This means that a sequential combination of the methods is suitable for projects with little initial sequence information, while the MTPX method is preferred in projects with advanced initial sequence information. The amount of known initial sequence information depends on how well the antibody genes of an organism are known. The human antibody genes compared to other model organisms are well described. Thus, to study human antibody genes it is sufficient to use the second method. The antibody genes of mice and monkeys are less well characterized. Therefore, in other model organisms both methods can be employed to generate optimal information about the antibody genes.

In the third study, I used the findings of the second study to characterize human antibody genes. I generated a map showing all germline-encoded antibody segments of thirty individuals with rheumatoid arthritis. My data indicated the expected structural variation of some gene segments and some novel gene variants, which had not been described in known databases. Analyses of individual chromosomes, so called haplotype analysis, showed that one common variable gene was seemingly only expressed from one of the chromosomes in some individuals. This was due to a variant nucleotide in this variable gene. To determine if this gene variant was also found in healthy individuals, I investigated this in a control group comprising thirty individuals

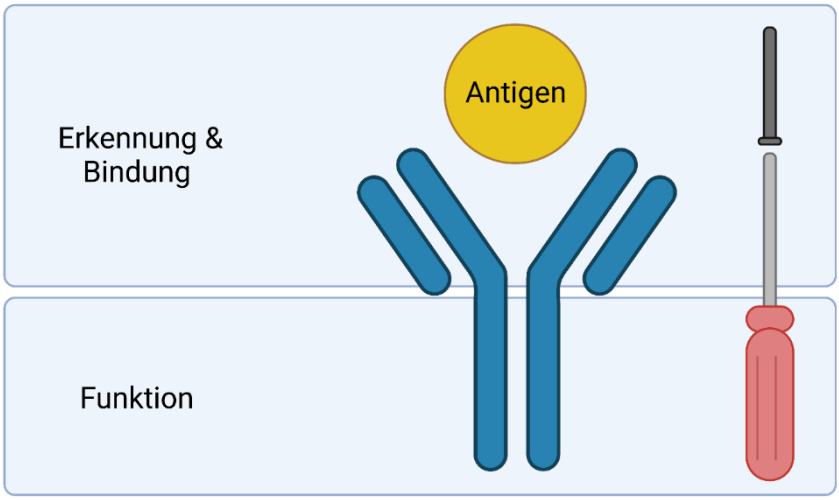
where I also found it at a similar frequency. Using a database describing how often variants occur in different populations, I found that this variant was often observed in the Finnish population. These studies show the importance of having a good understanding of the population's baseline genetic content before embarking on disease association studies.

In conclusion, I found matured memory cells and antibody-producing cells in early rheumatoid arthritis, and I describe approaches that can be used to investigate if antibody germline gene variations play a role in the development of rheumatoid arthritis. More studies are needed to understand disease risk in larger population groups. Such studies may guide the selection of therapeutic approaches to prevent rheumatoid arthritis disease progression. A preventative approach would be similar to the treatment of hypertension to prevent cardiovascular disease.

POPULÄRWISSENSCHAFTLICHE ZUSAMMENFASSUNG (DEUTSCH)

Rheumatoide Arthritis, kurz Rheuma, ist eine Autoimmunerkrankung, die sich durch eine Entzündung in den Gelenken auszeichnet. Dabei greift das Abwehrsystem des Körpers körpereigene Substanz an. Dies geschieht unter anderem indem sogenannte B-Zellen Antikörper ausschütten, die diese Substanz erkennen und ihre natürliche Funktion beeinträchtigen.

Antikörper haben zwei strukturelle Abschnitte: 1) Das variable Fragment erkennt und bindet Antigene. Antigene sind Strukturen, die von Antikörpern gebunden werden. In Abhängigkeit der Spezifität des Antikörpers erkennt das variable Fragment ein einziges oder mehrere Antigene. Das Ensemble der Antikörper kann mit seinen variablen Fragmenten verschiedenste Antigene erkennen. 2) Das konstante Fragment aktiviert gemeine Effektormoleküle, welche die Funktion der Antikörper vermitteln. Antikörper kann man sich in etwa wie einen Schraubenzieher vorstellen: Das variable Fragment entspricht der Spitze des Schraubenziehers, der auf den Kopf der Schraube, das spezifische Antigen, passt, während das konstante Fragment dem Griff entspricht, mit dem die Funktion des Schraubenziehers, das Lösen oder Anziehen der Schraube, in einer durch den Menschen ausgeführten Bewegung vermittelt wird.



Antikörper

Bei Autoimmunerkrankungen gehören zu den erkannten Antigenstrukturen unter anderem körpereigene Proteine, welche durch die Antikörper angegriffen werden. Proteine sind Moleküle, die durch eine einzigartige Kombination von zwanzig verschiedenen Aminosäuren zusammengesetzt sind und eine strukturelle Kette bilden. Die Aminosäuren haben unterschiedliche Eigenschaften in ihren Seitenketten, welche die Aminosäuren voneinander unterscheiden. Proteine kann man sich in etwa wie eine Kette mit unterschiedlichen Anhängern vorstellen. Bei Rheuma erkennen die Antikörper sogenannte zitrullinierte Proteine, d.h. Proteine mit einem Zitrullin als Anhänger. Antikörper, die zitrullinierte Proteine erkennen, werden nach dem Englischen „anti-citrullinated protein antibodies“ als ACPAs bezeichnet. Zusätzlich gibt es andere Antikörper, die die konstanten Fragmente der Antikörper selbst erkennen. Diese werden als rheumatoider Faktor, kurz RF, bezeichnet. RF neigt dazu ganze Komplexe zu bilden, bei denen ein RF, immer das jeweils entgegengesetzte Ende des nächsten RFs bindet wie bei einem Knäuel von Magneten, bei denen die entgegengesetzten Pole immer aneinander haften.

Etwa zwei Drittel der Rheumapatienten entwickeln ACPAs und/oder RF, die im Blut nachgewiesen werden. Diese Patienten werden dann als “seropositiv” bezeichnet. Sowohl seropositive, als auch seronegative Patienten können Antikörper ausbilden, die gegen andere körpereigene modifizierte Proteine gerichtet sind. Diese sind aber heutzutage nicht entscheidend für eine Klassifizierung als rheumatoide Arthritis. Welche Rolle die verschiedenen Antikörper und die B-Zellen, die in das Gelenk eindringen, in der Krankheitsentstehung spielen, ist bisher noch unklar. Es ist möglich, dass die verschiedenen Antikörper gesunde Prozesse stören, krankhafte Prozesse fördern oder dass sie nur Nebeneffekt einer Abwehrreaktion sind. Der ACPA-Spiegel steigt unmittelbar vor dem Ausbruch der Krankheit und bleibt unabhängig vom therapiebedingten Rückgang der Symptome konstant. Es wird vermutet, dass B-Zellen bereits in der frühen Krankheitsphase an der Krankheitsentstehung beteiligt sind und dass eine bestimmte Ausprägung der Antikörpergene das Risiko der Entstehung rheumatoider Arthritis erhöhen. Aufgrund dieser noch offenen Vermutungen habe ich mich in dieser Arbeit auf folgende Themen konzentriert:

In der ersten veröffentlichten Studie habe ich die B-Zellen untersucht, die zum Zeitpunkt der Feststellung der rheumatoiden Arthritis in den Gelenken präsent waren. Dabei stellte ich fest, dass ausgereifte Gedächtniszellen und antikörperproduzierende Zellen sowohl bei seropositiven, als auch bei seronegativen Patienten vorlagen. Aufgrund bisheriger Studien würde man erwarten, dass solche Zellen eher in seropositiven Patienten zu finden seien. Die Erkenntnis, dass solche Zellen in beiden Patientengruppen vorlagen, suggeriert, dass auch seronegative Patienten zu Beginn ihrer Krankheit von Medikamenten profitieren könnten, die die Ausreifung der Zellen in ihrer Umgebung stören. Außerdem konnte ich herleiten, dass ACPAs, die im Blut nachgewiesen werden, auch von B-Zellen in den betroffenen entzündeten Gelenken stammen, wo diese Zellen immer wieder angeregt werden, da sie dort auf ihre spezifischen Antigene treffen, und folglich kontinuierlich ACPAs ausschütten. Wie man diese ACPA-produzierenden B-Zellen spezifisch und lokal medikamentös beeinflussen könnte, ohne das gesamte Immunsystem zu schwächen, steht noch offen.

Im Rahmen der zweiten veröffentlichten Studie habe ich zwei Methoden verglichen, die es erlauben, Antikörpergene zu sequenzieren, das heißt die Abfolge der genetischen Bestandteile zu eruieren. Für diese Sequenzierung verglich ich zwei Methoden, die mit der Antikörper RNA beginnen und mit einem Algorithmus analysiert werden. Dabei war die RACE Methode geeignet unbekannte Sequenzinformationen zu generieren, wohingegen die MTPX Methode mithilfe der Erkenntnis dieser neuen Sequenzinformationen zu einer besseren Repräsentation der Antikörpergene in den Versuchen geführt hat. Das bedeutet, dass letztlich eine sequentielle Kombination der Methoden für Projekte mit wenig Anfangsinformation geeignet ist, wohingegen bei fortgeschrittenen Projekten die MTPX Methode bessere Ergebnisse liefert. Der Umfang der bekannten Anfangsinformation hängt im Wesentlichen davon ab, wie gut die Antikörpergene des Organismus beschrieben sind. Die Antikörpergene des Menschen sind im Vergleich zu den Modellorganismen relativ gut beschrieben. Daher ist die Anwendung der MTPX Methode ausreichend. Weniger gut beschrieben dagegen sind die Antikörpergene von Affen und Mäusen. In anderen Modellorganismen können beide Methoden verwendet werden, um ein gutes Abbild der Antikörpergene zu generieren.

In der dritten Studie verwendete ich die Erkenntnis der zweiten Studie, um die Antikörpergene des Menschen zu untersuchen. Dazu generierte ich eine Karte, die alle variablen Antikörpergensegmente von dreißig Individuen mit rheumatoider Arthritis aufzeigt. Meine Daten indizierten die erwarteten strukturellen Variationen von einigen Gensegmenten und einige neue Genvarianten, die noch nicht in bekannten Datenbanken beschrieben sind. In Analysen der individuellen Chromosomen, der sogenannten Haplotyp Analyse, konnte ich zeigen, dass ein allgemein häufiges variables Gen wirkte als sei es nur auf einem Chromosom präsent. Dies konnte ich auf eine genetische Variante zurückführen, die die Erkennungssequenz des variablen Gens betrifft. Um herauszufinden, ob diese genetische Variante auch in gesunden Individuen vorlag, untersuchte ich eine Kontrollgruppe mit dreißig Individuen, in der ich das Gen in einer ähnlichen Frequenz fand. In einer Datenbank, die beschreibt, wie oft genetische Varianten in verschiedenen Populationen auftreten, fand ich, dass die beschriebene Variante ebenfalls häufig in der finnischen Population auftrat. Diese Studien zeigen, wie wichtig es ist,

eine gutes Verständnis für die Zusammensetzung der genetischen Varianten in den verschiedenen Populationen zu haben, bevor man krankheitsbezogene Assoziationsstudien beginnt.

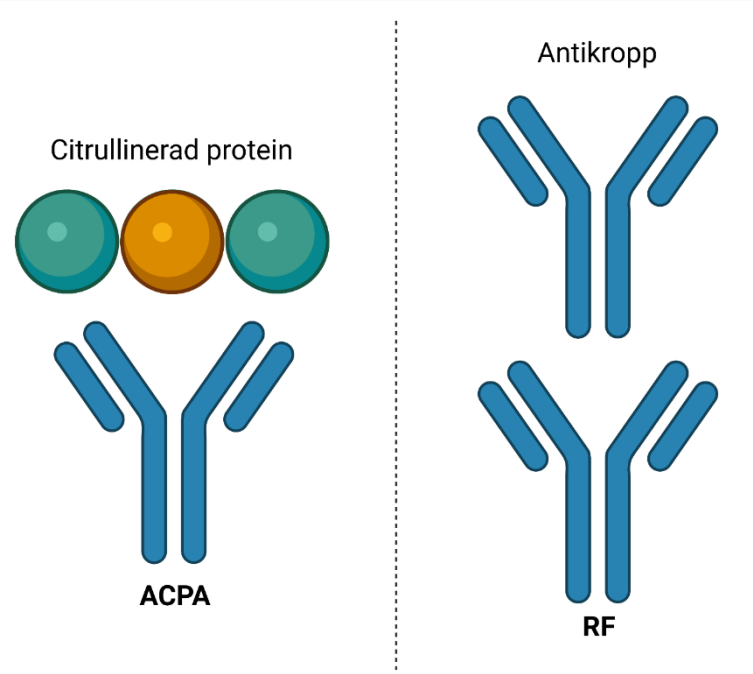
Zusammenfassend zeigte ich ausgereifte Gedächtniszellen und antikörperproduzierende Zellen in früher rheumatoider Arthritis und ich fand Antikörpergene, die die Basis des genetischen Make-Ups einer schwedischen rheumatoiden Arthritis Studiengruppe bilden. Es werden jedoch noch weitere Studien größerer Gruppen benötigt, in denen das Krankheitsrisiko und die Krankheitsentstehung bestätigt wird. Solche Studien könnten die Therapieansätze lenken, mit denen versucht werden kann, die Krankheitsentwicklung rheumtatoider Arthritis‘ aufzuhalten. Präventive Behandlung von Risikogruppen wäre ähnlich einer Bluthochdruck Behandlung bei Herz-Kreislauf-Erkrankungen.

POPULÄRVETENSKAPLIG SAMMANFATTNING (SVENSKA)

Ledgångsreumatism är en autoimmun sjukdom som kännetecknas av att kroppen blir attackerad av sitt eget immunförsvar. Detta inträffar till exempel när B-celler i immunförsvaret frisätter antikroppar som riktar sig mot kroppens egna proteiner som finns i ledytorna, proteinernas funktion försämras och slutligen leder det till att ytorna i lederna förstörs.

Antikroppar kan delas in i två delar: En del, den variabla, känner igen och binder till antigen. Antigener är strukturer som antikroppar binder till. Beroende på antikroppens specificitet känner variabla delen igen ett eller flera antigen. Tillsammans kan antikroppar känna igen en mängd olika antigen. Den andra delen av antikroppen, det konstanta fragmentet, aktiverar molekyler som förmedlar olika funktioner. Antikroppar kan ses ungefär som en skruvmejsel: den variabla delen motsvarar spetsen på skruvmejseln som passar på skruvhuvudet, det specifika antigenet, medan den konstanta delen motsvarar handtaget som utför skruvmejselns funktion att lossa eller dra åt skruven.

Antigen som känns igen i autoimmuna sjukdomar är kroppens egna proteiner, som attackeras av antikropparna. Proteiner är molekyler som består av en unik kombination av tjugo olika aminosyror som bildar kedjor som veckar sig på olika sätt. Aminosyrorna har olika egenskaper. Vid ledgångsreumatism känner antikropparna igen så kallade citrullinerade proteiner, som är modifierade till citrullin. Antikroppar som känner igen citrullinerade proteiner benämns ”anti-citrullinerade protein antikroppar”, kort ACPA. Dessutom finns det andra antikroppar som känner igen de konstanta fragmenten av själva antikropparna. Dessa är kända som ”reumatoid faktor”, kort RF. RF tenderar att bilda komplex, där en RF alltid binder den motsatta änden av nästa RF som ett knippe magneter, varav de motsatta polerna alltid fastnar vid varandra.



Ungefär två tredjedelar av patienter med ledgångsreumatism utvecklar ACPA och/eller RF, som upptäcks i blodet. Dessa patienter kallas för ”seropositiva”. Både seropositiva och seronegativa patienter kan utveckla antikroppar riktade mot andra modifierade proteiner. Dessa är dock inte avgörande för klassificeringen av ledgångsreumatism. Det är fortfarande oklart varför B-celler går in i leden och producerar olika ACPA och vad de spelar för roll i utvecklingen av sjukdomen. Det är möjligt att de olika antikropparna stör hälsosamma processer, främjar sjukdomsframkallande processer eller att det är en bieffekt av immunförsvaret. ACPA-nivåerna stiger strax innan början av sjukdomen och är konstanta oavsett behandling. Man antar att B-celler är involverade i utvecklingen av sjukdomen i en tidig fas och att förekomst av vissa typer av antikropparna ökar risken för att utveckla ledgångsreumatism. På grund av dessa antaganden har jag koncentrerat mig på följande ämnen i detta arbete:

I den första publicerade studien undersökte jag B-celler som fanns i lederna vid den tidpunkt då ledgångsreumatism diagnostiserades. Jag fann att mogna minnesceller och B-celler som är antikroppsproducerande fanns hos både seropositiva och seronegativa patienter. Baserat på tidigare studier förväntades det att sådana celler framför allt ska hittas hos seropositiva patienter. Upptäckten att sådana celler fanns i båda patientgrupperna tyder på att seronegativa patienter i början av sin sjukdom också skulle ha nytta av läkemedel som blockerar mognaden av dessa celler. Dessutom kunde jag härleda att ACPA som upptäcks i blodet också kommer från B-celler i de drabbade inflammerade lederna, där B-celler stimuleras upprepade gånger eftersom de möter sina antigen där, och därmed frisätts ACPAs kontinuerligt. Det återstår att se hur dessa ACPA-producerande B-celler kan påverkas specifikt och lokalt med medicinering utan att påverka andra delar av immunförsvaret.

I den andra publicerade studien jämförde jag två metoder för att sekvens bestämma antikroppars gener. Jag jämförde två metoder där man utgår från antikropparnas RNA för att producera bibliotek av antikroppar som sedan sekvenseras och analyseras. Jag fann att RACE metoden är lämplig för att generera okänd sekvensinformation, medan MTPX metoden leder till en bättre representation av antikroppsgenerna. Detta innebär att en kombination av metoderna i slutändan lämpar sig för projekt med lite bakgrundsinformation, medan MTPX metoden ger bättre resultat för avancerade projekt. Omfattningen av den redan kända informationen beror i huvudsak på hur väl antikroppsgenerna för organismen i fråga är beskrivna. De humana antikroppsgenerna är väl beskrivna jämfört med de från modellorganismer och därför är det tillräckligt att använda MTPX metoden för humana prover. Antikroppsgener hos till exempel apor och möss är å andra sidan mindre väl beskrivna, där bör båda metoderna användas för att skapa en bra bild av antikroppsgenerna.

I den tredje studien använde jag metoderna vi beskrev i den andra studien för att studera mänskliga antikroppsgener. Jag skapade en karta som visar alla variabla antikroppsgener från trettio individer med reumatoid artrit. Resultaten visade förväntade strukturella variationerna av vissa gener och några nya genvarianter som ännu inte beskrivits i kända databaser. Genom att bestämma vilka genvarianter som låg på vilken kromosom, så kallad haplotyp-analys, upptäckte jag att en ofta förekommande variabel gen endast uttrycktes från en av kromosomerna. Detta berodde på en enskild förändring i en nukleotid i den variabla genen. Jag undersökte sedan en kontrollgrupp på trettio individer och såg att denna genvariant fanns även där och i samma frekvens som i reumatoid artrit gruppen. Genom en databas som undersöker frekvensen av varianter i olika populationer kunde jag visa att den upptäckta varianten var vanlig i den finska befolkningen. Dessa studier visar vikten av förstå hur vanliga olika genvarianter är i populationer som ska studeras för att kunna undersöka genetiska riskfaktorer.

Sammanfattningsvis hittade jag mogna minnesceller och antikroppsproducerande B-celler i tidig reumatoid artrit och jag beskriver tillvägagångssätt som kan användas för att undersöka om antikropps genvariationer spelar en roll i utvecklingen av reumatoid artrit. Ytterligare studier behövs för att förstå sjukdomsrisk i större befolkningsgrupper. Sådana studier kan vägleda valet av terapeutiska metoder för att förhindra utveckling av reumatoid artrit. Ett förebyggande tillvägagångssätt skulle likna behandlingen av högt blodtryck för att förhindra hjärt-kärlsjukdom.

ABSTRACT

Rheumatoid Arthritis (RA) is a complex autoimmune disease characterized by inflammation of the peripheral joints. The exact disease pathology remains to be elucidated, but it has been hypothesized that autoreactive B cells and autoantibodies play a major role in the etiology. In this thesis, I studied synovial B cells at the time of RA diagnosis using single cell RNA sequencing, spatial transcriptomics and isolation of monoclonal antibodies. In a separate track, I optimized Next Generation Sequencing library preparation methods to study antibody repertoires, infer germline antibody alleles and I applied these methods to a local disease study group.

Using single cell RNA sequencing and spatial transcriptomics from patient paired tissue pieces, we found mostly memory B cells and naïve B cells in lymphoid rich biopsies. $CD27^{++}$ plasma cells were found to have a strong identity signal as shown by both methods. Similarily, both methods showed a strong signature for T-B cell crosstalk that we discuss to drive differentiation to memory and plasma cells. Furthermore, we found the CXCL12-CXCR4 axis to be an active component of a plasma cell niche in synovial tissue. Studies of the B cell receptor from the single cell RNA sequencing data imply a clonal recall response from memory cells differentiating into plasma cells due to recurrent antigen exposure. We also identified an anti-citrullinated protein reactivity towards modified vimentin in the tissue biopsies.

Moreover, we compared 5' rapid amplification of cDNA ends (5'RACE) antibody library preparation methods to a 5' multiplex (5'MTPX) approach. We detected a 5'RACE amplicon length limitation that occurs when the sequences of the antibody heavy chain repertoire have long 5' untranslated regions (UTRs) or long complementary determining regions 3 (CDR3s). Furthermore, we tested a single lambda constant chain primer versus a mix of single lambda constant chain specific primers and found that there was good detection of independent variable-joining segment (V-J) recombination that allows identification of lambda chain germline alleles.

Finally, we used the 5'MTPX approach to study the antibody heavy chain alleles of thirty individuals from the local Epidemiological Investigation of Rheumatoid Arthritis (EIRA) study. We found common structural variations as well as few novel alleles. In haplotype analysis, we observed that the IGHV4-34 gene was seemingly hemizygous. This could be linked to a polymorphism in the recombination signal sequence of one chromosome leading to the absence of the allele in the expressed repertoire. Notably, this variation was also found in thirty control subjects from the EIRA study group and in the SNP data from the Finnish population group in the 1000 Genomes Project. Hence, we stress the importance of population stratification when performing disease association studies.

In summary, I discuss RA therapy with respect to T-B cell interaction and a plasma cell survival niche, suitable antibody library generation methods for deep repertoire studies and germline gene inference studies, and the relevance of this for larger disease association studies.

LIST OF SCIENTIFIC PAPERS

I. **Uta Hardt***, Konstantin Carlberg*, Erik af Klint, Peter Sahlström, Ludvig Larsson, Annika van Vollenhoven, Susana Hernandez Machado, Lena Israelsson, Khaled Amara, Karine Chemin, Marina Korotkova, Gunilla B. Karlsson Hedestam, Anca I. Catrina, Sarah A. Teichmann, Patrik L. Ståhl, Vivianne Malmström. *"Integrated single cell and spatial transcriptomics reveal autoreactive differentiated B cells in joints of early rheumatoid arthritis."* Scientific Reports – 2022 Jul; 12(1):1-14.

   * These authors contributed equally.

II. Néstor Vázquez Bernat, Martin M. Corcoran, **Uta Hardt**, Mateusz Kaduk, Ganesh E. Phad, Marcel Martin and Gunilla B. Karlsson Hedestam. *"Highquality library preparation for NGS-based immunoglobulin germline gene inference and repertoire expression analysis."* Frontiers in Immunology – 2019 Apr; 10:660.

III. **Uta Hardt**, Martin M. Corcoran, Sanjana Narang, Vivianne Malmström, Leonid Padyukov, Gunilla B. Karlsson Hedestam. *"Analysis of IGH allele content in a sample group of rheumatoid arthritis patients demonstrates population heterogeneity."* Submitted manuscript.

SCIENTIFIC PAPERS NOT INCLUDED IN THE THESIS

I. **Uta Hardt**, Anders Larsson, Iva Gunnarsson, Robert M. Clancy, Michelle Petri, Jill P. Buyon, Gregg J. Silverman, Elisabet Svennungsson, Caroline Grönwall. *"Autoimmune reactivity to malondialdehyde adducts in systemic lupus erythematosus is associated with disease activity and nephritis."* Arthritis Research & Therapie – 2018 Feb; 20(1):36.

II. Caroline Grönwall, Khaled Amara, **Uta Hardt**, Akilan Krishnamurthy, Johanna Steen, Marianne Engström, Meng Sun, A. Jimmy Ytterberg, Roman A. Zubarev, Dagmar Scheel-Toellner, Jeffrey D. Greenberg, Lars Klareskog, Anca I. Catrina, Vivianne Malmström, Gregg J. Silverman. *"Autoreactivity to malondialdehyde-modifications in rheumatoid arthritis is linked to disease activity and synovial pathogenesis."* Journal of Autoimmunity – 2017 Nov; 84:29-45.

III. Caroline Grönwall, **Uta Hardt**, Johanna T. Gustafsson, Kerstin Elvin, Kerstin Jensen-Urstad, Marika Kvarnström, Giorgia Grosso, Johan Rönnelid, Leonid Padykov, Iva Gunnarsson, Gregg J. Silverman, Elisabet Svenungsson. *"Depressed serum IgM levels in SLE are restricted to defined subgroups."* Clinical Immunology – 2017 Oct; 183:304-315.

CONTENT

LIST OF ABBREVIATIONS

ABCs	Autoimmune-/Age-associated B Cells
ACPA	Anti-citrullinated Protein Antibodies
AID	Activation Induced cytidine Deaminase
BCR	B Cell Receptor
bp	base pair
CCP	Cyclic Citrullinated Peptide
CD	Cluster of Differentiation
CDR	Complementary Determining Region
CLP	Common Lymphoid Progenitor
CSR	Class Switch Recombination
DN	Double Negative (B cells)
EIRA	Epidemiological Investigation of Rheumatoid Arthritis
ERCC	External RNA Controls Consortium
F_{ab}	antigen binding Fragment
F_c	crystallizable Fragment
FWR	Framework Region
GC	Germinal Center
GWAS	Genome-Wide Association Studies
HC	Heavy Chain
HIV	Human Immunodeficiency Virus
HLA	Human Leucocyte Antigen
HSC	Hematopoietic Stem Cells
IARC	Inferred Allele Review sub-Committee
Ig	Immunoglobulin
IgH	Immunoglobulin Heavy chain
IGH	Immunoglobulin Heavy chain locus
IgL	Immunoglobulin Light chain
IMGT	The international ImMunoGeneTics system
ITAM	Immunoreceptor Tyrosine Activation Motif
LC	Light Chain

lncRNA	long non-coding RNA
MHC	Major Histocompatibility Complex
MTPX	Multiplex
mRNA	messenger RNA
MZB	Marginal Zone B cell
ng	nanogram
NGS	Next Generation Sequencing
nt	nucleotide
OA	Osteoarthritis
ORF	Open Reading Frame
PBMC	Peripheral Blood Mononuclear Cell
PC	Plasma Cell
PCA	Principal Component Analysis
PCR	Polymerase Chain Reaction
PDB	Protein Data Bank
PRR	Pattern Recognition Receptor
RA	Rheumatoid Arthritis
RACE	Rapid Amplification of cDNA Ends
RF	Rheumatoid Factor
RFLP	Restriction Fragment Length Polymorphism
RSS	Recombination Signal Sequence
SE	Shared Epitope
SHM	Somatic Hypermutation
sIgD	surface Immunoglobulin D
SLE	Systemic Lupus Erythematosus
SNP	Single Nucleotide Polymorphism
SPR	Surface Plasmon Resonance
TCR	T Cell Receptor
TLR	Toll-Like Receptor
UMI	Unique Molecular Identifier
UTR	Untranslated Region

1 INTRODUCTION

1.1 PRELUDE

Science and arts are closely related. Both are built on presenting pictures or snapshots of reality, and both are inspiring to each other. Yet, they depend on the creativity of the scientist or artist. While I was pursuing my PhD studies, I visited art museums in Stockholm, London, Cambridge, Boston, Paris, Amsterdam, Berlin, Munich, New York and Madrid, always on the hunt of my favorite Renoirs. Why Renoirs? Pierre Auguste Renoir is one of many famous persons who developed RA. Thus, it got the simile of my work.

RA is a chronic, painful and disabling autoimmune disease that about 1% of the population suffers from. My goal was always to help people without being the physician. So, I chose to get into life sciences to generate new basic understanding, which may improve physicians' opportunities to treat patients. In my PhD work, I focused on the B cell component in the joint tissue in early RA, methodological approaches to study the immunoglobulin germline and repertoire composition and on the baseline characteristics of these loci in a group of individuals. This is how I hope to contribute to the scientific knowledge.

2 LITERATURE REVIEW

2.1 RHEUMATOID ARTHRITIS

2.1.1 Disease and therapy

RA is a chronic disabling autoimmune disease primarily manifesting in classical inflammation and progressive destruction of the joints. Inflammation is characterized by five signs comprising heat, pain, redness, swelling and loss of function *(lat. calor, dolor, rubor, tumor, and functio laesa)*. The joints affected in RA are specific and include the metacarpophalangeal (MCP) and proximal interphalangeal (PIP) joint of the hands and feet (compare Figure 1), but can also include the wrist, ankle, elbow, shoulder, knee and hip joints[1]. Systemic manifestations of RA involve the eyes, the lungs, the heart and other organs. Cardiovascular disease is common among RA patients[2]. RA reduces health-related quality of life, thereby affecting individuals both physically and mentally[3].

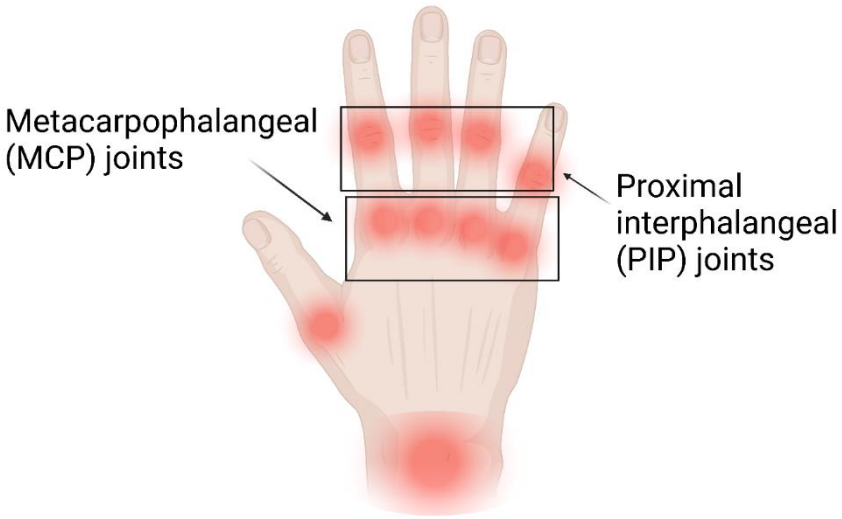


Figure 1 Scheme of the affected joints in RA

Treatment of RA aims at disease remission maximizing physical functionality of the joints in established disease, limiting permanent joint damage and requiring frequent disease monitoring. This treatment strategy involves the use of conventional Disease Modifying Antirheumatic Drugs (DMARDs) such as the synthetic DMARD methotrexate that dampen the inflammatory process and prevent joint destruction early on. Non-Steroidal Anti-Inflammatory Drugs (NSAIDs) targeting cyclooxygenase may be used symptomatically to reduce pain in the very early phase of RA. When the DMARDs do not exert the full anti-inflammatory effect, glucocorticoids can be added to the treatment for a maximum of 4-5 months. Glucocorticoids are effective but have a debated risk-to-benefit profile. When the conventional therapy conveys insufficient effects, RA patients may receive biological disease modifying therapy[4] from biological DMARDs. There are five modes of action among the biological DMARDs: TNF inhibition (adalimumab[5], certolizumab pegol[6], etanercept[7],

golimumab[8], and infliximab[9]), T cell co-stimulation blockade (abatacept[10]), IL-6 receptor inhibition (tocilizumab[11] and sarilumab[12]), B cell depletion (rituximab[13]) and interleukin 1 inhibition (anakinra[14]). Furthermore, targeted synthetic DMARDs inhibiting Janus kinase (JAK, tofacitinib[15] and baricitinib[16]) and biosimilars are also approved for RA. Antigen-specific treatments aiming at the induction of tolerance in RA are currently under investigation[17]. In these studies, the aim is to induce regulatory cells that are specific for the antigens associated with RA, so that they would dampen the specific autoimmune response.

2.1.2 Prevalence and incidence

RA has a prevalence of 0.5-1% across populations[18]. Interestingly, the prevalence of RA is higher in Northern compared to Southern Europe[19]. Among the Native American populations, a high prevalence of 5% is reported[20]. The annual incidence of RA is 40 per 100,000[21]. RA most commonly occurs between the age of 40 and 70 and affects females 2.5 times more often than males[22].

2.1.3 Protective and risk factors

Given the prevalence characteristics of RA, age and female sex are important risk factors. Further risk factors that are associated with RA include genetic and environmental factors.

RA has an important genetic contribution. Concordance rates among twins indicate a heritability of about 40% for RA[23], which is 20% less than older reports indicated[24], but about 30% more than other studies have shown[25]. There are about 150 risk loci associated with RA, which have been reported in genome wide association studies[26].

The class II human leukocyte antigen loci show a very strong association with RA[27]. The risk variants are collectively referred to as shared epitope (SE) alleles due to a common amino acid motif[28]. The presence of SE alleles account for about 8-9% of the phenotypic variance of RA[29].

Non class II human leukocyte antigen loci associated with RA account for about 5% of the heritability of RA and are shared risk variants with other autoimmune conditions. The risk loci with the highest odds ratios include the *PTPN22*, the *IL20RB*, the *NFKBIE*, the *TNFAIP3*, the *TYK2*, the *ICAM1* and the *ILF3* genes[30].

60% of the risk to develop RA is attributed to environmental factors. A prominent environmental risk factor is tobacco smoking. For men, ever, current and past smoking doubles the risk of RA compared to non-smokers. For women, ever, current and past smoking markedly increases the odds ratios of developing RA[31]. SE positive RA patients have a 20-fold increased risk compared to non-smokers that do not carry the SE[32].

Similarly to tobacco smoking, dust inhalation has been associated with RA. Silica exposure leads to an increased risk[33]. Emergency responders exposed to dust of mixed materials from the World Trade Center that collapsed in New York have increased risk of systemic autoimmune diseases including RA[34]. Equally, occupational textile dust inhalation is associated with elevated risk for developing RA[35].

Another environmental factor that confers risk for RA is the constitution of the microbiota. RA is associated with periodontal disease which is thought to be partly mediated by the oral microbiota *Porphyromonas gingivalis*[36] and *Aggregatibacter actinomycetemcomitans*[37]. Furthermore, there is an increase of *Prevotella copri* in the microbiota of the gut of RA patients[38].

Finally, modifiable lifestyle factors have been implicated in RA. Obese with a Body Mass Index of ≥ 30 kg/m^2 have an Odds Ratio of 1.45 compared to those with a body mass index of ≥ 25 kg/m^2 [39].

2.1.4 Trigger and pathogenesis

Based on the associations with genetic and environmental risk factors, RA is regarded as a complex immune-mediated disease that is triggered through environmental stimuli in genetically susceptible individuals. Still, the exact etiology of RA is only partially understood.

The environmental risk factors hint at a trigger at the mucosal surfaces in the airways, the oral cavity and perhaps also the intestine. Smoke and dust particles may induce the post-translational enzymatic citrullination of proteins in the lungs[40]. Bronchial biopsies from patients with untreated early RA show signs of immune activation and local inflammation[41]. Similarly, the citrullination may be mediated by microbiota such as *Porphyromonas gingivalis* in the gums during peridontitis[42]. In this context, the immune response may target a foreign antigen that mimics a self-antigen. Interestingly, peptides from intestinal *Prevotella copri* and from autoantigens that were isolated from class II human leukocyte antigen molecules of RA patients share significant sequence homology[43].

The current understanding of how the local immune response to neo-antigens progresses to systemic autoimmunity is limited. The fact that women are more likely to develop RA may be attributed to the stimulatory effects of estrogen on the immune system[44]. High estrogen levels are also associated with parity, while early age at first pregnancy associates with increased risk for RA[45]. On the contrary, increased risk for RA was also found for early menopause, which is associated with low levels of estrogen. Decreased risk for RA was found associated with duration of oral progestogen use in pre-menopause, supporting the immunosuppressive effect of progestogens on the immune system[46].

Following the genetic risk factors that are associated with RA, a role for the adaptive immune system is implicated. The classical immune feature in RA is the presence of anti-citrullinated protein antibodies and anti-immunoglobulin G (anti-Fcγ) antibodies known as rheumatoid factor (RF) in the blood. These autoantibodies can precede the clinical onset of RA by years[47–49].

The reason for the primary manifestation of the autoreactive response specifically in the joints is poorly understood. The site-specific manifestations include joint pain *(lat. arthralgia)*, inflammatory infiltration of the synovial tissue *(lat. synovitis)* and bone erosion.

The presence of arthralgia and anti-citrullinated protein antibodies increases the risk for the onset of RA[50]. Although there is no evidence for a direct effect of anti-citrullinated protein antibodies on sensory neurons, intravenous administration of pooled antibodies from anti-citrullinated protein antibody- (ACPA-)positive RA patients induced pain-like behavior in mice[51].

Synovitis in RA is characterized by hypertrophy of the synovial lining, neo-angiogenesis and infiltration of inflammatory molecules and immune cells. The composition of immune cells infiltrating the synovial tissue varies from scarce infiltration of monocytes or macrophages to the establishment of ectopic lymphoid structures perpetuating the adaptive autoimmune response[52]. These ectopic lymphoid structures primarily occur in the sublining of the synovial tissue, but can also occur in the lung[53] and the bone marrow[54].

Untreated RA eventually leads to bone destruction at the cortical areas of the bones and bone loss. This effect is pronounced in ACPA-positive RA patients[55]. Mechanistic studies have confirmed that ACPAs can directly bind and activate osteoclasts contributing to the clinical effect[56].

A summary scheme of the proposed etiology of RA is shown in Figure 2.

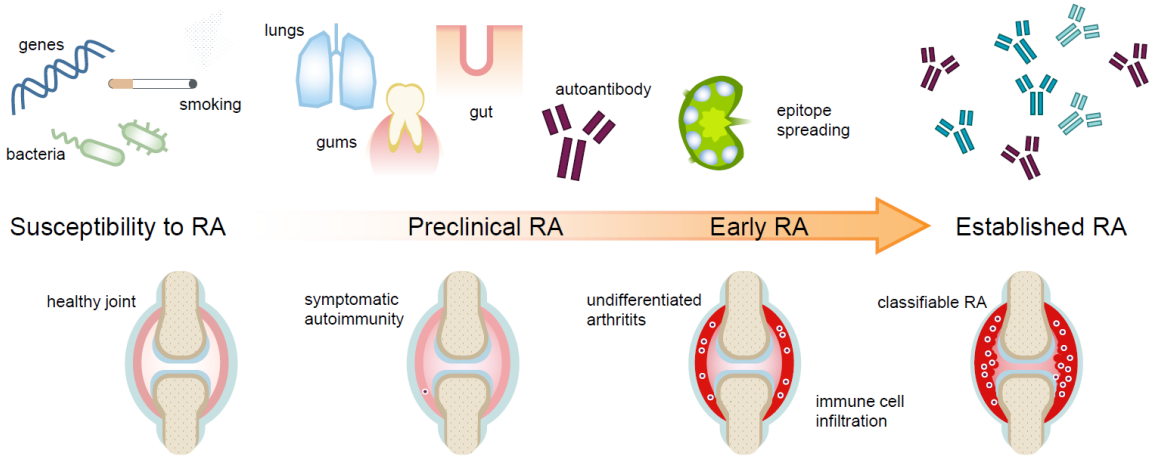


Figure 2 RA is a complex immune-mediated disease that is triggered by environmental stimuli in genetically susceptible individuals. Autoimmunity likely develops around the surfaces, where there is interaction with the environment. Just before the onset of RA, the immune response is perpetuated, which can be probed as epitope spreading. In established RA, the joints are marked by inflammation and an autoantibody response in circulation.

2.1.5 ACR/EULAR 2010 classification criteria

In order to conduct comparable clinical studies on RA around the globe, the American College of Rheumatology/European League Against Rheumatism proposed RA classification criteria that help to select suitable study cohorts[57]. These criteria have been developed in a data-driven consensus process and include clinical manifestations as well as serological components as in Table 1. The classification criteria may inform the clinical diagnosis by the rheumatologist, but it should be noted that the classification criteria are intended to stratify patient groups with high

specificity rather than to exert high sensitivity when selecting patients. In clinical diagnostics, ACPA reactivity is evaluated by a version of the cyclic citrullinated peptide (CCP) assay.

A. Joint involvement (swollen or tender)	
1 large joint	0
2-10 large joints	1
1-3 small joints	2
4-10 small joints	3
> 10 joints and ≥ 1 small joint	4
B. Serology	
RF- and ACPA- (seronegative)	0
RF-low or ACPA-low	2
RF+ or ACPA+ (seropositive)	3
C. Acute-phase reactants	
Normal CRP and ESR	0
Abnormal CRP or ESR	1
D. Duration of symptoms	
< 6 weeks	0
> 6 weeks	1

Table 1 The American College of Rheumatology/European League Against Rheumatism (ACR/EULAR) RA classification criteria can be applied to individuals presenting with ≥ 1 swollen joint. The individual is classified with RA with a score ≥ 6.

2.2 INNATE AND ADAPTIVE IMMUNE RESPONSE

The immune system can be divided into the innate and the adaptive immune branches. Innate immune responses are at their full effect within hours or days and recognize common structures, whereas adaptive immune responses take longer to mobilize, but are specific and create memory[58]. For a summary, compare Figure 3.

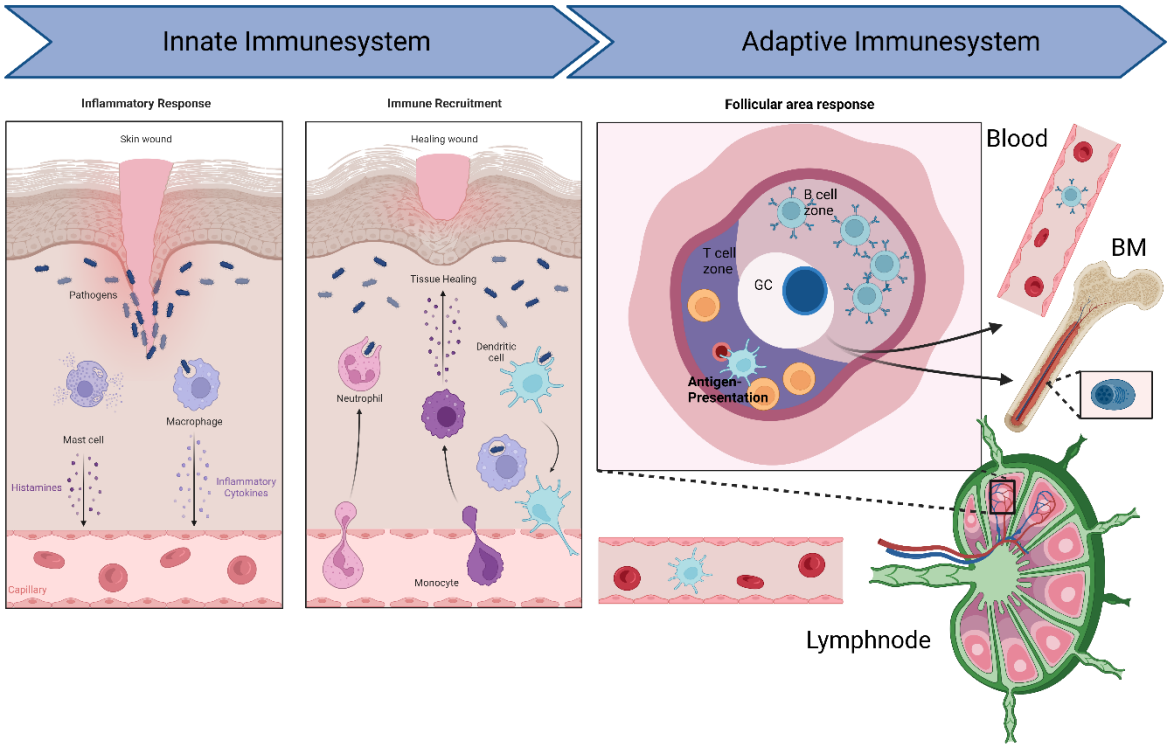


Figure 3 The scheme shows the innate and adaptive immune response in a nutshell. The inflammatory response leads to recruitment of monocytes and neutrophils. Dendritic cells carry the antigen to the secondary lymphoid organ (lymphnode) and presents to T cells. T and B cells meet, so that a germinal center (GC) with a dark and a light zone evolves. Memory B cells act as sentinels in circulation, while plasma cells home to the bone marrow (BM), where they can survive for decades.

The innate immune system involves primary physicochemical barriers such as the epithelial layer, secreted mucous and gastric acid. On a cellular basis, the first line of defense mainly involves tissue resident macrophages. These cells sense pathogens through their pattern recognition receptors (PRR)[59]. Upon stimulation, they release cytokines and chemokines that increase blood vessel permeability and attract larger numbers of neutrophils and monocytes that will further differentiate into macrophages. This process is recognized as inflammation. Macrophages and neutrophils further release peptides contributing to inflammation. Both macrophages and neutrophils also recognize pathogens through their complement receptors[60]. Complement is a system of proteolytic plasma proteins that opsonize microbial surfaces. Macrophages and neutrophils engulf their target and phagocytose it.

Antigen presentation bridges innate and adaptive immune responses. Dendritic cells are classical antigen presenting cells (APCs)[61]. Similar to macrophages, dendritic cells within the tissue recognize pathogens through PRRs, but they are also able to engulf pathogens in a receptor-independent manner by a process termed macropinocytosis. However, after engulfing the pathogen, it is not the primary goal to destroy it, but to carry processed antigen peptides to peripheral lymphoid organs and present them to T cells[62]. Antigen presentation occurs on a groove in class I and class II major histocompatibility complexes (MHC)[63]. In humans, major histocompatibility complexes are also termed human leukocyte antigen (HLA) molecules. As these terms allude to, these complexes are encoded by highly polymorphic allelic variants accounting for a high interindividual variability. All nucleated cells apart from homeostatic myocytes present endogenous peptides in steady state on class I histocompatibility complexes.

Only professional antigen presenting cells can present exogenous peptides on class II histocompatibility complexes. Professional antigen presenting cells include dendritic cells, macrophages and B cells.

T cells[64] in the paracortical zone of a secondary lymphoid organ that recognize antigen on class II MHC through their cognate T cell receptor (TCR) are primed to mediate T cell help in germinal centers (GCs). T cells that recognize peptides on class I MHC are able to exert cytotoxicity in tissues. Upon chronic antigen stimulation, T cells recognizing class II MHC can also acquire cytotoxic properties[65,66]. T cell receptors are inherently diverse and are subjected to selection pressure upon development in the thymus and the periphery. Adapted clones that persist after immune responses can form memory, which can be recalled upon recurrent immune challenge.

B cells[67] recognize antigen through their B cell receptor (BCR), but also have important roles in antigen presentation and cytokine production. The secretory counterparts of the B cell receptor are antibodies that are typically secreted by plasma cells that are terminally differentiated B cells. Similar to TCRs, BCRs are inherently diverse and are subjected to selection pressure, first during development in the bone marrow and later in germinal centers in the periphery. Adapted and persisting clones act as sentinels in circulation and tissues, or reside in the bone marrow in the form of long-term memory.

2.2.1 Roles and stages of the B cell

To study autoimmune B cells and their BCRs in RA stems from the idea that these components are pathogenic. It is clear that B cells actively participate in disease processes as evidenced by the efficacy of anti-CD20 therapy, rituximab. Surprisingly, the success of B cell depletion therapy is not due to the elimination of autoreactive antibodies, as plasma cell numbers are maintained during treatment[68]. B cells are a highly heterogeneous population and the manner of their involvement in pathogenesis is not completely clear.

2.2.1.1 B cell differentiation

B cell development starts in the fetal liver and is continued in the bone marrow throughout life. B cells differentiate from hematopoietic stem cells to immature B cells in a number of distinct steps (see Figure 4). These processes are characterized by the expression of surface markers, the assembly of the B cell receptor and selection mechanisms. Hematopoietic stem cells (HSC) are characterized by CD34 expression and differentiate into $CD127^+$ common lymphoid progenitors (CLP). These differentiate further into $CD34^+CD38^+CD10^+CD19^+CD127^+$ progenitor (pro-) B cells[69]. During this stage V(D)J recombination is started by rearrangement of the D_H and J_H segments followed by rearrangement of the V_H gene segment to the D_HJ_H segment. After successful recombination, the immunoglobulin heavy chain (IgH) associates with the surrogate light chain leading to the translocation of the pre-B cell receptor complex to the cellular membrane, where they associate with the Igα and Igβ chains. Pre-B cell receptor complex signaling is required for precursor (pre-) B cell survival and serves as a checkpoint for successful recombination[70]. Thereafter, in the pre-B cell stage, rearrangement of the

immunoglobulin light chain (IgL) gene segments occurs. The Igκ gene loci are preferentially rearranged before the Igλ loci. This sequential recombination ensures that a given B cell will only express one B cell receptor. The underlying mechanism is referred to as allelic exclusion[71]. After expression of a functional B cell receptor, the pre-B cell becomes an immature B cell that is $CD10^{+}CD19^{+}CD20^{+}CD127^{+}sIgM^{+}$. These immature B cells leave the bone marrow to mature in an antigen-dependent manner in the periphery, where they also start to express surface immunoglobulin D (sIgD) B cell receptors.

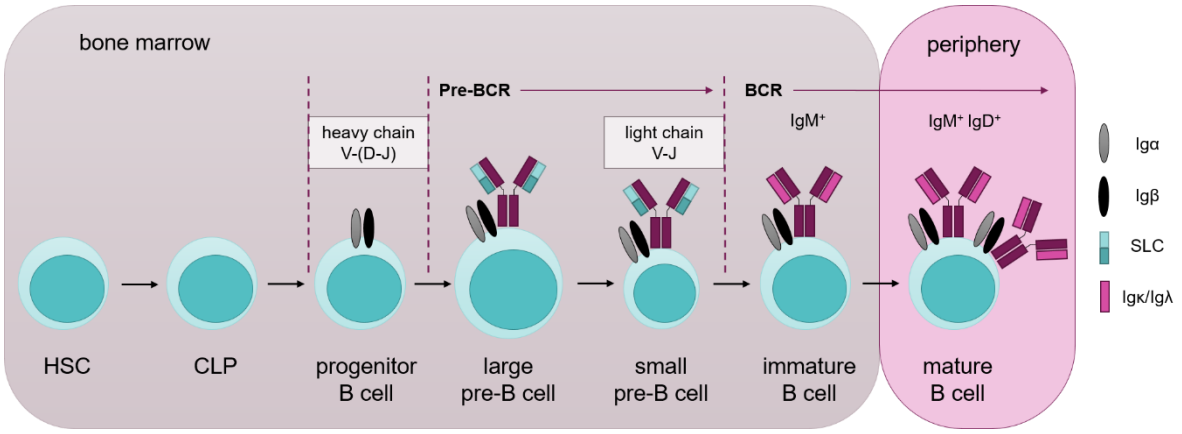


Figure 4 Stages of the B cell development and maturation process are shown. Hematopoietic stem cells (HSC) differentiate into common lymphoid progenitors (CLP), before entering the B cell lineage. During the progenitor B cell stage, V-(D-J) rearrangement of the heavy chain occurs. The Pre-BCR with the surrogate light chain (SLC) forms in the pre-B cell stage. Rearrangement of the light chain leads to expression of the BCR as surface IgM. The immature B cell differentiates into a mature B cell, which can be traced by surface IgD expression.

2.2.1.2 T cell-independent responses

In T cell-independent B cell responses, we can differentiate between T cell-independent type I and type II antigens. T cell-independent type I antigens, such as LPS, CpG and viral RNA signal through Toll-like receptors (TLR) to activate B cells[72,73]. Type II antigens are multivalent antigens that extensively ligate B cell receptors to activate B cells. These include bacterial capsular polysaccharides from *Streptococcus pneumoniae*, *Haemophilus influenza* type b, and *Neisseria meningitides*, but also repetitive determinants on viral particles[74,75].

B cells that respond in a T-cell independent manner include B-1 cells and marginal zone B cells (MZBs). B-1 cells have predominantly been described in the mouse, but are thought to be translatable to a particular phenotype in humans[76]. The generation of B-1 cells occurs mostly in the fetal liver and a stable B-1 cell population is maintained *in situ*[77]. They reside in the peritoneal and pleural cavities and spontaneously secrete polyreactive and broadly neutralizing IgM, and are therefore the main source of natural antibodies found in serum[78]. These natural antibodies are specific for self-antigens[79,80] and common bacterial and viral antigens containing multivalent epitopes[81]. Thus, B-1 cells are important for homeostasis and providing a first line of defense against infections.

MZBs are important for humoral immune responses against blood-borne pathogens and can rapidly differentiate into plasmablasts and short-lived plasma cells. They populate the splenic marginal sinus, express CD1d allowing them to present lipid antigens to NK T cells and

produce anti-lipid antibodies[82]. Upon capture of antigen bound as immune-complexes, MZBs can shuttle to the follicular area for antigen presentation to naïve T cells and promote T cell dependent responses[83,84].

2.2.1.3 T cell-dependent responses

The hallmark of T cell-dependent B cell responses are germinal center reactions. T cell-dependent B cell responses are elicited by protein antigens recognized by the B cell receptor and presented to T cells as an MHC-peptide complex. Besides the primary antigen signal, full B cell activation occurs upon secondary and tertiary signaling via CD40L-CD40 interaction and secretion of cytokines such as IL-21, IL-4 or IFN-γ[85,86]. In a germinal center reaction, interactions between cognate B and T follicular helper cells and follicular dendritic cells are facilitated in a structural microenvironment[87]. These interactions support affinity maturation and selection in the light zone of the germinal center[88]. In the dark zone of a germinal center, B cells proliferate and undergo somatic hypermutation (SHM). Germinal center B cells induce the expression of the enzyme activation induced cytidine deaminase (AID), which mediates SHM and class-switching[89–91]. The outcome of a germinal center T cell-dependent response are memory B cells that circulate throughout the periphery and induce rapid recall responses and long-lived plasma cells that home to the bone marrow niche, where they can confer protective immunity for years.

In RA, B cell responses can also be elicited by help from T peripheral helper cells in the inflamed synovium. T peripheral helper cells are like T follicular helper cells PD-$1^{hi}CD4^{+}$ cells that produce IL-21 and CXCL13[92]. However, they differ in their phenotype especially with respect to CXCR5, migratory capacity and transcriptional regulation[93]. T peripheral helper cells have now also been detected in other conditions including lupus, type I diabetes, celiac disease, autoimmune hepatitis, allograft rejection as well as in HIV and some tumors[94–97].

A summary scheme of the B cell activation and differentiation is shown in Figure 5.

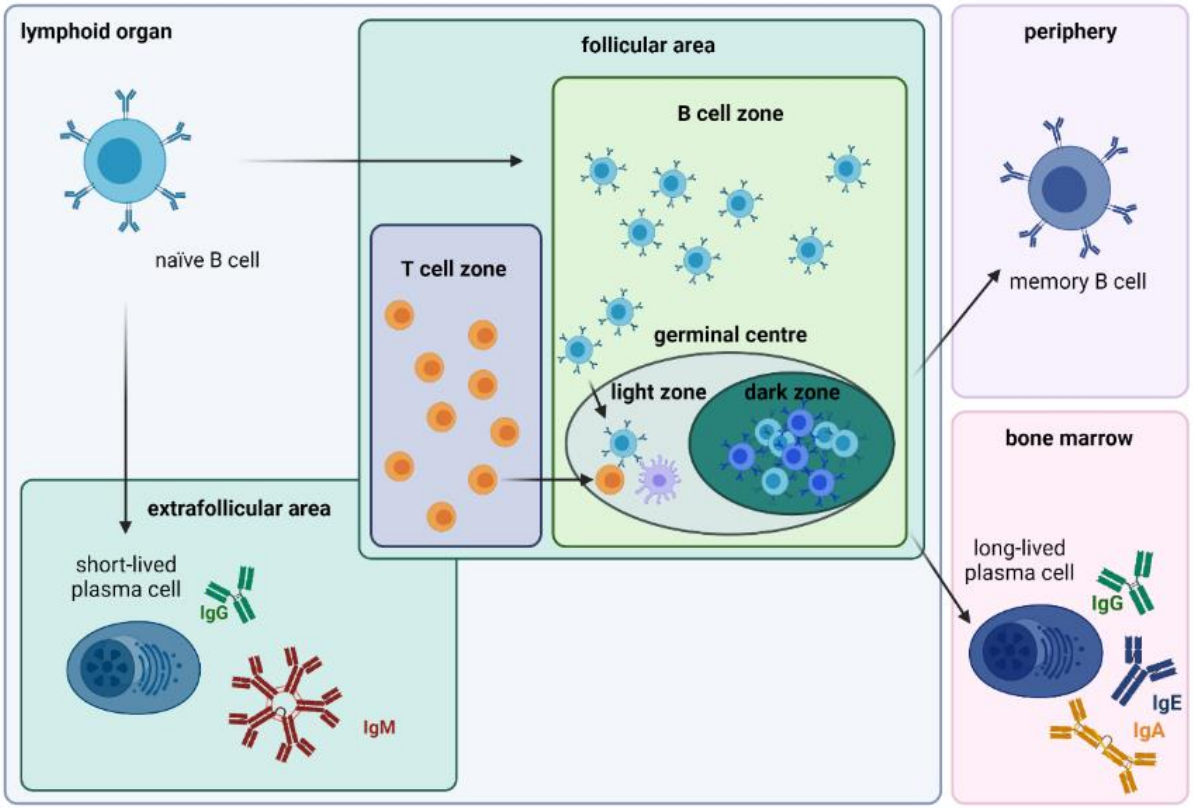


Figure 5 B cell activation and differentiation are shown.

Memory B cells are a heterogeneous population of cells that provide rapid and enhanced secondary immune reactions. Memory B cell generation depends on epigenetic imprinting of antigen-experienced B cells after a germinal center reaction[98]. In the memory phase, cells are resting, but upon antigen recall responses, they can quickly differentiate into antibody secreting plasma cells. Memory B cells reside in distinct tissues, but also circulate throughout the body. A prominent memory B cell population is found in the spleen[99].

Failure to eliminate antigen coincides with an increase of a subpopulation of memory B cells termed atypical memory B cells. Historically, this population has been considered anergic or exhausted as it has been found in chronic infection[100–102] and autoimmune disease[103–107]. Consistent with this idea, atypical memory B cells express high levels of inhibitory receptors, such as those belonging to the Fc-receptor-like (FCRL) family molecules, have muted BCR signaling and limited capacity to differentiate into plasma cells following BCR stimulation *in vitro*[101,108,109]. However, data from systemic lupus erythematosus patients suggest that these cells are short-lived activated cells, in the process of differentiating into plasma cells[110]. Numerous studies on malaria[111], influenza[112,113] and HIV[114] suggest that these cells are part of the functional antigen-specific immune response. Notably, in seropositive RA, it was shown, that there was an increase in $CD27^{-}IgM^{-}IgD^{-}$ B cells exhibiting a signal similar to activated or memory B cells. The increase of those cells in seropositive RA may indicate a more constant antigenic exposure, for example to citrullinated peptides, compared to seronegative patients. Those findings are consistent with a previous report that $CD27^{-}IgD^{-}$ (double negative, DN) B

cell numbers are increased in RA. This report equally showed, that the $CD27^-$ class-switched subset exhibits high levels of SHMs[115]. Moreover, these cells had a decreased frequency of the CD21 marker compared to healthy donors[116]. A low frequency of the CD21 marker also coexist with B cells expressing an autoreactive BCR in RA[103]. Atypical memory B cells have been shown to secrete autoimmune antibodies targeting erythrocyte membrane proteins leading to anemia[117].

In more granular analyses, autoimmune-/age-associated B cells (ABCs) have been described as TLR7/IL-21 driven $CD19^+IgD^-CD27^-CD21^-CD11c^+T\text{-}bet^+$ cells[118–122]. In RA, ABCs have been associated with disease activity measured by DAS28 scores, with treatment response and with T follicular helper cells in PBMCs[120]. Additionally, ABCs may contribute to fibroblast-like synoviocyte activation via TNFα-mediated ERK1/2 and JAK-STAT1 pathways[123]. These findings are consistent with the idea that ABCs contribute to autoimmunity not only by production of autoantibodies, but also by production of inflammatory cytokines and/or the presentation of antigen to and stimulation of T cells[124]. Another more granular analysis of similar cell subsets in systemic lupus erythematosus (SLE) has led to the definition of the DN1 and DN2 phenotype. DN1 are described as B cell follicle homing $CD19^+IgD^-CD27^-CXCR5^+$ cells, while DN2 are described as TLR7/IL-21 driven $CD19^+IgD^-CD27^-CD21^-CXCR5^-T\text{-}bet^+$ cells[110].

Plasmablasts are short-lived cycling antibody-secreting cells that are precursors of plasma cells. Plasmablasts typically occur in the circulation one to two weeks following infection or vaccination secreting large amounts of pathogen-specific antibodies[125–128]. Plasmablasts blasting out of switched memory B cells are mostly of the Immunoglobulin G (IgG) isotype.

Plasma cells are post-mitotic antibody secreting cells and can be distinguished into short-lived and long-lived plasma cells. They exist in all lymphoid organs, but the bone marrow is considered the site for their long-term maintenance over months, years or a lifetime.

Short-lived plasma cells arise from T cell-independent responses in the extrafollicular area of lymphoid organs. They typically secrete more antibodies of the Immunoglobulin M (IgM) isotype, but can also class-switch to the IgG isotype. The rapid generation of these short-lived plasma cells is essential for the early control of infections, while the adaptive T cell-dependent response is initiated.

Long-lived plasma cells arising from a germinal center response migrate to the bone marrow in a CXCR4-CXCL12 dependent fashion[129]. Signaling downstream of CXCR4 needs to be intact for terminal differentiation and longevity in the bone marrow niche[130,131].

In our studies (see below), we use known cluster of differentiation (CD) surface markers to define B cells. Alive B cells can be defined by lack of expression of phosphatidyl serine on the cell surface measured by absence of Annexin V staining, lack of CD3 and CD14 expression, which mainly define T cells and monocytes, and presence of CD19. We further define naïve B cells by expression of surface immunoglobulin D (sIgD) and lack of the memory marker CD27

and memory B cells by lack of sIgD expression and presence of CD27 on the cell surface. In CD27 high expressing B cells, we further distinguish between CD138- cycling plasmablasts and CD138+ non-cycling plasma cells.

2.3 IMMUNOGLOBULINS

2.3.1 Functional structure

Immunoglobulins have a dual role in immunity. They recognize antigens and mediate effector functions. In general, immunoglobulins have a Y-shaped configuration and consist of two identical heavy chains (50-70 kDa) and two identical light chains (25 kDa). The F_{ab}s (fragment antigen binding) contain the N-terminal part of the heavy chains associated with the light chains. The F_c (fragment crystallizable) is formed by the C-terminal part of the heavy chains. The F_{ab}s mediate recognition of antigens, while the F_c interacts with effector molecules such as C1q and F_c receptors triggering elimination mechanisms such as phagocytosis and lysis.

Human immunoglobulins (Igs) are present in 5 different isotypes, IgM, IgD, IgG, IgA and IgE. These differ in their heavy chain termed μ, δ, γ, α and ε heavy chains, respectively. The IgG class is further divided into IgG1, IgG2, IgG3 and IgG4, and the IgA class is further divided into the IgA1 and IgA2 subclasses. Allotypic variants that have been recognized as markers include the γ*1m*, γ*2m*, γ*3m*, α*2m* and *ε m* variants.

There are two types of light chains, κ and λ, that can associate with any of the heavy chain types. The κ to λ ratio in human serum of healthy individuals is about 2 to 1. There are three κ *m* allotypes and four λ isotypes.

Moreover, each chain folds up into domains consisting of 100 to 110 amino acids. The heavy chain contains four or five domains, whilst the light chain contains two domains. The N-terminal domain of both chains is highly variable and contains a β-sheet framework supporting 3 hypervariable loops or the complementary determining regions (CDRs) constituting the antigen binding site. The other domains are termed constant domains and display a modular structure with homologous elements shared by the different isotypes. The domains of immunoglobulins associate through multiple non-covalent lateral interactions and are stabilized by branched carbohydrate chains in the F_c fragment. A scheme of the IgG structure is shown in Figure 6.

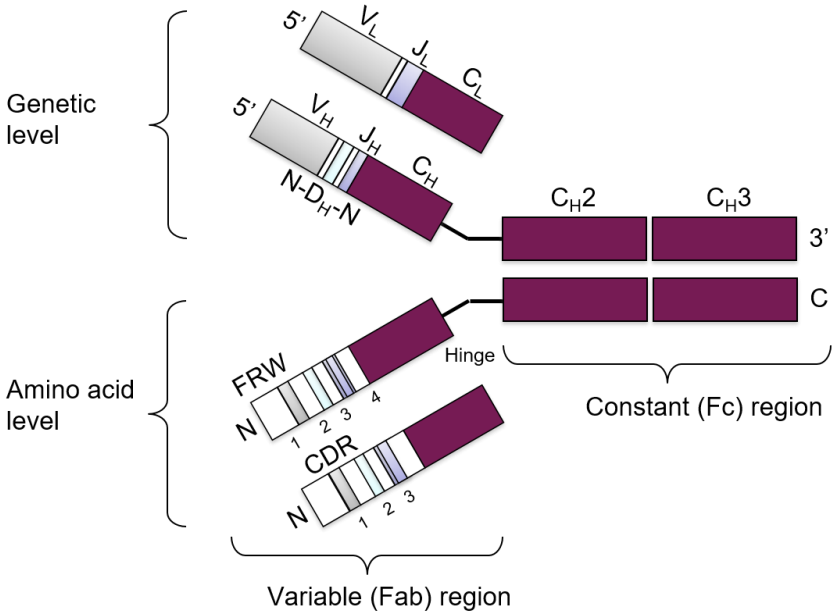


Figure 6 The immunoglobulin G antibody represented on a genetic level (top) and on amino acid level (bottom). The antibody is constituted of V, D and J segments for the heavy and light chain. On the amino acid level, the variable region can be defined by framework regions (FRW) and complementary determining regions (CDR).

Membrane immunoglobulins are present as identical monomers on the surface of a given B cell and serve as specific antigen receptors. The monomers associate with CD79a and CD79b molecules bearing immunoreceptor tyrosine activation motifs (ITAMs) that get phosphorylated upon activation of the B cell receptor leading to B cell activation.

Secreted IgG, IgD and IgE occur as monomers, while IgM presents a pentameric structure. Notably, IgD is mainly occurs as surface receptor. IgA is monomeric in serum and dimeric in seromucous secretions.

2.3.2 Loci

The immunoglobulin loci have unique features contributing to the recognition of diverse antigens. Antibodies are encoded by variability (V), diversity (D), joining (J) and constant (C) segments for the heavy chain and V, J and C segments for the light chain that both are recombined sequentially upon development of individual B cells. A summary of the variable loci composition is shown in Figure 7.

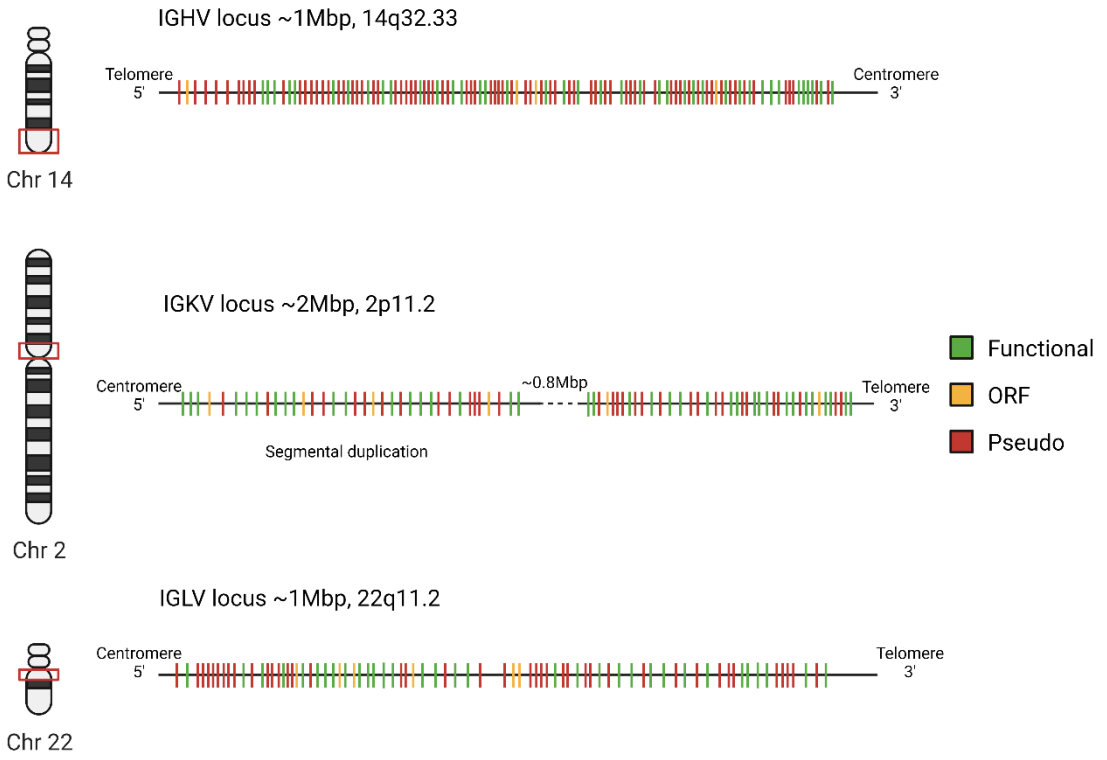


Figure 7 The scheme shows the composition of the variable heavy/ κ/ λ chain locus. They all contain many functional genes (Functional), open reading frames (ORFs) and pseudogenes (Pseudo).

The immunoglobulin heavy chain locus IGH spans approximately 10^6 base pairs on the telomeric end of the long arm of chromosome 14 and shows high interindividual variation with frequent duplications, deletions and multiple allelic variants of each segment. It includes 123-129 IGHV segments consisting of about 50 functional genes, 4-5 open reading frames and 79-81 pseudogenes. Further downstream, it includes 27 IGHD segments consisting of 23 functional genes and four open reading frames. The downstream sequence includes nine IGHJ segments with six functional genes and three non-functional pseudogenes. The IGHC locus furthest downstream includes 5-11 genes consisting of 5-9 functional, 0-1 open reading frames and 0-1 pseudogenes[132]. The international ImMunoGeneTics (IMGT) database nomenclature[133] of the IGHV and IGHD segments is in principle based on a digit designating a gene family, followed by a dash, followed by a digit designating the chromosomal position. There are exceptions to these rules. The digit of the IGHJ segments designates homology and position at once.

The immunoglobulin κ light chain locus is located on chromosome 2p11.2 and spans almost $2*10^6$ base pairs. It consists of two IGKV clusters that are presumably the result of a large duplication event and that are separated by $0.8*10^6$ base pairs. Not all haplotypes present both clusters. The distal IGKV cluster in the centromeric position contains 36 genes including 12-18 functional genes, 5-6 open reading frames and 13-18 pseudogenes. The genes are labeled with a D after the homology based family name. The proximal IGKV cluster in the telomeric position contains 40 genes including 17-19 functional genes, 4 open reading frames and 17-19

pseudogenes that are upstream of five functional IGKJ gene segments and the unique IGKC gene segment.

The immunoglobulin λ light chain locus spans approximately 10^6 base pairs and is located on chromosome 22 at band 22q11.2. It consists of 73-74 IGLV genes and 7-11 IGLJ genes that are interspersed between 7-11 IGLC genes. In contrast to the IGH and the IGK locus, the IGL locus is oriented in a forward direction. That means that the IGLC gene segments are the most telomeric. The IGLV locus has been divided in 3 IGLV gene clusters based on their IGLV family[134]. The IGLV gene segments contain 29-33 functional genes, 5-6 open reading frames and 35-39 pseudogenes. The IGLJ and IGLC gene segments contain 4-5 functional genes. The IGLJ locus contains 3 additional open reading frames, while the IGLC locus contains 2-3 additional pseudogenes.

2.3.3 Structure of the VDJ genes

The immunoglobulin genes occur in three configurations. They are present as genomic DNA in the germline, they occur rearranged on genomic DNA in individual B cells and they occur rearranged as mRNA in these B cells upon expression[135]. Notably, if they occur as rearranged mRNA, differential processing determines, whether the immunoglobulins will be membrane-bound or secreted.

An immunoglobulin germline V gene consists of a 5' UTR and a leader sequence. The leader sequence is variable and comprises the first part of the leader exon with a 3' terminal donor splice site, an intronic region and the second part of the leader exon with a 5' terminal acceptor splice site. At the 3' terminal end of the second part of the leader, the exon of the variable region starts. The V exon encodes the β-sheet framework supporting the three CDRs as mentioned before. Among the more conserved encoded residues, there are two conserved cysteines in position 23 and 104 that form an intramolecular disulfide bridge and a couple of hydrophobic amino acids in the FRWs including positions 4, 41, 50 and 100. At the 3' end of the V exon a recombination signal sequence follows. This site includes a conserved palindromic heptamer at its 5' end, a spacer of 23±1 or 12±1 base pairs and a conserved adenosine or thymidine rich nonamer in its 3' end. The 3' end of the V segment is terminated by a 3' UTR (see Figure 8).

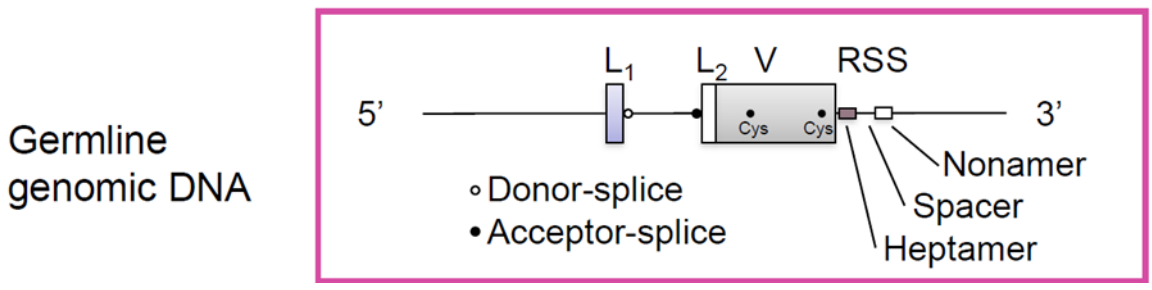


Figure 8 A representative V gene in the germline genomic configuration is shown.

The immunoglobulin germline D gene segments of the heavy chain contain recombination signal sites at both ends of the D exon. The conserved nonamers of the recombination signal sequences are at the 5' and 3' end. These are flanking the spacer and the conserved heptamers

that are located adjacent to the D exon. The immunoglobulin germline J gene segments contain their recombination signal sequences at the 5' end of the exon. The conserved nonamer is located upstream of the 23±1 or 12±1 base pair spacer that is followed by the conserved heptamer downstream. Adjacent to the 3' end of the J gene segments, there is a donor splice site for the acceptor splice sites at the constant chain segments.

The enzymatic rearrangement follows the 12/23 rule. That means that only gene segments that are flanked with a 12 base pair (bp) spacer in their recombination signal sequence can be recombined with a gene segment that is flanked with a 23 bp spacer in their recombination signal sequence and vice versa. This rule secures the correct rearrangement of the immunoglobulin gene segments. The rearrangement of the immunoglobulin VDJ segments on the DNA level occurs sequentially. First, the immunoglobulin heavy chain locus (IGH) is rearranged. The rearrangement starts with the joining of the IGHD and IGHJ gene segments, which is followed by the joining of the IGHV gene segment and the combined DJ segments. These rearrangement steps are accompanied by the deletion of the intermediary DNA as an excision loop.

Upon successful rearrangement and expression of the immunoglobulin heavy chain within the pre-B cell receptor, the rearrangement of the IGH locus is terminated and the rearrangement of the IGKV and IGKJ gene segments is initiated. If the rearrangement is productive, the IGL locus remains non-rearranged. If the rearrangement of the IGK locus remains unproductive, an IGLV gene segment is recombined with an IGLJ segment. This chronology is referred to as isotypic exclusion of the light chain. The rearrangement of either light chain is equally accompanied by the deletion of the intermediary DNA as an excision loop. Rearrangement occurs on one chromosome only. This feature is referred to as allelic exclusion and leads to the fact that generally only one chromosome 14 encodes a functional heavy chain and only one chromosome 2 or 22 encodes a functional light chain. Immunoglobulin genes on the other chromosomes are either non-rearranged or rearranged but non-productive or deleted. The rearranged genomic DNA in a B cell thus contains a 5' UTR, the first part of the leader exon with the donor splice site, the intronic region within the leader, the second part of the leader exon with the acceptor splice site adjacent to the intron, followed by the exonic rearranged VDJ segments in the heavy chain or the exonic rearranged VJ segments in the light chain with a donor splice site at the 3' end upstream of the 3' UTR (see Figure 9).

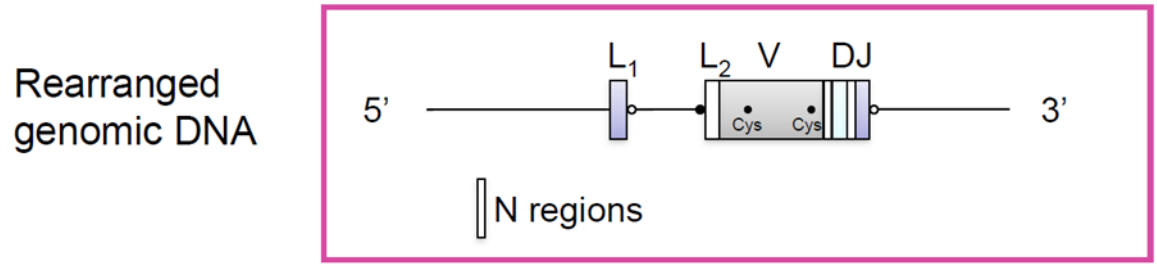


Figure 9 The rearranged genomic DNA of the immunoglobulin locus in a B cell is shown.

The rearranged VDJ segments and the IGHC segments or the rearranged VJ segments and IGKC or IGLC gene segments are then transcribed into pre-messenger RNA (pre-mRNA). The intronic segments within the leader and within the constant gene segments as well as the non-used J segments are excised by splicing. The mature mRNA thus contains a capped 5' UTR and poly-adenosine tailed 3' UTR and the spliced coding regions that are translated by the ribosomes (see Figure 10).

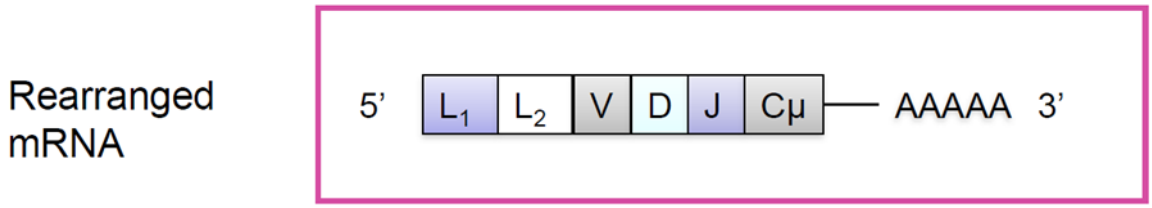


Figure 10 The mature IgM messenger RNA expressed in a B cell is shown.

The concept of the VDJ genes and their recombination form the molecular basis for the variability of the antigen recognition of an antibody. The different configurations of the immunoglobulin genes are crucial for understanding immunoglobulin sequencing. In addition to VDJ recombination, each rearranged VDJ (or each given BCR) sequence diversifies by additional processes as the B cell response evolves.

2.3.4 Diversification

The diversity of the immunoglobulins depends on several properties. This includes (1) the combinatorial and (2) junctional diversity upon VDJ or VJ recombination, (3) alternative splicing of the VDJ heavy chain pre-mRNA and the AID-mediated (4) SHM and (5) class switch recombination (CSR).

VDJ and VJ recombination is mediated by a recombinase complex that brings the recombination signal sites of two gene segments into close proximity. Thus, the first diversification of the F_{ab} of immunoglobulins is mediated by the number of available gene segments that are flanked with the correct recombination signal sites.

In a second step, the recombinase complex generates a DNA hairpin at the flanking sites of the coding ends. A nuclease then opens the hairpins at the coding ends generating palindromic DNA overhangs. The ends are modified by a deoxynucleotidyl transferase adding non-templated nucleotides at both overhangs. Eventually, the single-stranded overhangs pair, and unpaired nucleotides are removed by an exonuclease. The gaps are filled with nucleotides and ligated to form a coding joint. Thus, the Fab of immunoglobulins are further diversified by their junctional sequence.

Alternative splicing of the pre-mRNA further diversifies the heavy chain constant part by usage of an alternative 3' terminal exon. Membrane-bound heavy chains are anchored by a hydrophobic C-terminal sequence, while secreted immunoglobulins have a hydrophilic end. The co-expression of IgM and IgD is only possible due to the synthesis of a long VDJCμCδ

pre-mRNA that is alternatively spliced upon maturation. Other immunoglobulin isotypes are not co-expressed.

The next classical Fab diversifying mechanism is termed “somatic hypermutation” (SHM). SHM is mediated by AID. Upon immunoglobulin transcription in maturing B cells, AID binds to single stranded DNA and converts cytosine to uridine. The presence of uridine in DNA and the mismatch with the complementary strand of the DNA triggers mismatch- and base-excision repair pathways introducing mutations in the variable region.

The last immunoglobulin diversifying mechanism is termed “class switch recombination” (CSR). CSR is equally triggered by AID upon immunoglobulin transcription. Initially, all B cells express the membrane bound IgM and D and secrete IgM. Upstream of the IGHM, the IGHG, the IGHA and the IGHE gene segments, there are intronic interspersed switch regions. A switch region is about 2,000 base pairs long and contains tandem-repeated 20-80 nucleotide long motifs. Attack of AID in single-stranded R-loops of this region results in double-strand breaks and non-homologous recombination between the switch regions deleting intervening sequences and forming chimeric junctions. Thus, a different immunoglobulin isotype will be transcribed.

Taken together, the large initial variation of the germline immunoglobulin genes and the additional diversification processes result in a great inter- and intraindividual variation of immunoglobulins and hence an antigen selected antibody response. The diversification processes in the variable region i.e. the combinatorial, the junctional and the SHM processes are stored in the sequences of antibodies and hence are important for understanding immunoglobulin sequencing studies.

2.4 AUTOREACTIVITY

2.4.1 Presence of autoreactive B cells in health and disease

Autoreactivity is common in early immature B cells. Mechanisms that are collectively referred to as central tolerance provide protection against overt autoimmunity. These mechanisms include receptor editing, clonal deletion and functional inactivation. Despite elimination of highly autoreactive B cell clones during B cell development, modest self-reactivity has been shown to persist in 15-20% of the mature B cell repertoire of healthy individuals[136,137]. One possible peripheral tolerance mechanism includes the absence of T cell help that would stimulate a germinal center reaction and potentially promote the survival of autoreactive clones. This is mediated by the transient state of the B cell after antigen processing for presentation on MHC class II. The antigen-activated BCR molecules transmit a range of intracellular signals[138] that prepare the B cell to respond to stimuli delivered through other cell surface receptors. However, the period during which antigen-activated B cells remain receptive to such signals is transient. An apoptotic response is initiated, when the B cell fails to engage with a T helper cell via the peptide-MHC-complex within 12-24h[139,140]. Thus, a majority of autoreactive B cells are cleared by the T helper cell self-tolerance mechanisms[141]. Notably, these mechanisms do not prevent any B cell cross-reactivity, where the T helper cell reacts to foreign antigen, whereas

the B cell also reacts to self-antigen. It has also been demonstrated that self-tolerance is trained in the germinal center reaction. Counter selection of self-reactive B cells generated within the germinal center has been directly demonstrated[142]. Still, autoantibodies associated with various autoimmune diseases do have a germinal center origin, indicating that germinal center self-tolerance is by no means absolute[143].

The autoreactivity of mature B cells is crucial for homeostasis. Polyreactive antibodies that are also autoreactive, ensure tonic low grade BCR signaling for B cell survival, while protecting from common pathogenic structures in rapid responses. Furthermore, autoreactivity is important for apoptotic cell clearance and helps to prevent formation of atherosclerotic plaques. In healthy individuals, specific autoreactivity is limited by regulatory cells[144]. Interestingly, the human IGHV4-34 gene has been associated with self-reactivity as antibodies using unmutated IGHV4-34 recognize l- and i-antigen on red blood cells[145]. However, it also remains unclear whether autoreactivity occurs beyond the antigenic spectrum probed.

It is unclear, whether escaped autoreactive clones can give rise to pathogenic autoimmune B cell responses, but the frequency of such clones is increased in patients with RA compared to healthy individuals[146].

2.4.2 Anti-citrulline immunity

Citrullination or deimination is a physiological enzymatic process catalyzed by calcium-dependent peptidyl arginine deiminases (PADs). Thereby, a positively charged peptidyl arginine is post-translationally converted into a neutral peptidyl citrulline (see Figure 11). In steady state, citrullination has important roles in the normal function of the immune system, skin keratinization, neuron insulation, the plasticity of the central nervous system and in gene regulation[147]. Citrullination can occur extracellularly, in the cytoplasm or in the nucleus and influences a protein's ionic and hydrogen-bond forming capacity, that in turn may influence the structure, function and antigenicity of a protein.

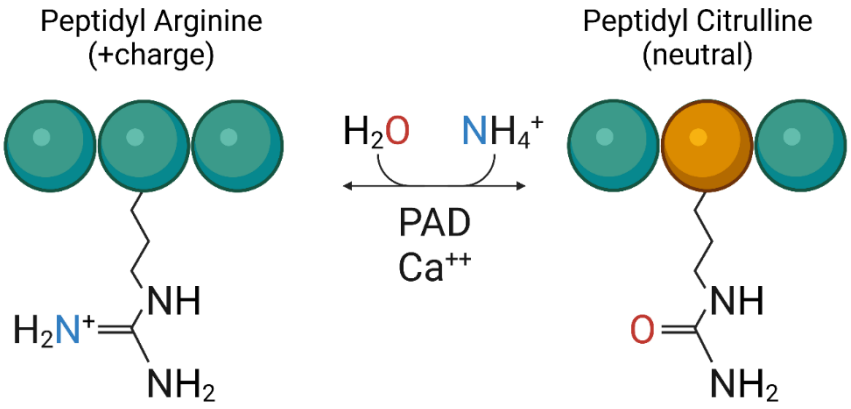


Figure 11 Peptidyl arginine deiminases convert peptidyl arginine to peptidyl citrulline in the presence of calcium ions.

PAD2 and PAD4 are so far identified in the RA synovium and their expression correlates with the degree of inflammation suggesting that these enzymes could play a role in RA

pathogenesis[148]. Furthermore, it has been reported that the periodontal pathogen *Porphyromonas gingivalis* expresses a PAD enzyme, denoted PPAD[42]. It has been proposed, that individuals predisposed to *Porphyromonas gingivalis* infection may be exposed to citrullinated antigens in the gum mucosa, that afterwards become systemic immunogens, that lead to systemic ACPA production and intraarticular inflammation[149].

ACPAs are a class of antibodies that recognize altered self in the form of citrullinated epitopes. Typical autoantigens include citrullinated fibrinogen[150], citrullinated vimentin[151], citrullinated collagen type II[152], citrullinated α-enolase[153] and citrullinated histone 4[154]. However, these antigen specificities are only proxies. ACPAs are multireacive[155] and the initiating antigen is unknown.

ACPAs represent the most specific humoral response in RA. The frequency of ACPA-positivity ranges from 50-60% in early RA and from 60-90% in long-standing RA[156–162]. Meanwhile, ACPAs are also detectable in about 1-3% of healthy individuals[23,163–165]. ACPAs do not only confer risk of RA in the pre-arthritis phase[166–168], but also confer risk of joint damage once disease is established[157,158]. ACPAs can trigger a variety of pro-inflammatory processes *in vitro* via interaction with the complement system or Fc receptors in the form of immune complexes[169–174]. It has also been suggested that ACPAs can interact directly with citrullinated antigens expressed by immune cells[169,175,176].

The initial trigger of the ACPA response is unknown. Individuals who develop arthritis have rising ACPA levels, an increase in the number of citrullinated antigens recognized and an expansion of immunoglobulin isotypes before the onset of the disease[177–182]. Currently, ACPA development and epitope spreading are described as the first and second hit in a pathogenesis model, where multiple immunological events are suggested. However, these events might not be sufficient to initiate and drive disease. Still, the ACPA response persists in established and clinically quiescent disease[183], putatively driven by ubiquitous presence of antigen[184]. B cells persist in peripheral blood at a frequency of about 1 in 12,500, which directly correlates with ACPA serum levels in RA patients[185]. A vast majority of these cells are class-switched post-germinal center memory cells and express IgG or IgA. About 7% of the circulating cells are plasmablasts[184], which more closely resembles the dynamics of antigen-specific responses upon secondary immunization. ACPA-expressing plasmablasts spontaneously secrete ACPA-IgG *ex vivo* and are more frequent in the synovial fluid than in peripheral blood[184,186]. The fate of these cells is unknown, but it is suggested that memory B cells or plasmablasts ‘carry’ the ACPA response to the site of inflammation in the joints.

ACPAs have one further quality that unites, but also distinguishes them from other antibody classes: ACPA-IgG molecules are glycosylated at the position 297 in the Fc tail[187], but also approximately 90% of ACPA-IgG are glycosylated in the variable domain[188–190]. Notably, of the 52 functional heavy chain variable region genes, only a few harbor germline encoded N-linked glycosylation sites. This points at either a restricted heavy chain variable region gene usage or the specific introduction of these sites during T cell-dependent SHM. As monoclonal antibodies from single cell studies have shown, that ACPAs use a variety of heavy chain

variable region genes and extensive SHMs[191], the latter option seems plausible. Further evidence is provided by a study showing that 88% of the ACPA-IgG clones in the repertoire had N-linked glycosylation sites[192]. In contrast, only 15-25% of conventional antibodies have variable domain glycosylation[193]. Additionally, ACPA-IgM molecules, which derive from B cells that presumably have not yet received T cell help, did not show increased glycosylation of the variable domain compared to non-ACPA-IgM[194]. It is suggested that ACPA-IgG harbor selective advantages for the survival and/or differentiation of ACPA-expressing B cells. ACPAs are low-avidity[195,196] and this quality could bypass potential conventional B cell selection mechanisms such as affinity maturation and as such lead to breach of tolerance.

2.4.3 HLA shared epitope

The MHC class II gene HLA-DRB1 is the strongest genetic risk factor for RA and many other autoimmune diseases. The HLA-DRB1*01, *04 and *10 alleles are associated with RA, especially ACPA-positive RA. These so-called SE alleles share sequences encoding five amino acids in position 70-74 of the third hypervariable region of the HLA-DR β chain[28]. A more recent study showed that amino acid 11 and 13 of the antigen-binding groove are associated with ACPA-positive RA[197]. It has been suggested, that citrullinated peptides bind to the positively charged binding groove of the HLA-DRB1*04:01 and HLA-DRB1*04:04 alleles and are presented to $CD4^+$ T helper cells[198,199]. An example of this presentation is shown in Figure 12. The HLA-DRB SE alleles do not only predispose for ACPA-positive RA, but also for disease severity[200–202] and not for ACPA positivity in healthy individuals[23,164]. This indicates that the initial breach of tolerance towards citrullinated proteins occurs independently of the HLA-DRB SE alleles, while the HLA-DRB SE alleles are important for perpetuation of the autoimmune response. Similar findings for the protective HLA-DRB1*13:01 allele support this notion[203,204]. However, a combination of the HLA-DRB1*13:01 and HLA-DRB1*03 alleles increase the risk for ACPA-negative RA[205]. The nature of the T cell help to ACPA-expressing B cells remains thus unclear. The possibility of non-self-reactive T cells providing help to autoreactive B cells by molecular mimicry offers a possible link to the microbiome or other sources of foreign antigens as drivers of autoimmunity. The presence of autoreactive T cells and the involvement of non-self-reactive T cells has been discussed elsewhere[43,206–214] and is reviewed[215].

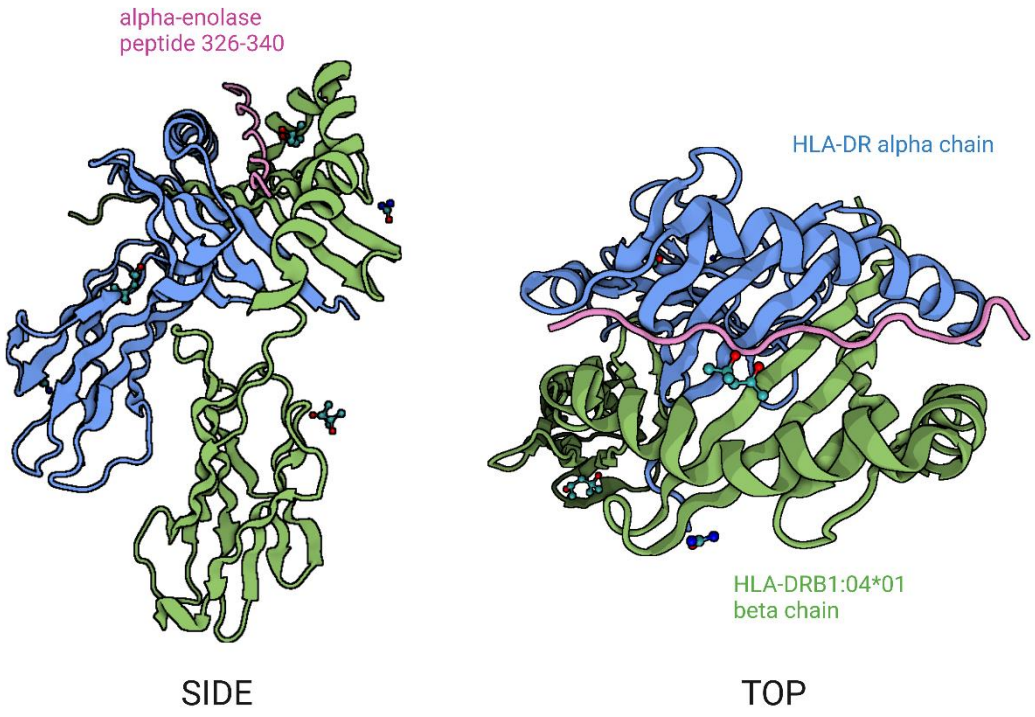


Figure 12 The ribbon diagrams show the HLA-DR4 molecules from the side and the top. The alpha-enolase peptide is positioned in the peptide binding cleft of the HLA class II molecule. (PDB-ID:5NI9[199])

2.5 SEQUENCING ANALYSIS

2.5.1 Rationale

Single cell transcriptomics is an excellent approach to study cellular phenotypes as well as B cell receptor sequences. However, the generation of full-length VDJ libraries coupled with bulk repertoire sequencing using the Illumina MiSeq platform to achieve sufficient read length can give more in-depth insight into phylogenetic relationships between antibody genes and B cell fates in the repertoire. These states can be analyzed in relation to the naïve immunoglobulin repertoire of an individual and thus give clues about the refinement of an antibody immune response. Repertoire studies do not give information about heavy and light chain pairing, but they do allow inference of the individualized germline configuration as a starting point of the VDJ analysis. Furthermore, single cell studies are often limited by the availability of single B cells and thus do not cover the same depth as unpaired expression immunoglobulin repertoire studies.

Measures that are currently used to characterize the genomic configuration of the immunoglobulin heavy chain locus include haplotyping. Haplotyping of the chromosomes is based on the concept of allelic exclusion. Another measure that is used to describe the immunoglobulin heavy chain repertoire is the SHM and the length of the CDR3. Both measures strongly depend on the correctness of the individualized germline database.

An important concept that has been discussed for expressed immunoglobulins is clonality. For single cell studies, the pairing of heavy and light chains gives additional information about the clonality of the B cell. B cells sharing clonal heavy and light chains are extremely likely to derive from a common selected B cell progenitor. However, this clonality is rarely seen in samples that are subjected to limited availability of material to start with. For in-depth bulk studies, a clone is often defined by sequences having identical CDR3s. However, this definition is not specific enough as sequences harboring identical CDR3s can have distinct V and J genes, pointing out a convergence of different B cell clones in their CDR3 and not the divergence of different B cell clones from a common ancestor. In our in-depth expression studies, we define clonally related sequences as having identical V and J gene usage and 80% identity of CDR3. The 80% homology of the CDR3 allows for some SHM over this region, which is important to cover mutated variants of a given immunoglobulin lineage, especially when sequencing IgG-switched B cells.

2.5.2 Single cell transcriptomics

Single cell RNA sequencing goes back to single cell quantitative polymerase chain reaction[216], alongside with single molecule fluorescent *in situ* hybridazation[217]. Subsequently, whole transcriptome analysis was performed using microarrays[218]. Advances in next generation sequencing (NGS) technology subsequently enabled bulk RNA sequencing that was quickly adapted to single cell approaches[219].

In essence, single cell RNA sequencing always consists of the capture and lysis of single cells, the amplification of minute amounts of mRNA, a pre-amplification, library preparation and sequencing. The data then requires in-depth computational processing making the interpretation of the data heavily reliable on bioinformatics.

Single cell capture can be performed by micromanipulation with micropipettes, laser capture microdissection, flow cytometry assisted cell sorting, micro droplets and micro fluidic systems. After lysis, reverse transcription can either be accomplished by first strand synthesis, poly-adenosine tailing and consequent second strand synthesis[219] or first strand synthesis with tailing, template switching and extension[220]. Most protocols currently use the Nextera kit for library preparation and the Illumina sequencing platform. Finally, single cell RNA sequencing data may allow the identification of cell types, but requires further experimental validation[221].

In 2014, the National Institutes of Health, the U.S. Food and Drug Administration, multiple biopharmaceutical and life science companies and non-profit organizations launched the Accelerating Medicines Partnership with projects in RA and SLE. As a result, Orange et al.[222] and Zhang et al.[223] published bulk and Cell Expression by Linear amplification single cell RNA sequencing data from synovial tissue in Arthritis & Rheumatology and Nature Immunology in 2019. The authors identified four distinct B cell clusters among 1,142 $CD45^+$ $CD3^-$ $CD19^+$ B cells passing quality control from 21 synovial tissue donors. These include 355 naïve $IGHD^+$ $CD27^-$ B cells (31%), 440 $IGHG3^+$ $CD27^-$ memory B cells (39%), 48 autoimmune-associated B cells with high expression of *ITGAX* (CD11c) (4%) and a plasmablast cluster of 299 B cells

(26%) with high expression of *IGHG* genes and *XBP1*, an important transcription factor for plasma cell differentiation.

In 2022, an update on these studies was published on BioRxiv by Zhang et al.[224]. The authors identified nine different B cell clusters with 30,697 B cells from 79 tissue donors that passed quality control. This study challenges the understanding of a 'cell state' as the authors describe the same B cell populations in a more granular fashion. Thus, naïve B cells are further divided into unswitched memory, naïve and marginal zone-like cells, and plasmablasts are further divided into IgG^+ plasma cells, HLA-DR^+ plasmablasts, IgM^+ plasma cells and germinal center-like cells. Autoimmune-associated B cells and switched memory B cells remain the same, but just gain more markers. Differentiation and functional assays are needed to support these specific distinctions. Visually, the UMAPs rather support only one clear distinction between B cells and plasma cells.

Notably, the authors recruited n=28 DMARD-naïve patients in their early disease course (mean 2.64 years), n=42 methotrexate or anti-TNF agent inadequate responders and n=9 osteoarthritis patients as controls.

In our studies (see below), we used IgD and CD27 to index sort $CD19^+$ B cells from treatment naïve early Rheumatoid Arthritis for single cell RNA sequencing. Thus, it is interesting to compare the profile of our single B cell RNA sequencing data to the subsets described by Zhang et al.[223,224]. In addition, our studies focus on the immunoglobulin composition of the analyzed B cells in detail.

2.5.3 T and B cell receptor reconstruction and clonality inference from single cell RNA sequencing (TraCeR/ BraCeR)

TraCeR[225] and BraCeR[226] are specialized tools to analyze the T and B cell receptor sequences from single T and B cell RNA sequencing. Both tools wrap different other programs and customized scripts to allow this analysis. BraCeR uses a similar approach as TraCeR but has a specific feature to allow SHM that only occurs in immunoglobulins and not in T cell receptors. The proof-of-concept publication for BraCeR uses 2*75 bp paired-end sequencing data from 67 human plasma cells. A summary of the tool pipeline is shown in Figure 13.

First, the authors use Salmon[227] to map the reads to the reference genome and the ERCC spike-ins. Then they use tximport[228] to quantify the reads and calculate a gene expression matrix, where rows represent genes and columns represent single cells. Afterwards, they use a filter to exclude low-quality cells. This filter consists of the following criteria: the authors keep cells where they find <20% spike-in RNAs, <10% of the reads mapping to mitochondrial genes, >1200 mapped genes and a mapping rate >40%. This removes 37 cells from the analysis. For the remaining 30 cells a BCR is reconstructed. Therefore, a theoretical recombinome containing germline V, J and C gene segments is constructed.

In this recombinome, each entry contains a sequence spanning the V, J and CH1 segments of all isotypes or the KC or LC sequences. At the 5' end and in between these segments, there are

ambiguous N nucleotides inserted to create a reference that assigns sequence space to the leader region and the CDR3. Meanwhile, the reads for the single cells are trimmed to remove adapter sequences and filtered to remove low-quality reads using Trim Galore (http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/) and Cutadapt[229]. Then, the reads are aligned to this recombinome using Bowtie 2[230]. This tool is especially suited, since it allows alignments against ambiguous N nucleotides within a reference, and can also introduce gaps into both, the reference and read sequences. Thus, it can align against the variable junctional regions and allows mismatches originating from the immunoglobulin diversification processes.

The heavy and light chains of the individual cell are then assembled using the reads that aligned to the appropriate recombinome with the Trinity RNA-seq assembly software[231]. Contigs assembled by Trinity are then submitted to BLAST[232] for constant chain detection and to IgBLAST[233] for assignment of V, D and J segments. Contigs are classified as BCRs, if their gene segments derive from the correct locus and if their top assignment E-values are $< 5x10^{-3}$ or $< 5x10^{-4}$ for IgH V segments. E-values are an estimate for the likelihood that a given sequence matches the reference by chance. If the alignment scores are > 96% for IgH or > 99 % for IgK or IgL, possible allele assignments are added to a list for a given contig. Furthermore, the called contigs are checked, whether they encode a productive chain by excluding the existence of stop-codons and by assessing, whether the chains are in the correct reading frame. After this positive selection, negative selection processes take place.

Highly similar sequences are collapsed based on the following criteria: intersecting V and J gene assignments, equal CDR3 length and the Hamming distance of the CDR3s normalized by length are < 0.07. These highly similar sequences can occur for example, if the antibody is expressed in different isotypes, if the contigs are of different length or if they contain PCR or sequencing errors. Finally, the expression counts for the reconstructed BCRs and the transcriptome are estimated using Kallisto[234].

Kallisto is a pseudoalignment-based algorithm that is ideal for this task, as it is fast in constructing the indices for the BCR and transcriptome reference file and in performing expression quantification. When BraCeR assigns multiple BCR sequences to a specific locus, the most highly expressed recombinant for each locus and its assignment is chosen for further analysis. However, one has to be vigilant, not to analyze potential multiplets. Using BraCeR, B cell clonality and fate can be linked. While BraCeR is an *in silico* tool to study expressed immunoglobulin genes, studies of the immunoglobulin locus form the basis for expressed repertoire studies.

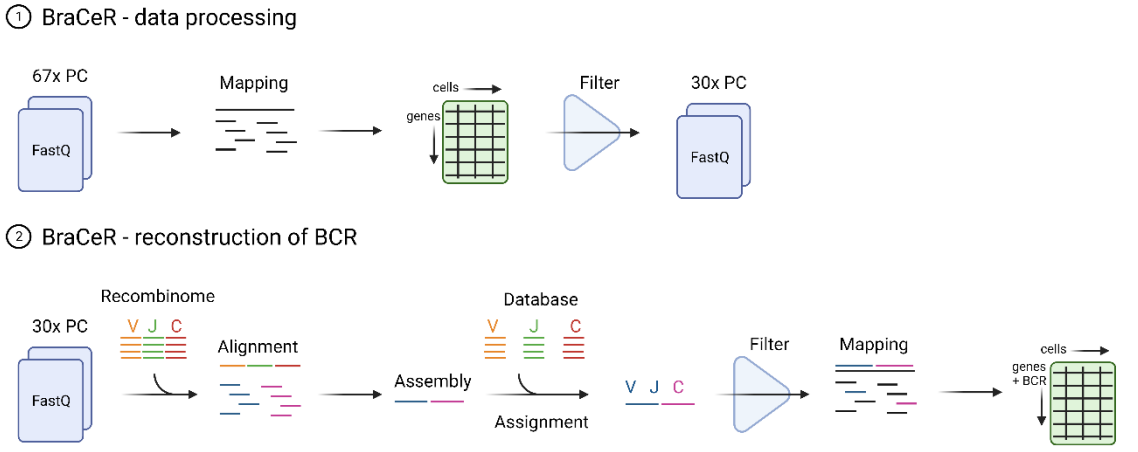


Figure 13 The scheme shows the BraCeR pipeline. Starting from fastq files from 67 plasma cells (PC), the tool takes 30 high quality cells to reconstruct the BCR.

2.5.4 Immunoglobulin loci sequencing efforts: long read and deep sequencing

Although the scientists running the Human Genome Project and Craig Venter and his company Celera Genomics published the first draft of the human genome in 2001 in Nature and Science, there are still highly variable and repetitive loci that are incompletely characterized on a population level. Although immunoglobulins are of great importance for infectious diseases, autoimmune diseases, allergies, cardiovascular disease and cancer, their germline configuration on a population level is still unknown. The repetitive structure of the immunoglobulin loci limits the use of short reads from 25 to 160 bp in length, which were used to assemble the human reference database listing the most common variants within the 1000 Genomes Project. On the other hand, creation of long reads spanning the entire immunoglobulin loci is so far limited to a small number of individuals. In fact, the IGH locus has only been completely sequenced twice using yeast artificial chromosome libraries and cosmids[235] and bacterial artificial chromosome libraries[132]. However, recently there are studies emerging to characterize human haplotype diversity using long-read PacBio sequencing[236,237]. Indeed, there is only one study that identified 6 single nucleotide polymorphisms (SNPs) in the IGHV locus, which correlate with Kawasaki syndrome susceptibility in a Han Chinese population[238]. With new Next-generation-sequencing- (NGS-) based methods arising, targeted sequencing of immunoglobulin libraries on the Illumina MiSeq platform allowing 2*300 bp paired-end reads may have the potential to map the variants in these highly repetitive loci in more individuals[239].

2.5.5 IgDiscover: germline inference and repertoire analysis

IgDiscover[240] is a versatile tool that allows inference of immunoglobulin germline alleles as well as analyses of immunoglobulin repertoires. To infer individualized germline alleles, IgDiscover assigns 2*300 bp reads from IgM libraries with about one million sequences per library to the closest sequences in V, D and J segment databases. Based on the distribution of sequences of percentage mismatches to a single database sequence, windowed and linkage clustering is applied to create a consensus-driven candidate germline gene database for each gene segment. This database is then filtered for example based on the number of barcodes and unique CDRs. The created database is used to replace the initial reference database and the process is iterated. For human samples with a reasonably large starting database, one iteration is sufficient for the analysis. The inferred individualized germline alleles can be used to determine features of the expressed repertoire such as clonality or gene usage. A scheme of the IgDiscover inference method is shown in Figure 14.

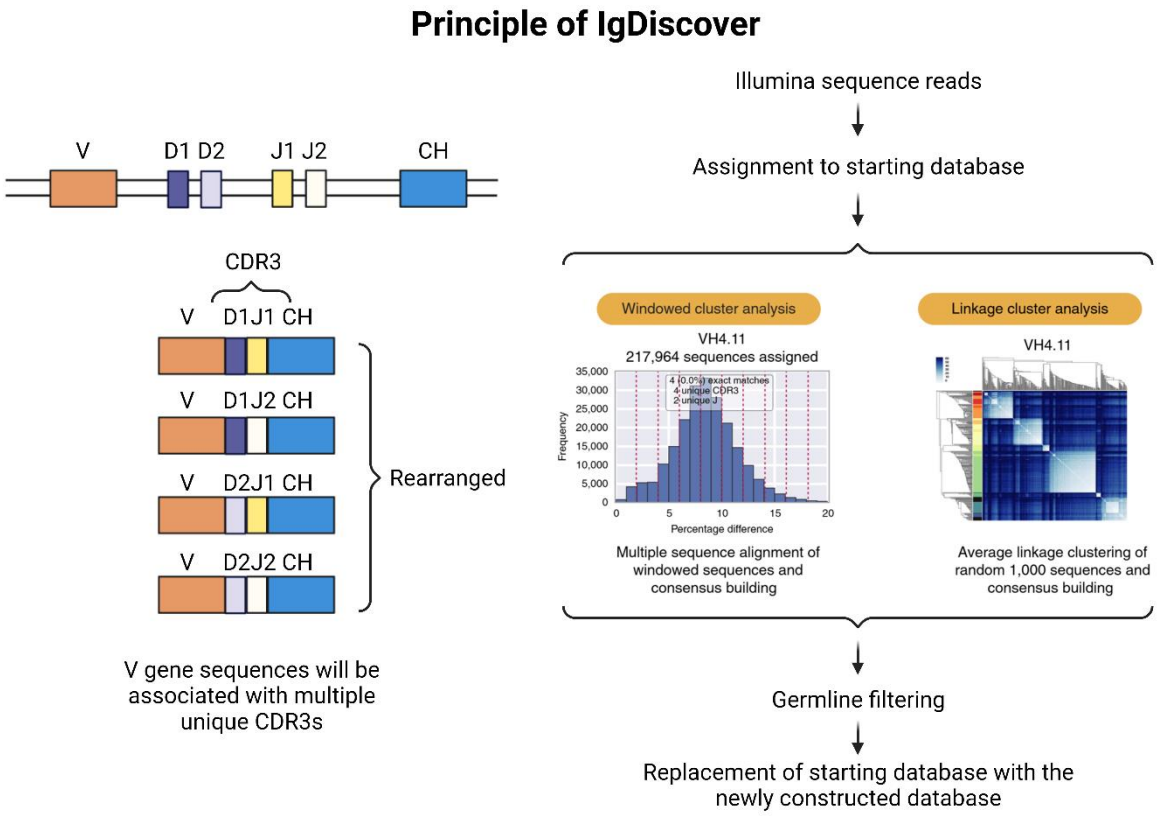


Figure 14 IgDiscover can infer germline genes from immunoglobulin M libraries.

Substantial immunoglobulin germline variation is prevalent by the number of identified novel V alleles that have been inferred by a number of repertoire studies[241–247]. Novel alleles identified through computational inference from expressed IgM libraries must be validated genomically to be added to the IMGT/GENE-DB, which currently serves as a reference database. So far, IMGT covers mainly individuals with European ancestry, a limitation that has been noted[248]. The Inferred Allele Review sub-Committee (IARC) has currently assessed and recommended 16 inferred heavy chain and 2 light chain immunoglobulin V alleles, which have

been approved and added to the IMGT/GENE-DB, because of sufficient evidence from multiple studies. Despite a large number of known germline V alleles, more are still being discovered.

The Chromium Single Cell Immune Profiling, which is offered by 10x Genomics, represents a commercialized targeted sequencing approach that aims to identify the paired heavy and light chain B cell receptor profile of up to 80,000 cells. 10x Genomics provides a public dataset generated from 17,000 $CD19^+$ B cells from a healthy donor (official 10x Genomics support). In this dataset, they identify 8,951 cells expressing at least one of the immunoglobulin chains. The analysis does not cover allelic variation nor detect more than 38 functional IGHV genes after basic filtering, while we identify all 50 functional IGHV genes per individual with our targeted bulk sequencing approach using IgDiscover. Thus, the 10x Genomics method is insufficient for in-depth studies of the immunoglobulin repertoire in great detail.

To our knowledge, one of the first published studies using the 10x Genomics approach for human B cells is, where the authors investigate the uniform B cell receptor repertoire of 1,300 plasma cells from a multiple myeloma patient and compare it to a more polyclonal bone marrow plasma cell pool from an individual with osteoarthritis that underwent hip replacement surgery[249]. Although paired immunoglobulin heavy and light chain information is retained, this study does not even identify half of the IGHV genes that we find using bulk repertoire sequencing. In a more recent study[250], the authors identify only 36 IGHV genes without the allelic resolution of 76,276 IgG^+ B cells from repertoires of three donors. In an optimal scenario, 10x Genomics could provide the paired sequences of about 40,000 $CD19^+$ B cells mapped to a public database, while bulk repertoire sequencing identifies up to 10^6 unpaired heavy and light chain sequences mapped to individualized germline databases, thus providing a more detailed picture.

In Rheumatoid Arthritis, repertoire studies are so far limited to the study of heavy chain CDR3 sequences, which are derived from comparatively small libraries with a minimum of 10,000 heavy chain sequences per individual and which are amplified with a limited number of IGHV primers[251]. In our studies, we use more than 15 IGHV 5'MTPX primers and find roughly 5-10% of unique CDR3 sequences per library with 10^6 sequencing reads. Thus, previous studies are insufficient to describe the immunoglobulin heavy chain repertoire in detail. Furthermore, the previous studies did not identify personal germline allele contents, but assigned the sequences to the IMGT database, which likely leads to several mis-assignments.

2.5.6 Polymorphisms or allelic variation in the immunoglobulin loci and disease associations

In early studies, immunoglobulin polymorphism analyses were performed by restriction fragment length polymorphism (RFLP). In these studies, genomic regions are amplified by targeted PCR and then digested with a restriction enzyme that is sensitive to only one allelic variant of the single nucleotide polymorphism. Then, one can expect different band sizes depending on the present variant in an electrophoretic gel. The study by Vencovsk et al.

suggested that a polymorphism in IGHV1-69 was linked to susceptibility to RA in a Czech cohort that lacked the HLA-DRB1 SE. However, this association was not found in a British Caucasian RA cohort[252].

Several other studies showed a possible role for the IGHV locus, especially the IGHV2 locus, in multiple sclerosis[253–255]. Polymorphisms in or absence of the complete IGHV3-30/IGHV4-31 region (hv3005) were found to be associated with RA[256] and SLE in Korean[257] and Caucasian[258] populations. This deletion was also enriched in chronic idiopathic thrombocytopenic purpura in Caucasians[259]. Similarly, polymorphisms in the IGHV2, IGHV4 and IGHV5 were found to be associated with type I diabetes in the Finnish population[260]. However, these studies are hard to interpret with the current knowledge of the immunoglobulin loci.

As previously mentioned, six SNPs in the IGHV1-69/IGHV2-70 gene region were found to mediate disease susceptibility to Kawasaki's disease in a genome-wide association study (GWAS) in a Han Chinese population[238]. In a GWAS from the indigenous populations of the South Pacific a haplotype of nonsynonymous variants in the IGHV4-61 gene segment corresponding to the *02 allele was found to be associated with increased risk for rheumatic heart disease[261].

A polymorphism, also leading to the absence of immunoglobulin variable gene, was reported for the recombination signal site (RSS) of the IGK2-29D gene in the κ locus for the Native American population, which was associated with increased susceptibility for *Heamophilus influenzae* type b infections in this population[262,263].

The importance of the immunoglobulin variable gene germline configureation has also been shown in antibody responses to different infections such as influenza, HIV, SARS-CoV-2, Ebola, Zika and malaria.

The IGHV1-69 alleles either contain a phenylalanine (F) or a leucine (L) at position 54, in the CDR2 region of the IGHV gene. Influenza neutralizing antibodies that target the conserved HA stem region have been shown to prefer using the F variant. Studies from the 1000 Genome Project investigated the IGHV1-69 polymorphism at the population level and found that the F/F genotype was high in Africa and lower in South Asian cases indicating the possibility of a population-specific protective response towards the influenza virus[264,265]. Another important IGHV germline gene for a protective influenza response is the IGHV3-30 gene, that provides a good basis for influenza-specific antibodies[266,267].

The VRC01 class of antibodies that have been isolated from individuals infected with HIV-1 describe a class of antibodies that use the IGHV1-2*02 allele and are broadly neutralizing[268]. The essential amino acid positions are W50, N58 in the CDR2 and R71 in the FRW3[269], which are present in the IGHV1-2*02 and IGHV1-2*04 alleles. Individuals lacking the germline alleles with the necessary W50, N58, R71 motif are unable to generate VRC01 class broadly neutralizing antibodies to HIV-1.

Antibodies that are SARS-CoV-2 neutralizing[270] often use the IGHV3-53 and IGHV3-66 alleles and are minimally somatically hypermutated. The two motifs that are important for binding the receptor binding domain (RBD) of the virus are an asparagine-tyrosine motif in position 32 and 33 in the CDR1 and a serine-glycine-glycine-serine motif in positions 53 to 56 in the CDR2. IGHV3-53 alleles with these motifs are found in all the population groups[271,272].

Neutralizing antibodies that are protective for Ebola virus infection have been isolated from Ebola virus survivors. Convergent antibody evolution was seen across multiple donors, particularly among IGHV3-13 neutralizing antibodies specific for the glycoprotein 1 core[273].

Several antibodies neutralizing Zika virus have been isolated and cloned from infected individuals. The majority of antibodies used a IGHV3-23 heavy and IGKV1-5 light chain gene. For the heavy chain contact with the virus, germline residue Y58 was found to be important[274].

In malaria, antibody responses are mounted towards an immunodominant NANP repeat region on a surface protein expressed by the *Plasmodium falciparum* parasite. These antibodies are mostly encoded by the germline IGHV3-33 heavy chain and IGKV1-5 light chain genes[275]. Repertoire analysis showed that amino acid W52 in the CDR2 is important for the contact of the antibody with the parasite[276].

These studies highlight the importance of germline gene-encoded variation in the immunoglobulin variable gene locus for autoimmune disease susceptibility and responses to infectious agents. Allelic variation that affect some of these polymorphisms may explain differences in B cell responses in these contexts between individuals or population groups.

3 RESEARCH AIMS

The specific aims for the individual studies were:

I. To investigate B cell phenotype, environment and clonality in synovial tissue of early RA.

II. To determine advantages and limitations of NGS-based immunoglobulin library preparation methods.

III. To study baseline characteristics of immunoglobulin heavy chain segments using NGS.

4 METHODOLOGICAL CONSIDERATIONS

4.1 TRANSCRIPTOMICS AS A PROXY OF FUNCTIONALITY

Single cell transcriptomics have become an important tool to globally dissect cellular phenotypes in an unbiased high throughput fashion. Therefore, transcriptomics constitute a feasible alternative approach compared to proteomics to dissect the functional states of individual cells. As the function of a cell is primarily attributed to its proteome, the transcriptome remains a proxy. However, single-cell proteomics are still subjected to methodological challenges and are limited to the most abundant proteins. Although single cell RNA sequencing is also biased to the more abundant transcripts, it remains more sensitive than single cell proteomics detecting up to 5000 genes per cell. An advantage of using flow cytometry assisted cell sorting in single cell transcriptomics is, that a target population can be enriched and the data from index sorting can be used to assess the relationship between the transcriptional and the protein profile.

Notably, RNA presence does not always mean protein presence. There are many regulatory networks that determine whether RNA is translated into protein. This can have advantages and disadvantages when trying to study the functionality of a cell. Advantages include the detection of many regulatory RNAs that also have a polyA tail. This includes for example long non-coding RNAs (lncRNAs). A disadvantage is that amounts of RNA may not be correlating to the amount of corresponding protein and hence that the functional cell state is not accurately depicted.

4.2 GENERATION OF SYNOVIAL B CELL SUSPENSIONS FOR DOWNSTREAM APPLICATIONS

Synovial tissue is a collagen-rich fibrous tissue that regulates the synovial fluid composition. Consistent live cell recovery after tissue dissociation is important to downstream applications such as flow cytometry and transcriptomic analysis. When the number of live cells was quantified using the ratio of cells per beads and cells per gram of biopsy, enzymatic and mechanical dissociation yielded considerably higher amounts of live cells than mechanic dissociation alone[277]. Another important consideration is that utilizing proteolytic enzymes can lead to cleaving of markers that are thus masked for downstream analysis. Notably, CD27 expression on B cells is reduced post-enzymatic digestion, but is re-established after 6h[278]. In our studies, we employed both enzymatic and mechanic dissociation, however, we did not test our protocol to specifically look at $CD19^+$ B cell recovery. Finding good overlap between the naïve and the memory B cell subset in the flow cytometry and the associated RNA sequencing data indicates that there was sufficient CD27 expression to characterize these subsets also on the protein level. As one of our reviewers of study I pointed out, the proportion of single cell RNA sequencing assigned memory cells in the IgD^+CD27^- gate from the flow cytometric analysis is about 16%. These may either be cells that already have transcript for the memory cell phenotype, but still appear IgD^+CD27^- as transcript expression comes before protein expression, or a mixture of cells that have lost CD27 expression due to masking. Notably, we

also stained for the plasma cell marker CD138, which gave a weak signal that did not correlate with the clear picture from the *SDC1* expression from the single cell RNA sequencing. Since CD138 is important for cell-matrix interactions, we hypothesize that upon matrix removal, CD138 is downregulated at the protein level.

4.3 CRYOPRESERVATION OF SYNOVIAL TISSUE BIOPSIES AND SINGLE CELL SUSPENSIONS

It has been shown that correct cryopreservation and thawing of synovial biopsies is essential for preserving cells. A previous study[279] highlighted that the total number of viable cells in fresh versus frozen samples were not significantly different. Neither were the trends for $CD19^+$ B cell numbers. However, among the three paired data points, there was a trend towards fewer $CD19^+$ B cells among dissociated cells that underwent the cryopreservation process. A different study[277] also showed similar percentages of live $CD45^+$ cells irrespectively of whether the tissue was dissociated fresh, cryopreserved and then dissociated or dissociated and then cryopreserved. Unfortunately, this study did not compare $CD19^+$ B cell recovery in detail. In our studies, we used fresh samples.

4.4 SINGLE CELL LIBRARY PREPARATION

To improve the quantitative nature of single cell RNA sequencing, unique molecular identifiers (UMIs) can be introduced upon reverse transcription. However, a 3' end coverage bias is inherent to most single cell RNA sequencing protocols. Barcoding and pooling prior to sequencing limits sample loss, reagents and work costs.

Additionally, to estimate technical variation, External RNA Controls Consortium (ERCC) mRNA is commonly spiked into the cell lysate. This allows to observe variances of counts between individual cells[280]. Moreover, there is variance originating from the biology. mRNA amounts can be influenced by cell-type, transcription kinetics and cell cycle for example. Considering all potential sources of noise, filtering criteria and analysis methods such as clustering and principal component analysis (PCA) must be chosen carefully.

4.5 B CELL REACTIVITY TESTING

B cell reactivity is classically evaluated by ELISpot, ELISA, immunofluorescence assay or surface plasmon resonance (SPR). The latter methods are all dependent on the generation of B cell receptor-derived monoclonal antibodies. However, these assays may not capture the full biophysical features of interactions between membrane-associated B cell receptors on naïve or memory B cells. Thus, some membrane-associated B cell receptor reactivity may be below the detection threshold of the previously mentioned methods. However, monoclonal antibodies may be adequate to determine the reactivity of plasma cell-derived secreted antibodies as in our studies. In this case, one limiting factor may be the influence of the immunoglobulin heavy chain isotype, as most monoclonal antibodies are expressed with an IgG1 constant isotype. Alternative approaches to study membrane-associated B cell receptors include flow cytometry

assisted cell sorting-based detection of antigen binding and cell-based calcium assays measuring antigen-induced calcium release, which indicate BCR-induced cell activation.

4.6 STATISTICAL CONSIDERATIONS

In my studies, I have drawn samples of study populations. There are statistical methods to decipher, whether the observed differences in the sample groups may be attributable to real difference or to random variation. Statistical tests and measures help to estimate how likely chance and random variation could have had an influence on the results. Scientific statistical analysis tells us the probability that the results in the study reflect the situation in the population. After identification of the appropriate statistical methods (regression analysis, odds ratios, comparisons of sensitivity and specificity, t tests, χ^2 analyses, analysis of variance), the sample size required can be predicted from power studies.

In my first study, sample size reduction was problematic with regards to be able to derive conclusions for ACPA- or ACPA+ disease. Although, we originally sampled a total of eight early onset patients, we only had high quality single cell RNA sequencing data of three patients. Thus, the study was underpowered to describe synovial B cell differences in ACPA- versus ACPA+ disease beyond an observational characterization of the B cell compartment in synovial tissue of early RA patients. Notably, there is a study that compares synovial tissue single cell samples of a total of twenty RA/OA patients with long-standing active disease that was powered to find differences between ACPA- and ACPA+ disease[281].

In my third study, sample size was an equally important factor to consider. The immunoglobulin locus is highly variable and 30 cases are not enough for association studies, for example to investigate if there are differences in germline allele content between ACPA+ RA and ACPA- RA. Before embarking on these studies, power calculations should be performed. These can be informed by the risk conferred by the SE alleles. There are current efforts in the laboratory to establish a method for personalized VDJ typing with higher throughput that will be applied to larger sample sets such as the EIRA study group in the future.

Hence, we illustrate the importance of power calculations, also when considering ethical requirements for a study. An underpowered study does not enable to derive conclusions from the studied sample, while a too large study sample raises ethical questions about the waste intensive approach and in case of human or animal studies, subjecting more individuals than necessary to the inconveniences.

4.7 ETHICAL CONSIDERATIONS

RA is a chronic, painful and disabling disease, where the etiology is incompletely understood. Knowledge on the pathogenesis of the disease may enable scientists to develop targeted therapeutics leading to clinical improvement for patients. However, there are no guarantees.

In these studies, we used blood as well as joint biopsy samples from RA patients and healthy donor blood samples. The use of these samples implies several important ethical

considerations. First, individuals are subjects that have the inherent right to autonomously decide. Thus, sample donations are provided under informed consent that can be withdrawn without reason at any time point.

Secondly, sensitive information such as identity, clinical assessments and genetics must be handled with care. This includes the use of appropriate identifiers whenever possible and the limited distribution of sensitive details. In our studies, I introduced additional pseudonymized identifiers upon sample reception, so that fellow study contributors cannot back trace any sensitive details. The identification key is securely stored.

The third argument deals with the type of samples used in these studies. 10 ml blood samples are easily accessible and do not cause additional physical or mental harm to individuals that may already suffer from a chronic disease that needs clinical attention. In contrast, joint biopsies are not taken routinely from RA patients. This procedure includes local anesthesia, takes about 2 hours and is performed by a specialized rheumatologist. As such, this procedure entails secondary post-operative effects and bares the risk of medical side effects. A critical issue to consider should be the extent of the risk of worsening the disease by inducing additional inflammation through the procedure. Apart from the physical effects, the mental effects of this procedure should be considered. Here, we used joint biopsies that are taken at the time point of diagnosis of a chronic disease. Thus, the procedure may put an additional mental burden to the patients at this stage.

Lastly, individuals that donate samples in these studies may not directly benefit from the results. It may be that the results are only implemented in large research efforts that ultimately lead to clinical improvements in the far future. However, there are individuals that actively want to contribute to this type of research and several projects have patient research partners.

A different set of ethical considerations concerns the integrity and nature of the science that is pursued in all research areas including these studies. These considerations relate to conflicts of interest or presumably unintentional misconduct that can affect research in different forms. This includes for example authorships in publications and faulty scientific results due to technical artefacts, flaws in experimental set-ups and biased views leading to misconceptions in the entire scientific community.

Important measures to address all ethical issues are the ethical review of scientific projects, transparency and good routines in research processes and peer review of scientific literature, as well as critical thinking towards one’s own research and the research from others.

5 RESULTS

5.1 PHENOTYPES OF SYNOVIAL B CELLS IN EARLY RA

B cells play a major role in established RA. Here, we used biopsy samples (n=8) from early RA patients to study the B cell composition and context at the time of diagnosis.

Using index sorting in flow cytometry, we found 20-45% of $IgD^{+}CD27^{-}$ naïve B cells, 24-63% $IgD^{-}CD27^{+}$ classical memory B cells, 1-4% $CD27^{++}$ B cells, 2-4% $IgD^{+}CD27^{+}$ B cells having an unswitched memory phenotype and 8-29% of B cells having an $IgD^{-}CD27^{-}$ phenotype. These proportions were irrespective of the ACPA status of the patients. In parallel, we found B cell rich infiltrates in two biopsies and more dispersed B cells in two other biopsies using CD19 and CD20 for immunohistochemistry, again, irrespective of the ACPA status.

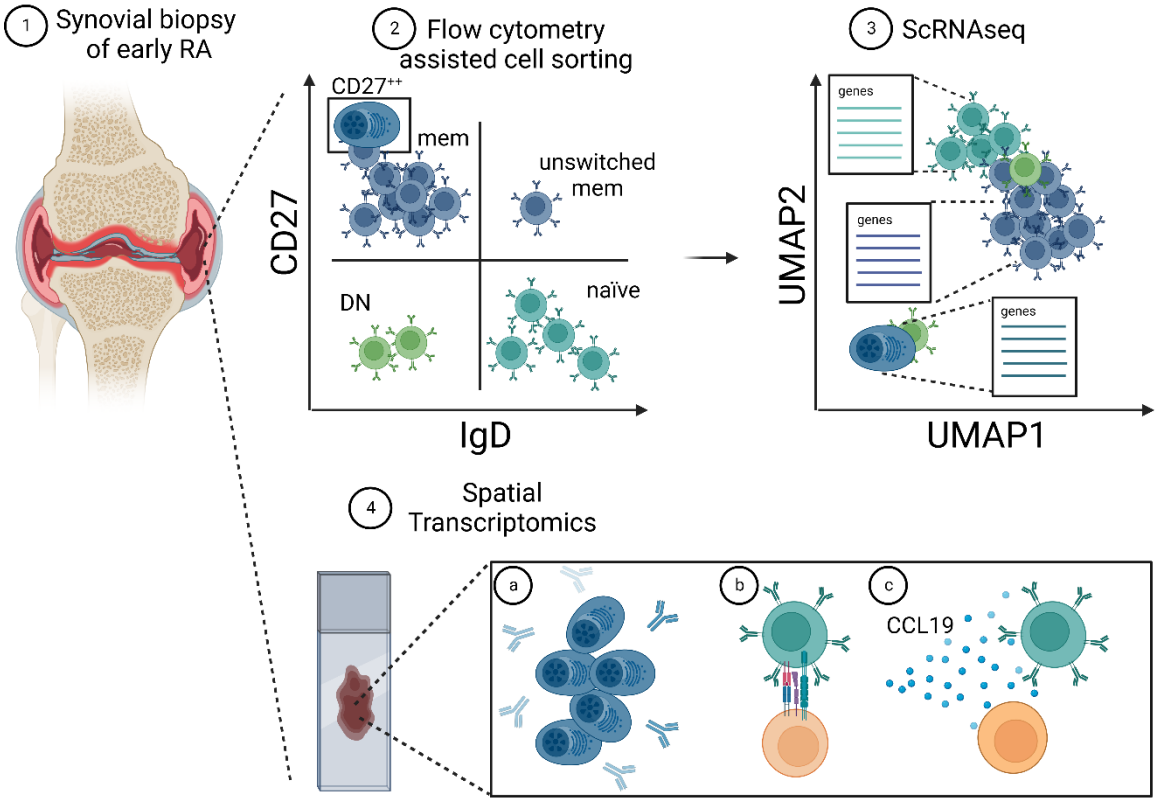


Figure 15 Synovial biopsies were processed for flow cytometry assisted cell sorting, single-cell transcriptomics (scRNAseq) and spatial transcriptomics.

In the single cell RNA sequencing data, we found 113 plasma cells (6%), 395 naïve B cells (21%) and 1388 memory B cells (73%) from one ACPA+ and two ACPA- patients. Reassuringly, these phenotypes translated back to the index labels from the flow cytometry sort. Notably, we could not find the signature of autoimmune-associated B cells in the biopsy material. Plasma cells were characterized by *XBP-1*, *SDC1* (encoding CD138), *SLAMF7* and *PRDM1* (encoding Blimp-1) expression. In contrast, the memory B cell cluster was defined by *ITGAM* (encoding CD11b) and *GPR183* (encoding EBI2) expression. The naïve B cell cluster had high expression of *IGHD* and *IL4R*, which is characteristic for naïve resting B cells.

We also subjected a set of biopsies to spatial transcriptomics analyses. Thereby, we found three clusters that had immune cell signatures. One cluster was in proximity of the infiltrate regions and B cell rich areas and had a strong plasma cell signature characterized by *JCHAIN*, IGH, IGL, *MZB1*, *SDC1* and *XBP1* expression. The second cluster was characterized by high expression of HLA class II genes and *CD74* indicating antigen presentation mechanisms. The last cluster had a weaker immune signature, which could mean that this cluster was not primarily made up by immune cells or that the present immune cells were less activated. Still, there was apparent expression of *CCL19* known to attract B and T cells. A summary of these results can be seen in Figure 15.

5.2 DIFFERENTIATION PROCESS AND NICHE OF SYNOVIAL B CELLS IN RA

Furthermore, we investigated B-T cell interaction networks implicating the antigen-presentation capacity of B cells. We found high expression of HLA-DR genes as well as expression of costimulatory receptors. Besides, the spatial transcriptomic approach revealed four B cell signatures and one T cell signature upon factor analysis. Submitting the driver genes for pathway analysis using NCATS BioPlanet 2019 resulted in identification of the BCR signaling pathway as well as the PD-1 signaling pathway as the top hit for the B-T cell interaction pathways. This finding supports the notion of B-T cell crosstalk with respect to the identification of T peripheral helper cells in RA.

Next, we studied the profile and spatial context of differentiated B cells, i.e. memory B cells and plasma cells. Memory B cells were marked by expression of *CCR6, CR2* (encoding CD21), *FCER2* (encoding CD23), *ITGB2* (encoding CD18), *IRF8*, *PAX5*, *SPIB*, *SKI*, *TNFRSF13B* (encoding TACI), *CD27* and *CXCR3*. Spatially, memory B cells were co-localized with other B cells and T cells increasing the likelihood of present B cell expansion despite the lack of clear proliferation signals such as Ki-67 and B-T cell crosstalk. Plasma cells were characterized by expression of the top-twenty differentially expressed genes as described before and *ICAM2*, *ITGA6* encoding CD49f, *CCR2*, *CCR10*, *CXCR4*, *TNFRSF17* encoding APRIL receptor BCMA, *IL6R*, *PECAM1* (encoding CD31), *SELPG* (encoding CD162) and CD98 encoded by *SLC3A2* and *SLC7A5*. Spatially, plasma cell location correlation estimates were somewhat lower between plasma cells and B and T cells compared to memory B cells and B and T cells. That means that plasma cells occurred near the lymphocyte infiltrates. Correlations with location of fibroblasts and monocytes were inverse. Since plasma cells were expressing CXCR4, we also investigated the overlap with CXCL12 in the tissue and found a high overlap, which indicates that the CXCL12-CXCR4 axis is probably active already at time of diagnosis.

A summary of these results is shown in Figure 16.

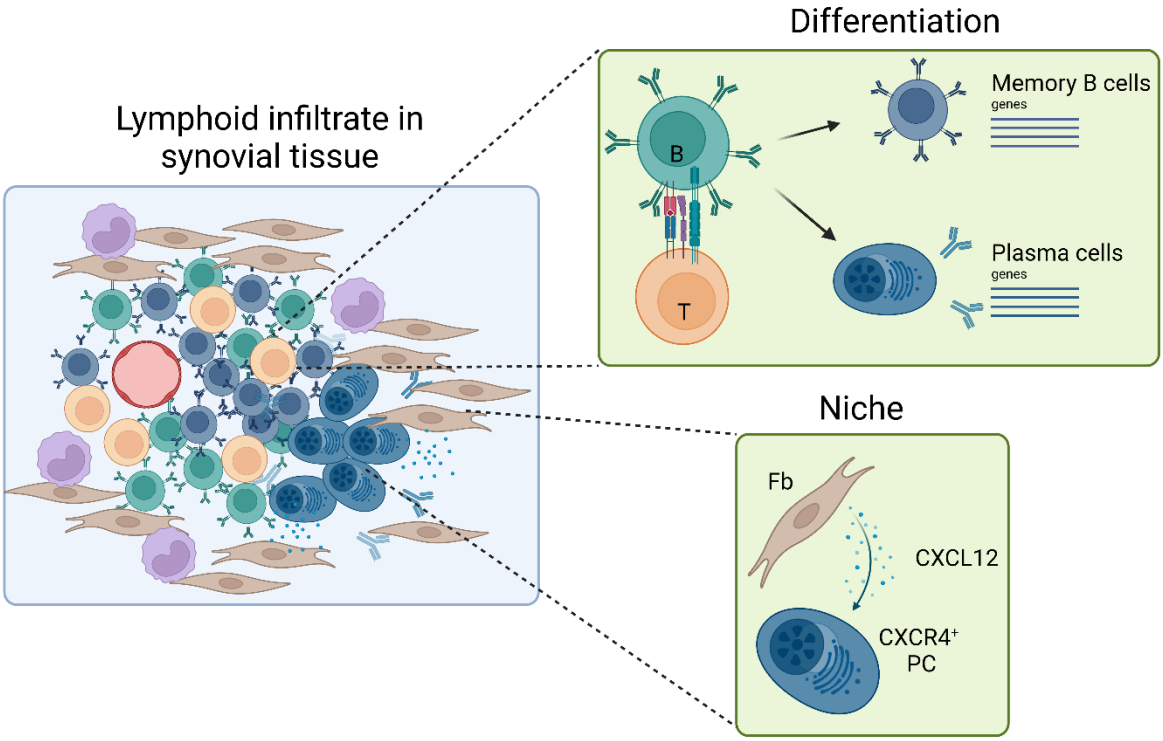


Figure 16 Inside the lymphoid infiltrate in the synovial tissue, there are differentiation processes as well as a niche structure.

5.3 BCR CLONALITY OF SYNOVIAL B CELLS IN EARLY RA

From the single cell data, we also studied the clonality and reactivity of plasma cells. For 9 out of 15 monoclonal antibodies derived from plasma cells, we found sibling cells with an identical variable and constant chain pair among the memory cell compartment. The five clonotypes from patient A2 covered 9% of the memory and plasma cell members, the two clonotypes from patient A3 covered 31% of the memory and plasma cell members and the seven clonotypes from patient A4 covered 10% of the memory and plasma cell members. Thus, the synovial B cells from patient A2 and patient A4 appeared more polyclonal, than from patient A3. Importantly, one expanded clone from the ACPA+ A2 patient was CCP2 reactive and not found to be polyreactive or reactive against other tested post-translational modifications. To further study the citrulline reactivity, we tested the reactivity to different citrullinated antigens. Reactivity to citrullinated mutated vimentin was found down to a concentration of 2 ng/ml. We also found vimentin expressed in the patient specific tissue section. The ACPA was characterized by SHM induced glycosylation sites as well as high SHM rates.

A summary of the findings is presented in Figure 17.

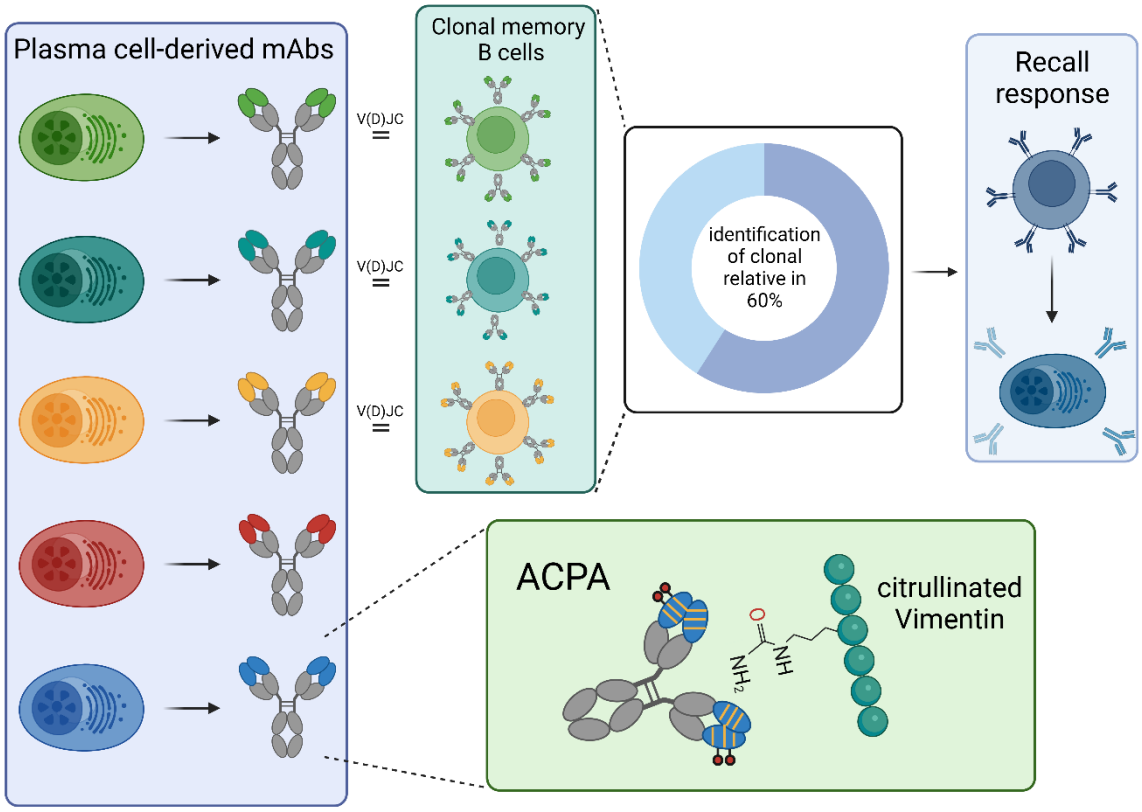


Figure 17 From the expressed plasma-cell derived monoclonal antibodies (mAbs), we identified clonal memory B cells in 60% of the cases. This indicates a recall response, where memory B cells differentiate into plasma cells. Furthermore, we found an ACPA towards modified citrullinated vimentin that has typical ACPA characteristics.

To better characterize the implicated phylogenetic relationships within clonotypes, we broadened our analysis to BCR repertoire sequencing that we used foremost to discover immunoglobulin germline information. These data may inform us about genetic predispositions for autoimmunity and disease, when using the proper population controls.

5.4 NGS LIBRARY PREPARATION METHODS

Immunoglobulin germline inference and repertoire sequencing analysis require robust NGS library preparation methods. The most used methods to produce NGS libraries include 5’ rapid amplification of cDNA ends (5’RACE) and 5’ multiplex (5’MTPX) protocols.

We improved the 5’RACE protocol by introducing Illumina’s READ1 sequencing primer as the universal amplification primer to enable template switching. This reduced the amplicon length by 20-25nt. Furthermore, we introduced a semi-nested PCR protocol to increase yield of the template switching reaction. To this end, we designed inner and outer constant region primers to use with the 5’READ1 universal primer in two consecutive PCR reactions with 20 and 10 cycles. The inner region primer additionally introduced the READ2 sequence to the 3’ end. Thus, the library product contains the V gene leader sequence and the 5’UTR, enabling design of full-length V(D)J primers for 5’MTPX libraries.

For our MTPX library preparation protocol, we designed leader region primers using the sequence information obtained from the 5’RACE libraries and available genomic references (GRCh37 and GRCh38). Thus, the product contains full-length V(D)J amplicons. Reverse

transcription introduces a UMI and READ2 sequences in the product. This reaction is followed by the amplification PCR, which introduces READ1 at the 5'end of the multiplex primer sequences followed by the indexing PCR to add the adapters for the Illumina MiSeq platform. Thus, the 5'MTPX protocol requires fewer steps than the 5'RACE protocol and yields shorter sequences.

A summary of the methods is shown in Figure 18.

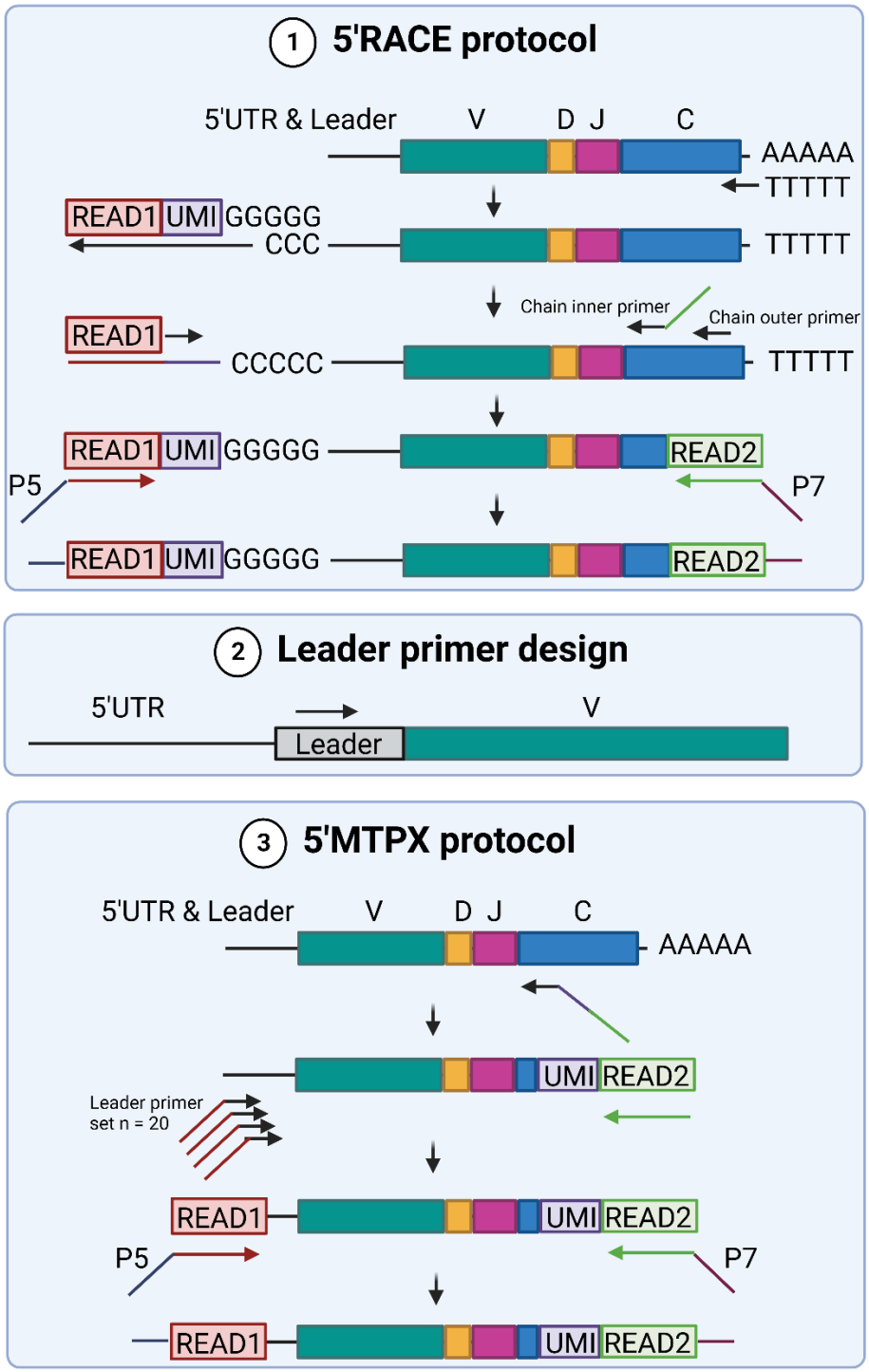


Figure 18 The scheme shows protocols for 5'RACE and 5'MTPX libraries.

5.5 AMPLICON LENGTH LIMITATION EFFECT ON HC LIBRARY QUALITY

The Illumina MiSeq 2x300 bp V3 kit has a sequencing capability of 600 nt, which is theoretically sufficient to cover the amplicon libraries of both preparation protocols. In practice, there is the need for a sequence overlap to merge Read1 and Read2 and the sequence quality decreases with a significant effect from 275 nt in the Read1 direction and 225 nt in the Read2 direction. So, sequences above 500 nt in length are associated with higher levels of instrument-based sequencing error.

The SMART 5'RACE library generated amplicons of 580 – 600 nt in length, while our improved 5'RACE method generated amplicons between 550 – 560 nt in length. In contrast, the 5'MTPX approach generated amplicons below 500 nt. Longer sequencing amplicons resulted less often in merged reads in the PEAR merger within the IgDiscover analysis. Amplicon length mainly depends on the length of the CDR3 and the length of the 5'UTR in the 5'RACE protocol libraries. Longer 5'UTRs in the IGHV3 family were lost in 5'RACE versus 5'MTPX due to reduced merging of Read1 and Read2.

The quality of the 5'RACE and 5'MTPX libraries could also be assessed by investigating the number of sequences that are identical to the inferred germline genes in each library preparation protocol. Sequences that are not identical to the inferred germline genes are either somatically mutated or arise from instrument error. The amount of SHM detected in both methods should be equal, while the amount of sequencing error is higher in 5'RACE due to length limitations as described earlier. In the IgM libraries from our six test subjects, we found a mean difference of 23 ± 3 % between the lower numbers of identical sequences in the 5'RACE library compared to the higher numbers of identical sequences in the 5'MTPX library. In line with these observations, the Phred quality score was lower in the 5'RACE library in the end of the V segment compared to the 5'MTPX library due to the smaller sequence overlap in the merging procedure. The Phred quality score estimates the base calling error probabilities in a 10-fold logarithmic fashion. For example, if Phred assigns a quality score of 30 to a base, this is equivalent to the probability of an incorrect base call of 1 per 1000 base calls. A lower quality score of 20 equals to an incorrect base call once in every 100 base calls. Since one objective of the library analysis is to infer germline alleles, we could also directly compare the output as per number of alleles found in the libraries from both protocols and in either one. We found that most alleles were inferred in the libraries from both protocols, but on average, we identified 11 more alleles by the 5'MTPX approach compared to the 5'RACE approach.

The length limitation by the 5'RACE approach is visualized in Figure 19.

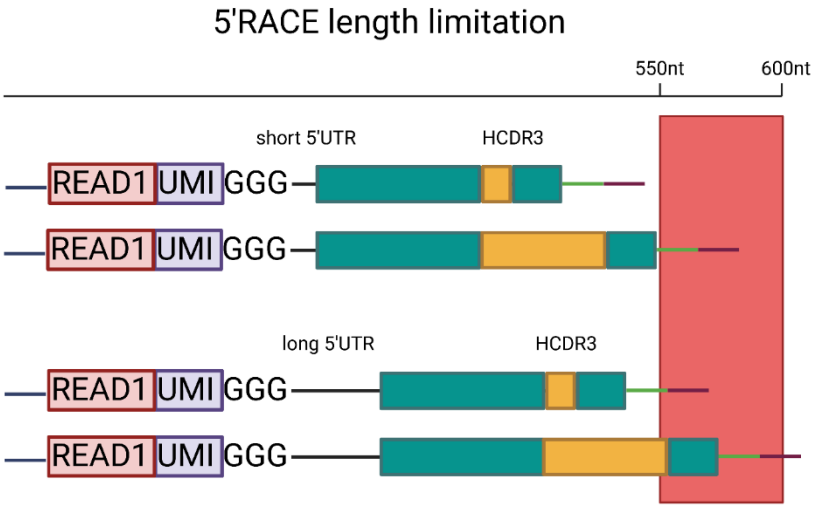


Figure 19 The RACE library preparation method creates amplicons that are too long for the practical Illumina MiSeq sequencing length capability. Sequences with long 5'UTRs and/or long HCDR3 risk of being low quality and underrepresented in the repertoire.

5.6 EFFECT OF STARTING MATERIAL ON LIBRARY QUALITY

Next, we also tested the amount of mRNA from peripheral blood mononuclear cells (PBMC) necessary to produce diverse 5'MTPX libraries. We found that the number of unique CDR3s saturates, if using more than 200 ng of RNA. Similarly, we found more unmutated D segments per V allele suggesting increased diversity and quality of libraries generated with 200 ng of RNA.

5.7 IMMUNOGLOBULIN LIGHT CHAIN LIBRARY QUALITY

Producing light chain libraries provides advantages and disadvantages compared to the heavy chain library production. Light chain amplicons are shorter in length and thus remove the length limitation associated with heavy chain libraries. However, the variability of the CDR3 is lower, which results in lower amounts of identified unique recombination events during the germline inference process. Another metric to use to identify candidate germline V sequences is to count the number of J genes associated with a given V gene allele. In this analysis, we found that for IgL repertoires, we could use a single constant region primer to target all the constant regions and that J1, J2 and J3 together with C1, C2 and C3 were the most frequently used λ genes.

A summary of the light chain generation approach is shown in Figure 20.

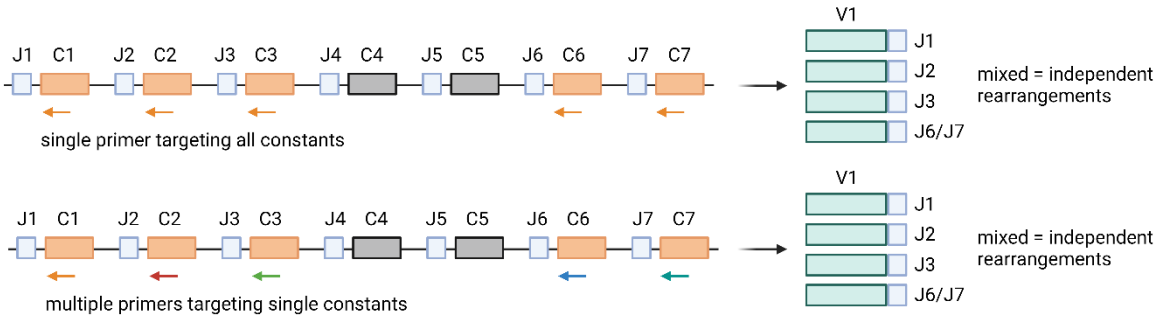


Figure 20 Using a single primer for the lambda chain constant regions is as good as using multiple primers targeting each constant region for detecting independent rearrangements.

5.8 INFERENCE OF INDIVIDUALIZED IMMUNOGLOBULIN ALLELES

Finally, we used our 5'MTPX primer sets to create an individualized germline database for one individual. We found three novel IGHV alleles, one novel IGKV allele and six novel IGLV alleles that were not described in the IMGT database. The novel alleles were genomically validated and found to be expressed at comparable levels to those of known alleles of the same gene.

Using this information for library preparation, we went on and applied the 5'MTPX method to study baseline characteristics of IgH genes from cases of a local RA study group.

5.9 BASELINE CHARACTERIZATION OF IGH GENES FROM REPERTOIRE STUDIES

Here, we inferred the immunoglobulin heavy chain germline genes from 30 female cases from the EIRA study. Inclusion criterion was that the cases had available whole blood RNA for repertoire library preparation. Individuals of possible non-European origin assessed by genetic analysis were excluded. Libraries were generated using an IgM constant-specific primer and the 5'MTPX approach. Around one million sequences per library were produced on the Illumina MiSeq platform. To infer immunoglobulin heavy chain germline genes, we used the IgDiscover tool. This analysis enabled us to infer the complete genetic and allelic composition of the IGH locus for the individuals. The complete process is depicted in Figure 21.

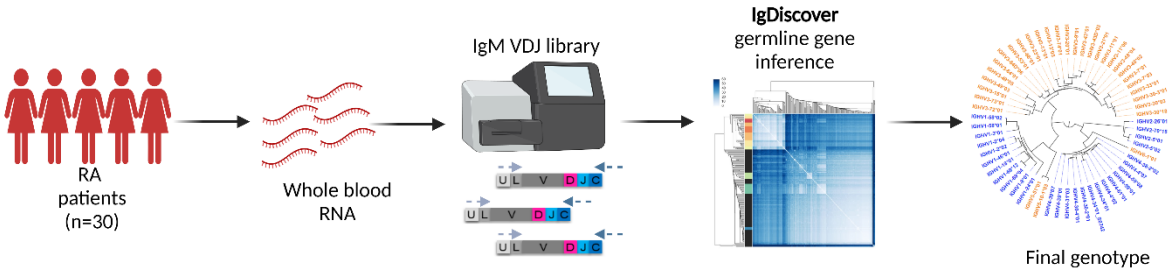


Figure 21 The scheme shows the study setup from the baseline characterization of the immunoglobulin heavy chain allele. Whole blood from 30 RA patients was used to generate IgM VDJ libraries. These libraries were used for genotype inference by IgDiscover.

At baseline, we observed some common structural variation as well as several novel alleles in our cases. We found structural variation around the IGHV3-30 and IGHV1-69 families as well as a mosaic pattern of genotypes for the more centromeric IGHV1-8, IGHV3-9, IGHV3-64D

and IGHV5-10-1 genes. Furthermore, we identified 8 novel IGHV alleles, which underlines the incompleteness of the IMGT reference database.

5.10 INFERENCE OF HAPLOTYPES BASED ON IGHJ6 HETEROZYGOSITY

IGHJ6 heterozygosity was found in 9 out of our 30 individual cases. As VDJ recombination only occurs along one chromosome at a time, association of the IGHJ6*02 or *03 alleles with a particular IGHV allele in the expressed repertoires allows inference of the haplotype status of the IGHV alleles. In this manner, IGHV genes can be inferred as homozygous, heterozygous, duplicated or deleted. This inference technique is depicted in Figure 22.

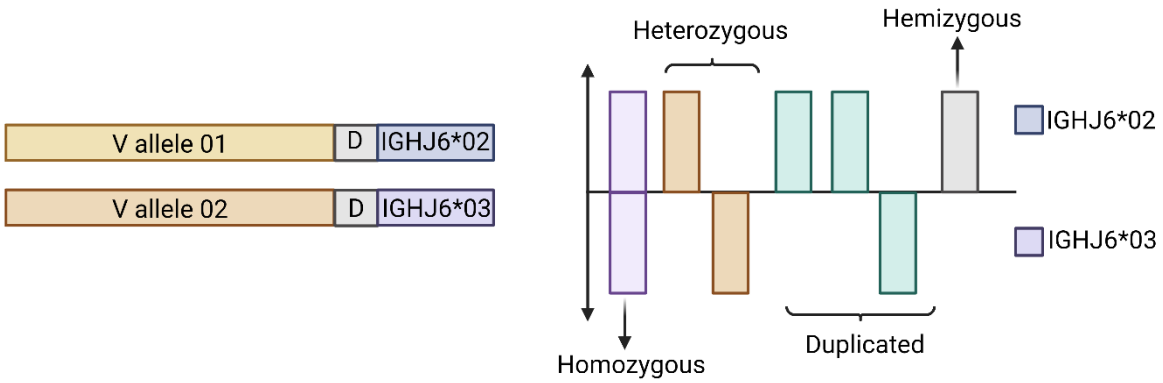


Figure 22 The IGHJ6 gene heterozygosity allows V allele haplotyping. Depending on the co-occurrence in the sequences from the libraries, V alleles are inferred to be homozygous, heterozygous, duplicated or hemizygous.

We found an unexpected hemizygosity for the IGHV4-34 gene with the IGHV4-34*01 allele only present on one chromosome in two individuals of our cohort. IGHV4-34*01 is a very common allele, that shows low levels of allelic variation. This hemizygosity can be either explained by genomic deletion of IGHV4-34 on one chromosome or by mechanisms that interfere with recombination, which results in the absence of the allele in the expressed IgM repertoire.

5.11 RSS SNP VARIANT ASSOCIATING WITH DECREASED IGHV4-34 EXPRESSION

Genomic amplification and Sanger sequencing revealed a novel variant of the IGHV4-34 gene that was 100% identical with the IGHV4-34*01 allele along the entire V sequence but had a SNP variant in the second position of the RSS heptamer sequence. The CACAGTG RSS sequence is replaced by the non-canonical CTCAGTG variant, which presumably results in defective V-DJ recombination explaining the appearance of IGHV4-34 as hemizygous in these individuals.

As 21 out of the 30 cases could not be haplotyped, we were curious, whether the RSS variant could be found in any of those additional individuals. The RSS variant introduced a DdeI restriction site in the genomic sequence, so that we could amplify an 84bp region spanning the RSS variant and make a RFLP analysis. Presence of the RSS variant would result in two diagnostic bands of 53 and 31bp in length after digestion with DdeI. A summary of this RFLP with DdeI is shown in Figure 23. We found this variant in one additional RA case and in 3 out

of 30 additionally tested healthy controls from the EIRA study group. This gave a prevalence of 10% in our Swedish study group.

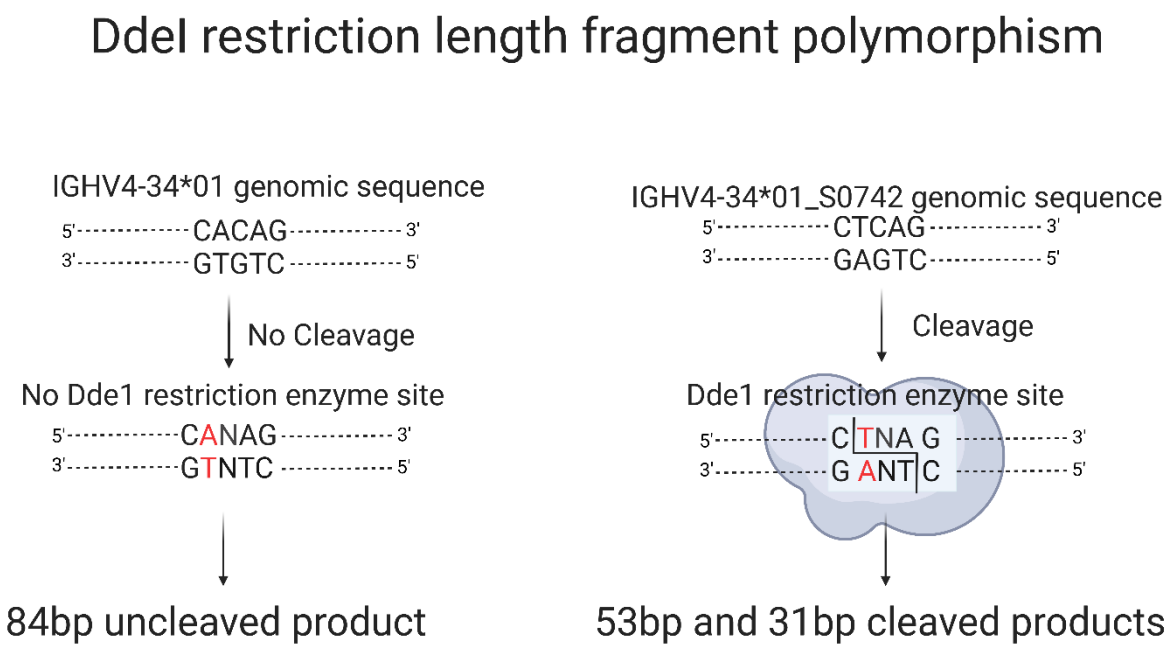


Figure 23 The figure shows the sequence information, on which the DdeI digest is build up upon.

The variant was consistent with the SNP rs148342179 T described in the 1000 Genomes Project. It is present in 0.5% of all samples (ALL) but is enriched in the European sample group with 1.8% (EUR) and especially in the Finnish population group with 4.5% (FIN). It was not found in either the African (AFR) or East Asian samples (EAS) and only in small frequencies in the American (AMR: 0.6%) and South Asian (SAS: 0.1%) samples.

5.12 IDENTIFYING A NOVEL IGHJ6 ALLELE SUITABLE AS A HAPLOTYPING ANCHOR

Besides identifying novel IGHV alleles, we also identified a novel IGHJ6 allele, which was characterized by the deletion of a triplet base GGT on the coding genomic minus strand. This deletion, which was validated with targeted genomic PCR and Sanger sequencing, results in the deletion of a glycine at the protein level and is described as SNP variant rs74454466.

The presence of this variant IGHJ6 allele provided an additional haplotyping possibility for the specific individual. This individual was also haplotypable by a heterozygous IGHD gene, but there is additional value from haplotyping by a heterozygous IGHJ6 gene since the sequence counts for the recombinations with each allele are higher for the few IGHJ genes. Haplotype analysis with either IGHD or IGHJ anchor was possible in 15 out of our 30 RA cases.

6 DISCUSSION

6.1 B CELL YIELDS AND DIFFERENTIATION STAGES IN EARLY RA

We investigated synovial B cells from two ACPA+/RF+ and two ACPA-/RF- RA patients obtained within two days of diagnosis and before the start of therapy. Thereby, we found differentiated B cell subsets that matured from crosstalk with T cells and resided in a survival niche. The differentiation process could be followed by clonotype analysis identifying shared clones between the memory B cell and the plasma cell compartment. Another characteristic feature that we found was citrulline reactivity of an expanded plasma cell clone.

We combined single cell RNA sequencing with spatial transcriptomics analysis to study the spatial context of the predicted single cells in the tissue. This distinguishes our study from previous studies, where dispersed synovial tissue was analyzed at mixed time points of the disease[121,282,283]. Similar to these studies, we found that not all RA patients had high amounts of B cells in the biopsy, which also limited the phylogenetic analysis of the B cell lineage. B cell yields correlated with disease activity at the time of sampling and amount of tissue obtained but was independent of ACPA status. This is intriguing since the genetic HLA-DR association and the presence of autoantibodies suggest a stronger B-T cell drive in ACPA+ patients. In contrast, RF marks an antibody response that is independent of the B-T cell drive[284]. Thus, we hypothesize that the RF response is generated outside the synovial lymphoid infiltrates. We have shown that also the synovial biopsies from ACPA-/RF- patients are attractive for T cell-dependent B cell differentiation, potentially also for unspecific immune responses. Indeed, ACPA- patients may also benefit from B and T cell targeted therapy, although ACPA+ patients are more likely to do so[285,286].

Furthermore, our data implies local expansion and differentiation of plasma cells as shown by the CCP2+ monoclonal antibody that was found in two plasma cell copies. Thus, the joint seems to support a local immunological B cell maturation process via T cell help involving PD-1 and CXCL13 as previously implicated for longstanding ACPA+ RA[93]. Another B cell feature that has been described mostly in SLE in recent years, but also in RA, is atypical memory[115,287]. Although, we found double negative B cells in the biopsies, we could not identify a B cell cluster that would have markers such as CD11c or T-bet or autoimmune associated B cells. Instead, this subset may be associated with chronic immune responses or even immunosuppressive therapy, a stage that the patients in our study have not yet reached.

Additionally, we describe the nature of differentiated B cells and find supporting niche structures in the joint in the vicinity of the lymphocyte infiltrates. Still, it was not clear, whether plasma cells were generated in these areas by differentiation processes or whether they were homing into these areas. Our data suggests both processes as we find evidence for local B cell differentiation by our clonality studies, as well as BAFF- and APRIL-dependent B cell survival and strong CXCR4 expression in plasma cells, and elevated CXCL12 expression in the supporting tissue. The CXCR4-CXCL12 axis is known to be crucial for plasma cell survival in the bone marrow[288], but has not been described at the site of joint inflammation. However,

we could not identify colocalization signals from potentially supporting fibroblasts and plasma cells, which may be a consequence of the limited resolution of the spatial transcriptomics.

6.2 LIMITATION EFFECTS OF SMALL SAMPLE SIZES

Another limiting factor of our study is the small sample size, that does not allow for drawing conclusions about RA in its full heterogeneity, when it comes to the composition of the joint inflammation[289]. Still, our study provides a detailed picture of the composition and context of synovial B cells for two ACPA- and two ACPA+ samples and might as such be used for further studies.

6.3 REACTIVITY AND CLONALITY OF DIFFERENTIATED SYNOVIAL B CELLS IN EARLY RA

Interestingly, we found a novel plasma cell derived ACPA, that displays some classic features such as N-linked glycosylation sites in the F_{ab} fragments and increased SHM[188,290,291]. However, this ACPA has a narrow reactivity pattern towards a mutated vimentin peptide originally identified in RA joint tissue[292] in comparison to previously identified ACPAs, which are broadly citrulline reactive and either identified in other compartments or from other B cell subsets[155,214,293–296].

Lastly, 60% of our plasma cell derived monoclonal antibodies from expanded clones had representatives in both the memory and the plasma cell compartment. This further suggests a local B cell differentiation process and recent recall responses from memory cells. Among our plasma cell derived monoclonal antibodies, we also found six polyreactive clones. Polyreactivity has been described in different settings such as HIV infection and vaccination[297,298], but also in healthy donors[299]. Currently, we do not understand whether polyreactive clones contribute to RA disease features or whether they are a result of a dysregulated B cell compartment.

6.4 EARLY DISEASE B CELL PROCESSES IN RA

In summary, our study of synovial biopsies collected from untreated RA patients within two days of diagnosis shows that there are many similarities to processes that are shown to be active at later disease stages. The prominent lymphocyte infiltrations suggest local B-T cell crosstalk and the immunoglobulin sequence analysis demonstrates local plasma cell differentiation. This was true independently of the ACPA status, however our study only included B cell rich tissues and our findings may thus be not completely representative. We equally identified a clonally expanded ACPA expressing plasma cell in the inflamed joint that was reactive to modified citrullinated vimentin, a specificity associated with osteoclast activation[155]. Plasma cell survival niches were already present at this early time point of the disease supporting the notion, that such structures are attractive therapeutic targets in RA[300].

As individuals at risk are genetically predisposed, we further wanted to understand the immunoglobulin germline configuration, which can be inferred from NGS immunoglobulin libraries.

6.5 NGS LIBRARY PREPARATION METHODS

There are two library preparation methods for repertoire sequencing analysis: 5'RACE and 5'MTPX. While 5'RACE introduces a universal amplification primer site and should theoretically amplify the target in an unbiased manner[301], we show that length considerations limit the 5'RACE performance on the Illumina MiSeq. The more widely used 5'MTPX libraries[302,303] are informed by knowledge on the 5' leader sequences, but do not exceed technical amplicon length limitations that may lead to lower quality base calling. This issue is partially mitigated by use of a template switch primer containing the Illumina Read1 sequence in the 5'RACE libraries. However, long 5'UTRs in the IGHV3 family and long CDR3s may persistently worsen the 5'RACE library quality.

6.6 ADVANTAGES AND PITFALLS INFLUENCING HIGH QUALITY LIBRARY PRODUCTION

Minimizing the amplicon length for the Illumina MiSeq system bears the advantage of sequencing in a higher throughput manner compared to longer read approaches, such as PacBio[304]. We described library preparation methods providing full-length V(D)J sequences suitable for V gene allele inference in contrast to protocols using primers targeting the Immunoglobulin FRW1[305].

By investigating the amount of RNA used for cDNA synthesis as a factor to ensure high quality library production, we found less diverse libraries using 50 or 100ng of RNA, while 200 or 500ng of RNA were adequate. These libraries are superior when inferring germline genes, as they are more sensitive in detecting independent CDR3 and thus independently recombined sequences. This may be important when preparing libraries from sorted cells, because in that case the amount of RNA is limited[306].

6.7 SUITABLE LIBRARY PREPARATION METHODS FOR DIFFERENT PURPOSES

From our studies, we conclude that 5'MTPX approaches are preferable, if 5' leader sequences are relatively well-characterized such as in humans. However, there are research animal models such as rabbits, guinea pigs, macaques, rats and mice, with the exception of C57BL/6 mice, where a prior 5'RACE library is warranted to identify the 5' leader sequences. While much more work is required to characterize the immunoglobulin loci in these non-human research models, there is also a need for a more comprehensive immunoglobulin germline database for humans, including from different population groups[248]. In addition to the known alleles, we found three novel IGHV, one novel IGKV and six novel IGLV alleles from one individual in our cohort. The identification of the full variation within the human Ig germline repertoire will require extensive analysis of a large number of individuals from different populations[240,246,247,264,307,308].

In summary, we have presented the advantages and limitations of two library preparation methods, which we believe will be of broad utility for the immune repertoire sequencing

community. In the following, we will present a study, where we used a high-quality library preparation approach to understand the baseline IgH genotypes in a local RA study group.

6.8 GENOTYPING THE IGHV LOCUS FOR ASSOCIATION STUDIES

Genotyping the complex IGHV locus with the aim to perform disease association studies is a challenge. It requires complete and accurate identification of genotypes from each individual including deleted or duplicated genes and novel variants that may be present in the population. Particular confounding bias may occur, when case and control groups from disease cohorts are not population-stratified[309]. The goal of association studies is to find specific polymorphisms that are at different frequencies in cases versus controls.

SNP arrays and whole genome sequencing have revolutionized the field of complex disease genetics. GWAS have helped to identify causal genes and mechanisms, drug targets, biomarkers and to enable risk prediction. In RA, there are 150 risk loci associations implicated by GWAS from different populations[26]. The class II human leukocyte antigen loci show the strongest association. Together with efficacy of anti-CD20 therapy, rituximab[68], these data implicate T and B cells in RA pathogenesis.

Currently, GWAS are not informed by the highly complex immunoglobulin loci including the heavy and light chain loci. Still, immunoglobulins are important B cell effector molecules in RA, which is why genetic variants in their loci within the general population may be an important component predisposing to disease. Hence, the genetic variation may be a valuable source for disease association studies. It has been shown for example, that each copy of IGHV4-61*02 is associated with an increased risk of rheumatic heart disease in an indigenous population from the South Pacific[310]. Our NGS approach identifying the genotypes of individuals forms the basis for analyzing IGHV variation in any disease. However, interpretation of IGHV allelic variation should be handled with caution as population-based variations might change allelic frequencies. Therefore, a first step towards this goal is to better understand the immunoglobulin germline gene variation in relevant study groups.

6.9 IMPLICATIONS OF THE NOVEL IDENTIFIED ALLELES

We identified two novel alleles in our study by haplotyping 30 individuals. We found a novel IGHV4-34 allele that was present in 10 percent of cases and controls and that was characterized by a SNP in the RSS region, which likely interfered with the recombination, so that the IGHV4-34 gene appeared hemizygous in the haplotype analysis. Furthermore, we found a novel IGHJ6*05_S6029 allele, which was characterized by a deletion of a triplet base and allowed to haplotype one additional individual by using the J6 heterozygosity as an anchor.

The participants in the EIRA study were recruited in the middle and southern parts of Sweden. Data from the 1000 Genomes Project show that the Finnish population has an increased frequency of the variant allele that we identified in the EIRA study. A prevalence of about 10 percent results in around 1 percent of the Finnish/Scandinavian population being homozygous for this variant allele, which leads to complete absence in the expressed repertoire. This raises

additional questions for the antibody response in these individuals. Similarly, homozygous deletions were reported for the IGHV3-30/IGHV4-31 region (hv3005) in RA patients[256] and SLE patients of Korean[257] and Caucasian[258] ethnicity.

The IGHV4-34 gene was previously associated with autoimmunity[145], but the specificities responsible for this remain unclear. Interestingly, the IGHV4-34*01 allele contains a germline-encoded N-linked glycosylation site[311]. Variable domain glycans have been associated with autoantibodies in RA[312]. Simultaneously, antibodies produced by a self-reactive B-cell lost the capacity of autoantigen binding, when N-linked glycosylation was introduced[313]. Here, we identified a non-canonical RSS heptamer that led to non-productive recombination of the IGHV4-34 gene, which was present in 10 percent cases and controls.

A similar observation, where an RSS polymorphism affected V gene usage, has been described for a variant Igκ V gene, IGKV2-29D, enriched in a Native American population, which was associated with *Haemophilus influenzae* type b susceptibility[262,263]. Finding an additional IGHJ6 allele highlights the fact that the genetic diversity of both, IGHV and IGHJ genes, is incompletely characterized. Our study was not powered for association analysis, however the EIRA study group offers the possibility of extending our study to large numbers[314].

6.10 PREVIOUS ASSOCIATION STUDIES AND LESSONS FROM OUR STUDY

Association studies for the immunoglobulin alleles are limited and reviewed elsewhere[315,316]. The impact of this variation on disease risks is understudied[308,317].

Our results implicate that population based genetic variance of IG alleles is likely to be common[261,264,318,319] and without the adequate baseline information that we provided for few alleles, apparent but erroneous associations may be identified. Likewise, true associations may only become apparent when accurate information of allelic frequencies in the target population is known.

7 CONCLUSIONS

In this thesis, I present three studies that have helped me to better understand the scientific method and research process, and which provide new information to the research community.

I studied the synovial B cell compartment and context of early Rheumatoid Arthritis patients. These studies taught me to consider advantages and limitations of study design. The biopsies were analyzed in depth, however practical limitations made our sample size and power very small. Still, we were able to detect processes such as B-T cell crosstalk, B cell differentiation footprints, citrulline reactivity and plasma cell survival niches in these early biopsy materials. These processes give us clues about the early disease development. Notably, we could only suggest little differences between ACPA- and ACPA+ patients with lymphocyte rich infiltrates in the synovial tissue, which would suggest a differential pathogenesis mechanism between ACPA- and ACPA+ disease. However, a previous study that differentiated between lympho-myeloid, diffuse-myeloid and pauci-immune fibroid tissue pathotypes using bulk RNA sequencing also reported a correlation between the plasma cell gene module and ACPA or RF titer[320], which is consistent with the notion that local *in situ* plasma cell differentiation and perpetuation are associated with high ACPA titer[321,322]. A study with larger sample size should be performed in order to elucidate, whether B cell activation and maturation is different between those two groups as genetic studies imply[323].

Furthermore, I evaluated two library preparation methods that primed my methodological rigor in the laboratory by setting up parallel 5'MTPX and 5'RACE reactions for the same set of healthy donor samples. Then, I found several ways of comparing the resulting libraries in the analysis asking, which method would be preferable. The answer was conditional in that sense, that despite the opposite claim, the 5'RACE would give more IGHV family biased libraries, while the 5'MTPX libraries resulted in a more diverse repertoire representation. Still, if sequence information for primer design is limited, 5'RACE is helpful in generating that information for 5'MTPX primer design.

Lastly, I studied immunoglobulin heavy chain VDJ genotypes in 30 individuals from the Swedish EIRA study. These studies taught me the importance of population stratification in disease association studies. There is a great need of describing disease associations especially in the overlooked immunoglobulin loci in different populations. This might give insights into many disease study areas. This study contributes to the baseline characterization of the immunoglobulin heavy chain locus and illustrates the importance of additional investigations in this area.

8 POINTS OF PERSPECTIVE

Clinical translation of the performed biopsy studies is still in its infancy. Compared to diagnostic criteria in oncology, RA is not diagnosed with the help of tissue biopsies. It is up for debate, whether it would be useful to perform clinical assessments with routine biopsies, while the oncology field is drifting towards liquid biopsy examinations. It is questionable, whether we could explore additional patient stratifying characteristics that would justify the cost of taking routine biopsies. Although we had very small sample sizes, our studies indicated that there were signs of B-T cell crosstalk, local B cell differentiation and plasma cell survival mechanisms at play in lymphoid rich biopsies independent of the ACPA status. Based on our study, we suggest further exploration of two treatment regimens:

First, we implicate that ACPA- patients that are non-responsive to methotrexate treatment could also benefit from B-T cell crosstalk targeting therapy such as abatacept[285,286], if they present with lymphocyte rich tissue. Interestingly, a previous study[324] has shown that abatacept treatment downregulates CD38 in synovial tissue, a surface marker that we found highly expressed on synovial tissue plasma cells. Furthermore, abatacept reduced $CD20^+$ B cell counts in synovial tissue[325] indicating that the impaired B-T cell crosstalk affects the B cell presence and differentiation in the joints. Hence, early treatment with abatacept might prevent disease chronicity that is driven by local antigen-specific plasma cells in lymphoid rich tissue. Notably, the experienced clinician could also predict the yield of B cells that were obtained from the biopsies by his examination, thus potentially bypassing the actual biopsy requirement. However, this observation also needs further validation and objective criteria determination.

Secondly, we suggest the plasma cell niche structure in the joint as a therapeutic target. In mice, the specific CXCR4 antagonist AMD3100 reduced the incidence of collagen-induced arthritis and modestly inhibited disease severity, however AMD3100 treatment seemed to not act via interference with humoral or cellular responses to collagen type II, but rather through inhibition of CXCL12-elicited cell migration[326]. Plasma cells have been targeted in RA *ex vivo* using the CD38 binding daratumumab[327] and by atacicept targeting BAFF and APRIL[328]. However, atacicept did not reduce the ACPA levels[329] indicating the escape of citrulline specific plasma cells to treatment. Still, it is up for debate whether ACPA producing B cells should be the target of RA therapy, especially with emerging data, that ACPAs may be protective[330].

For future developments, we envision to perform the described genetic antibody analysis on larger disease study groups including the population matched controls. This will allow identification of preferentially used immunoglobulin alleles and localization of disease associated SNPs in the immunoglobulin locus. These studies might enable stratification of at-risk individuals that might have the propensity to develop a specific antibody response, for example the ACPA response. Recently, it has been discussed to also treat at-risk individuals to prevent disease or to alleviate disease severity[331]. This would be a paradigm shift for RA treatment, but in line with clinical practice in other chronic conditions such as cardiovascular

disease, where statins and other antihypertensive medications are widely prescribed to reduce risk.

9 ACKNOWLEDGEMENTS

This PhD journey would not have been possible if I didn't have had the right people around me that kept encouraging me.

TO MY SUPERVISORS

Vivi and **Nilla**, thanks for nourishing the spark in me when I was lost, when I started my PhD with you and on the course of it. Both of you have contributed to keep me in academia by valuable advice and as role models. Thank you for giving me freedom within my projects and for guiding me when I asked for it.

Martin, thank you for being such a dedicated scientist with a hint of black humor. You are very knowledgeable, and it is very entertaining to be your desk neighbor.

Leonid, thank you for being very approachable. It is very easy to talk to you and it is great, that you enabled me to progress quickly.

TO SAMPLE DONORS AND CO-AUTHORS

Thanks to all the **patients** and **donors** that provided samples. My research would not have been possible without you. Thank you also to all my **co-authors** for fruitful collaborations, open minds and constructive feedback. Thank you to the **funding agencies** that enabled my work.

TO MY COLLEAGUES AT MTC

Sharesta, thank you for your care, when things got rocky on my road, and thank you for making so many days enjoyable. Thank you also for proofreading my thesis. **Sanjana**, thank you for all the kind words and acts and for sharing your food culture with me. Thank you also for proofreading my thesis. **Pradee**, thank you for your happy and open nature, which made me share all my struggles and successes with you. **Marco**, thank you for being so excited about food. It truly makes life so much more enjoyable. **Mateusz**, thank you for the geeky talks with coffee in the mornings. It is great to exchange ideas about analysis with you. **Moni**, thank you for being so kind and reliable. It is great to be able to share all ups and downs with you. **Martina**, thanks for keeping the lab in order. **Ben**, thank you for taking care of me, when I struggled. **Mark**, thank you for always knowing the wifi back-up password. **Xaquin**, thank you for sharing your enthusiasm for science with the office. **Sung Yong**, thank you for being such a kind colleague. **Mariia**, thank you for friendly lunches. **Dan**, thank you for friendly encounters in the corridor. **Ganesh**, thank you for your encouraging words, when I felt lost. **Néstor**, thank you for always pushing me to the limits. It was fun arguing with you. **Camilla**, thank you for great scientific discussions on B cells and immunoglobulins in the coffee corner.

TO MY COLLEAGUES AT CMM

Kostas, thank you for the great collaboration on the B cell manuscript. **Sara**, thank you for teaching me laboratory skills, when I was lost, and for sharing desk space and a room with me during the KiiM retreat. **Karine**, thank you for sharing your passion for science and France with me. It is fun to work with you. **Lina**, thank you for being such a fun person and for sharing your excitement about science and bioinformatics. **Angeles**, thank you for kindly offering your desk to me and for bringing a Mexican vibe to parties. **Erik**, tack för att jag kan öva min svenska med dig. Det var kul att få biopsier. **Alexandra**, thank you for your kind and happy nature. **Christina**, thank you for sharing the EULAR experience with me. **Anatoly**, thank you for always questioning everybody's work. It helped me grow. **Gena**, thank you for challenging my biochemistry knowledge. **Heidi**, thank you for your Finnish humorous soul. **Ravi** and **Bruno**, thanks for your confidence in my bioinformatics skills. **Khaled**, thank you for introducing me into antibody production and for helping me practice my French. **Annika**, thank you for your reliable sorting skills and your compassion when finding more or less B cells in every sample. **Marianne**, thank you for your help with the biopsies. **Vijay**, thank you for teaching me antibody harvesting and your great smile. **Johanna**, **Ragnhild** and **Lena**, thank you for the good spirit and keeping the monoclonal antibodies running and in order. Thank you also for the help with the Swedish popular science summary. **Katy**, thank you for all the fun times indulging the lab. **Yan**, thank you for your relaxed nature. **Peter**, thank you for your help running the ELISAs. **Caroline,** thank you for kick-starting my journey. **Johanna**, thank you for taking me to a proper Swedish midsommar and for the help sewing. **Julia,** thank you for friendly chats over coffee. **Yunbing,** thank you for fun walks in nature. **Vlad**, thank you for awesome coffee and your positive attitude. **Jyoti**, thank you for your fun company.

TO FRIENDS AND FAMILY

Henna, thank you for being such a fun and outgoing Finnish friend. You have truly supported my well-being as a peer. Thanks for proofreading my thesis. **Meike** und **Tilo**, **Carmen** und **Börge**, danke für spannende Dinner und Spieleabende. **Tobias** och **Helena**, tack för att vara mina danslärare or att för att undervisa mig i svenska. Ni har öppnat svensk kultur för mig och det var väldigt viktigt under min resa. Thanks to the badminton crew for many fun matches: **Maria**, **Minna**, **Shalina**, **Mansi**, **Tom**, **Mauricio**, **Walentin**, **Vaish**, **Alek**, **Mark**, **Henna** and **Tommy**. **Sara**, thank you for taking care of me, when I was finding my way through academia. **Johanna** und **Natalie**, danke für gemeinsames Problemlösen beim Klettern und in allen Lebenslagen. **Steffi**, danke für deine Gastfreundschaft in Boston. Danke an die **Glykos** von der Studienstiftung und die **Kurs**- und **Akademieleiter** von der JGW für die vielen gemeinsamen Erlebnisse innerhalb und außerhalb der Wissenschaft. **Florian**, danke für den Teil des Weges, den wir gemeinsam bestritten haben. **Rosanna**, **Anna, Lucas** und **Denise**, danke, dass ich mich immer auf euch verlassen kann und danke für's gegenlesen. **Susanne, Peter, Barbara** und **Lisi**, danke, dass ihr mich so lieb aufgenommen habt und danke für die Care-Pakete. **Omi Moni**, **Opa Heinz**, **Omi Christine** und **Opa Klaus**, danke, dass ihr immer an mich geglaubt habt und meine Reise immer unterstützt habt. **Mama**, **Papa** und **Pia**, danke, dass ihr immer für

mich da seid, mich machen lasst und mich unterstützt, mich zu verwirklichen. **Tommy**, danke, dass du mich gefunden hast, dass du mir immer kleine Botschaften zum Frühstück hinterlässt und mich zu allen Zeiten unterstützt.

10 REFERENCES

1. Smolen, J. S. *et al.* Validity and reliability of the twenty-eight-joint count for the assessment of rheumatoid arthritis activity. *Arthritis Rheum.* **38**, 38–43 (1995).

2. Pujades-Rodriguez, M. *et al.* Rheumatoid Arthritis and Incidence of Twelve Initial Presentations of Cardiovascular Disease: A Population Record-Linkage Cohort Study in England. *PLoS One* **11**, e0151245 (2016).

3. Matcham, F. *et al.* The impact of rheumatoid arthritis on quality-of-life assessed using the SF-36: A systematic review and meta-analysis. *Semin. Arthritis Rheum.* **44**, 123–130 (2014).

4. Findeisen, K. E., Sewell, J. & Ostor, A. J. K. Biological Therapies for Rheumatoid Arthritis: An Overview for the Clinician. *Biologics* **15**, 343–352 (2021).

5. Safety and Effectiveness of Adalimumab in Patients Diagnosed With Rheumatoid Arthritis - Full Text View - ClinicalTrials.gov. Available at: https://clinicaltrials.gov/ct2/show/study/NCT01078571?term=adalimumab&cond=Rheumatoid+Arthritis&draw=2&rank=3. (Accessed: 21st September 2022)

6. Sarzi-Puttini, P. *et al.* Clinical, Ultrasound, and Predictability Outcomes Following Certolizumab Pegol Treatment (with Methotrexate) in Patients with Moderate-to-Severe Rheumatoid Arthritis: 52-Week Results from the CZP-SPEED Study. *Adv. Ther.* **35**, 1153–1168 (2018).

7. Hobbs, K. *et al.* Randomized, double-blind, placebo-controlled study to evaluate the efficacy and safety of etanercept in patients with moderately active rheumatoid arthritis despite DMARD therapy. *Springerplus* **4**, 1–7 (2015).

8. Emery, P. *et al.* Efficacy of golimumab plus methotrexate in methotrexate-naïve patients with severe active rheumatoid arthritis. *Clin. Rheumatol.* **33**, 1239–1246 (2014).

9. Elliott, M. . *et al.* Randomised double-blind comparison of chimeric monoclonal antibody to tumour necrosis factor α (cA2) versus placebo in rheumatoid arthritis. *Lancet* **344**, 1105–1110 (1994).

10. A Phase III Study of Abatacept in Patients With Rheumatoid Arthritis and an Inadequate Response to Methotrexate - Full Text View - ClinicalTrials.gov. Available at: https://clinicaltrials.gov/ct2/show/NCT00409838?term=abatacept&cond=Rheumatoid+Arthritis&draw=2&rank=7. (Accessed: 21st September 2022)

11. Use of Tocilizumab for Rheumatoid Arthritis (RA) in Daily Routine - Full Text View - ClinicalTrials.gov. Available at: https://clinicaltrials.gov/ct2/show/NCT02809833?term=tocilizumab&cond=Rheumatoid+Arthritis&draw=2&rank=1. (Accessed: 21st September 2022)

12. Fleischmann, R. *et al.* Long-term safety and efficacy of sarilumab over 5 years in patients with rheumatoid arthritis refractory to TNF inhibitors. *Rheumatology (Oxford).* **60**, 4991–5001 (2021).

13. Study to Examine Efficacy and Safety of Rituximab in Participants With Rheumatoid Arthritis - Full Text View - ClinicalTrials.gov. Available at:

https://clinicaltrials.gov/ct2/show/NCT02699892?term=rituximab&cond=rheumatoid+arthritis&draw=2&rank=3. (Accessed: 21st September 2022)

14. Kineret in the Treatment of Rheumatoid Arthritis - Full Text View - ClinicalTrials.gov. Available at: https://clinicaltrials.gov/ct2/show/NCT02915094?term=anakinra&cond=Rheumatoid+Arthritis&draw=2&rank=2. (Accessed: 21st September 2022)

15. Tanaka, Y. *et al.* Modified- versus immediate-release tofacitinib in Japanese rheumatoid arthritis patients: a randomized, phase III, non-inferiority study. *Rheumatology (Oxford).* **58**, 70–79 (2019).

16. Taylor, P. C. *et al.* Safety of baricitinib for the treatment of rheumatoid arthritis over a median of 4.6 and up to 9.3 years of treatment: final results from long-term extension study and integrated database. *Ann. Rheum. Dis.* **81**, 335–343 (2022).

17. Clemente-Casares, X. *et al.* Expanding antigen-specific regulatory networks to treat autoimmunity. *Nature* **530**, 434–440 (2016).

18. Malemba, J. J. *et al.* The epidemiology of rheumatoid arthritis in Kinshasa, Democratic Republic of Congo--a population-based study. *Rheumatology* **51**, 1644–1647 (2012).

19. Tobón, G. J., Youinou, P. & Saraux, A. The environment, geo-epidemiology, and autoimmune disease: Rheumatoid arthritis. *J. Autoimmun.* **35**, 10–14 (2010).

20. Peschken, C. A. & Esdaile, J. M. Rheumatic diseases in North America's indigenous peoples. *Semin. Arthritis Rheum.* **28**, 368–391 (1999).

21. Myasoedova, E., Crowson, C. S., Kremers, H. M., Therneau, T. M. & Gabriel, S. E. Is the incidence of rheumatoid arthritis rising?: Results from Olmsted County, Minnesota, 1955-2007. *Arthritis Rheum.* **62**, 1576–1582 (2010).

22. Lee, D. M. & Weinblatt, M. E. Rheumatoid arthritis. *Lancet* **358**, 903–911 (2001).

23. Hensvold, A. H. *et al.* Environmental and genetic factors in the development of anticitrullinated protein antibodies (ACPAs) and ACPA-positive rheumatoid arthritis: an epidemiological investigation in twins. *Ann. Rheum. Dis.* **74**, 375–380 (2015).

24. MacGregor, A. J. *et al.* Characterizing the quantitative genetic contribution to rheumatoid arthritis using data from twins. *Arthritis Rheum.* **43**, 30–37 (2000).

25. Svendsen, A. J. *et al.* On the origin of rheumatoid arthritis: the impact of environment and genes--a population based twin study. *PLoS One* **8**, (2013).

26. Padyukov, L. Genetics of rheumatoid arthritis. *Semin. Immunopathol.* **44**, 47–62 (2022).

27. Stastny, P. Mixed lymphocyte cultures in rheumatoid arthritis. *J. Clin. Invest.* **57**, 1148–57 (1976).

28. Gregersen, P. K., Silver, J. & Winchester, R. J. The shared epitope hypothesis. an approach to understanding the molecular genetics of susceptibility to rheumatoid arthritis. *Arthritis Rheum.* **30**, 1205–1213 (1987).

29. Lenz, T. L. *et al.* Widespread non-additive and interaction effects within HLA loci modulate the risk of autoimmune diseases. *Nat. Genet.* **47**, 1085–1090 (2015).

30. Kim, K., Bang, S.-Y., Lee, H.-S. & Bae, S.-C. Update on the genetic architecture of rheumatoid arthritis. *Nat. Rev. Rheumatol.* **13**, 13–24 (2017).

31. Sugiyama, D. *et al.* Impact of smoking as a risk factor for developing rheumatoid arthritis: a meta-analysis of observational studies. *Ann. Rheum. Dis.* **69**, 70–81 (2010).

32. Källberg, H. *et al.* Smoking is a major preventable risk factor for rheumatoid arthritis: estimations of risks after various exposures to cigarette smoke. *Ann. Rheum. Dis.* **70**, 508–511 (2011).

33. Stolt, P. *et al.* Silica exposure is associated with increased risk of developing rheumatoid arthritis: results from the Swedish EIRA study. *Ann. Rheum. Dis.* **64**, 582–586 (2005).

34. Webber, M. P. *et al.* Nested Case-Control Study of Selected Systemic Autoimmune Diseases in World Trade Center Rescue/Recovery Workers. *Arthritis Rheumatol.* **67**, 1369–1376 (2015).

35. Too, C. L. *et al.* Occupational exposure to textile dust increases the risk of rheumatoid arthritis: results from a Malaysian population-based case–control study. *Ann. Rheum. Dis.* **75**, 997–1002 (2016).

36. Kharlamova, N. *et al.* Antibodies to *Porphyromonas gingivalis*Indicate Interaction Between Oral Infection, Smoking, and Risk Genes in Rheumatoid Arthritis Etiology. *Arthritis Rheumatol.* **68**, 604–613 (2016).

37. Konig, M. F. *et al.* Aggregatibacter actinomycetemcomitans-induced hypercitrullination links periodontal infection to autoimmunity in rheumatoid arthritis. *Sci. Transl. Med.* **8**, 369ra176 (2016).

38. Scher, J. U. *et al.* Expansion of intestinal Prevotella copri correlates with enhanced susceptibility to arthritis. *Elife* **2**, e01202 (2013).

39. Ljung, L. & Rantapää-Dahlqvist, S. Abdominal obesity, gender and the risk of rheumatoid arthritis – a nested case–control study. *Arthritis Res. Ther.* **18**, 277 (2016).

40. Makrygiannakis, D. *et al.* Smoking increases peptidylarginine deiminase 2 enzyme expression in human lungs and increases citrullination in BAL cells. *Ann. Rheum. Dis.* **67**, 1488–1492 (2008).

41. Reynisdottir, G. *et al.* Signs of immune activation and local inflammation are present in the bronchial tissue of patients with untreated early rheumatoid arthritis. *Ann. Rheum. Dis.* **75**, 1722–1727 (2016).

42. McGraw, W. T., Potempa, J., Farley, D. & Travis, J. Purification, characterization, and sequence analysis of a potential virulence factor from Porphyromonas gingivalis, peptidylarginine deiminase. *Infect. Immun.* **67**, 3248–3256 (1999).

43. Pianta, A. *et al.* Two rheumatoid arthritis–specific autoantigens correlate microbial immunity with autoimmune responses in joints. *J. Clin. Invest.* **127**, 2946–2956 (2017).

44. Alpizar-Rodriguez, D. *et al.* Female hormonal factors and the development of anti-citrullinated protein antibodies in women at risk of rheumatoid arthritis. *Rheumatology* **56**, 1579–1585 (2017).

45. Salliot, C. *et al.* Female hormonal exposures and risk of rheumatoid arthritis in the French E3N-EPIC cohort study. *Rheumatology* **60**, 4790–4800 (2021).

46. Cutolo, M. *et al.* Sex hormones influence on the immune system: basic and clinical aspects in autoimmunity. *Lupus* **13**, 635–638 (2004).

47. Aho, K., Heliövaara, M., Maatela, J., Tuomi, T. & Palosuo, T. Rheumatoid factors antedating clinical rheumatoid arthritis. *J. Rheumatol.* **18**, 1282–4 (1991).

48. Rantapää-Dahlqvist, S. *et al.* Antibodies against cyclic citrullinated peptide and IgA rheumatoid factor predict the development of rheumatoid arthritis. *Arthritis Rheum.* **48**, 2741–2749 (2003).

49. Nielen, M. M. J. *et al.* Specific autoantibodies precede the symptoms of rheumatoid arthritis: A study of serial measurements in blood donors. *Arthritis Rheum.* **50**, 380–386 (2004).

50. Bos, W. H. *et al.* Arthritis development in patients with arthralgia is strongly associated with anti-citrullinated protein antibody status: a prospective cohort study. *Ann. Rheum. Dis.* **69**, 490–494 (2010).

51. Wigerblad, G. *et al.* Autoantibodies to citrullinated proteins induce joint pain independent of inflammation via a chemokine-dependent mechanism. *Ann. Rheum. Dis.* **75**, 730–738 (2016).

52. Pitzalis, C., Kelly, S. & Humby, F. New learnings on the pathophysiology of RA from synovial biopsies. *Curr. Opin. Rheumatol.* **25**, 334–344 (2013).

53. Rangel-Moreno, J. *et al.* Inducible bronchus-associated lymphoid tissue (iBALT) in patients with pulmonary complications of rheumatoid arthritis. *J. Clin. Invest.* **116**, 3183–94 (2006).

54. Bugatti, S. *et al.* Involvement of subchondral bone marrow in rheumatoid arthritis: Lymphoid neogenesis and in situ relationship to subchondral bone marrow osteoclast recruitment. *Arthritis Rheum.* **52**, 3448–3459 (2005).

55. van Steenbergen, H. W., Ajeganova, S., Forslind, K., Svensson, B. & van der Helm-van Mil, A. H. M. The effects of rheumatoid factor and anticitrullinated peptide antibodies on bone erosions in rheumatoid arthritis. *Ann. Rheum. Dis.* **74**, e3 (2015).

56. Krishnamurthy, A. *et al.* Identification of a novel chemokine-dependent molecular mechanism underlying rheumatoid arthritis-associated autoantibody-mediated bone loss. *Ann. Rheum. Dis.* **75**, 721–729 (2016).

57. Aletaha, D. *et al.* 2010 Rheumatoid arthritis classification criteria: an American College of Rheumatology/European League Against Rheumatism collaborative initiative. *Ann. Rheum. Dis.* **69**, 1580–1588 (2010).

58. Marshall, J. S., Warrington, R., Watson, W. & Kim, H. L. An introduction to immunology and immunopathology. *Allergy, Asthma Clin. Immunol.* **14**, 1–10 (2018).

59. Takeuchi, O. & Akira, S. Pattern recognition receptors and inflammation. *Cell* **140**, 805–820 (2010).

60. Vandendriessche, S., Cambier, S., Proost, P. & Marques, P. E. Complement Receptors and Their Role in Leukocyte Recruitment and Phagocytosis. *Front. Cell Dev. Biol.* **9**,

144 (2021).

61. Steinman, R. M. & Cohn, Z. A. Identification of a novel cell type in peripheral lymphoid organs of mice. II. Functional properties in vitro. *J. Exp. Med.* **139**, 380–97 (1974).

62. Chudnovskiy, A., Pasqual, G. & Victora, G. D. Studying interactions between dendritic cells and T cells in vivo. *Curr. Opin. Immunol.* **58**, 24–30 (2019).

63. Wieczorek, M. *et al.* Major histocompatibility complex (MHC) class I and MHC class II proteins: Conformational plasticity in antigen presentation. *Front. Immunol.* **8**, 292 (2017).

64. Kumar, B. V., Connors, T. J. & Farber, D. L. Human T cell development, localization, and function throughout life. *Immunity* **48**, 202 (2018).

65. Tian, Y. *et al.* Unique phenotypes and clonal expansions of human CD4 effector memory T cells re-expressing CD45RA. *Nat. Commun. 2017 81* **8**, 1–13 (2017).

66. Patil, V. S. *et al.* Precursors of human CD4+ cytotoxic T lymphocytes identified by single-cell transcriptome analysis. *Sci. Immunol.* **3**, (2018).

67. Cyster, J. G. & Allen, C. D. C. Leading Edge Review B Cell Responses: Cell Interaction Dynamics and Decisions. *Cell* **177**, 524–540 (2019).

68. Edwards, J. C. W. *et al.* Efficacy of B-Cell–Targeted Therapy with Rituximab in Patients with Rheumatoid Arthritis. *https://doi.org/10.1056/NEJMoa032534* **350**, 2572–2581 (2004).

69. Maddaly, R. *et al.* Receptors and signaling mechanisms for B-lymphocyte activation, proliferation and differentiation--insights from both in vivo and in vitro approaches. *FEBS Lett.* **584**, 4883–4894 (2010).

70. Mårtensson, I. L. & Ceredig, R. Review article: role of the surrogate light chain and the pre-B-cell receptor in mouse B-cell development. *Immunology* **101**, 435–441 (2000).

71. Nussenzweig, M. C. *et al.* Allelic exclusion in transgenic mice that express the membrane form of immunoglobulin mu. *Science* **236**, 816–819 (1987).

72. Coutinho, A., Gronowicz, E., Bullock, W. W. & Möller, G. Mechanism of thymus-independent immunocyte triggering. Mitogenic activation of B cells results in specific immune responses. *J. Exp. Med.* **139**, 74–92 (1974).

73. Bekeredjian-Ding, I. & Jego, G. Toll-like receptors--sentries in the B-cell response. *Immunology* **128**, 311–323 (2009).

74. Mond, J. J., Lees, A. & Snapper, C. M. T cell-independent antigens type 2. *Annu. Rev. Immunol.* **13**, 655–692 (1995).

75. García de Vinuesa, C., O'Leary, P., Sze, D. M., Toellner, K. M. & MacLennan, I. C. T-independent type 2 antigens induce B cell proliferation in multiple splenic sites, but exponential growth is confined to extrafollicular foci. *Eur. J. Immunol.* **29**, 1314–1323 (1999).

76. Kageyama, Y. & Katayama, N. Ontogeny of human B1 cells. *Int. J. Hematol.* **111**, 628–633 (2020).

77. Kantor, A. B., Stall, A. M., Adams, S., Watanabe, K. & Herzenberg, L. A. De novo development and self-replenishment of B cells. *Int. Immunol.* **7**, 55–68 (1995).

78. Hayakawa, K., Hardy, R. R. & Herzenberg, L. A. Peritoneal Ly-1 B cells: genetic control, autoantibody production, increased lambda light chain expression. *Eur. J. Immunol.* **16**, 450–456 (1986).

79. Chou, M.-Y. *et al.* Oxidation-specific epitopes are dominant targets of innate natural antibodies in mice and humans. *J. Clin. Invest.* **119**, 1335–1349 (2009).

80. Kulik, L. *et al.* Pathogenic natural antibodies recognizing annexin IV are required to develop intestinal ischemia-reperfusion injury. *J. Immunol.* **182**, 5363–5373 (2009).

81. Briles, D. E. *et al.* Antiphosphocholine antibodies found in normal mouse serum are protective against intravenous infection with type 3 streptococcus pneumoniae. *J. Exp. Med.* **153**, 694–705 (1981).

82. Leadbetter, E. A. *et al.* NK T cells provide lipid antigen-specific cognate help for B cells. *Proc. Natl. Acad. Sci. U. S. A.* **105**, 8339–8344 (2008).

83. Attanavanich, K. & Kearney, J. F. Marginal Zone, but Not Follicular B Cells, Are Potent Activators of Naive CD4 T Cells. *J. Immunol.* **172**, 803–811 (2004).

84. Cinamon, G. *et al.* Sphingosine 1-phosphate receptor 1 promotes B cell localization in the splenic marginal zone. *Nat. Immunol.* **5**, 713–720 (2004).

85. Bryant, V. L. *et al.* Cytokine-mediated regulation of human B cell differentiation into Ig-secreting cells: predominant role of IL-21 produced by CXCR5+ T follicular helper cells. *J. Immunol.* **179**, 8180–8190 (2007).

86. Cannons, J. L. *et al.* Optimal germinal center responses require a multistage T cell:B cell adhesion process involving integrins, SLAM-associated protein, and CD84. *Immunity* **32**, 253–265 (2010).

87. Huang, C. Germinal Center Reaction. *Adv. Exp. Med. Biol.* **1254**, 47–53 (2020).

88. Wan, Z., Lin, Y., Zhao, Y. & Qi, H. TFH cells in bystander and cognate interactions with B cells. *Immunol. Rev.* **288**, 28–36 (2019).

89. De Silva, N. S. & Klein, U. Dynamics of B cells in germinal centres. *Nat. Rev. Immunol.* **15**, 137–148 (2015).

90. Mesin, L., Ersching, J. & Victora, G. D. Germinal Center B Cell Dynamics. *Immunity* **45**, 471–482 (2016).

91. Victora, G. D. & Nussenzweig, M. C. Germinal centers. *Annu. Rev. Immunol.* **30**, 429–457 (2012).

92. Wacleche, V. S., Wang, R. & Rao, D. A. Identification of T Peripheral Helper (Tph) Cells. *Methods Mol. Biol.* **2380**, 59–76 (2022).

93. Rao, D. A. *et al.* Pathologically expanded peripheral T helper cell subset drives B cells in rheumatoid arthritis. *Nature* **542**, 110–114 (2017).

94. Bocharnikov, A. V. *et al.* PD-1hiCXCR5- T peripheral helper cells promote B cell responses in lupus via MAF and IL-21. *JCI Insight* **4**, (2019).

95. Ekman, I. *et al.* Circulating CXCR5−PD-1hi peripheral T helper cells are associated with progression to type 1 diabetes. *Diabetologia* **62**, 1681–1688 (2019).

96. Rao, D. A. The rise of peripheral T helper cells in autoimmune disease. *Nat. Rev. Rheumatol.* **15**, 453–454 (2019).

97. Christophersen, A. *et al.* Distinct phenotype of CD4+ T cells driving celiac disease identified in multiple autoimmune conditions. *Nat. Med.* **25**, 734–737 (2019).

98. Price, M. J., Scharer, C. D., Kania, A. K., Randall, T. D. & Boss, J. M. Conserved Epigenetic Programming and Enhanced Heme Metabolism Drive Memory B Cell Reactivation. *J. Immunol.* **206**, 1493 LP – 1504 (2021).

99. Mamani-Matsuda, M. *et al.* The human spleen is a major reservoir for long-lived vaccinia virus-specific memory B cells. *Blood* **111**, 4653–4659 (2008).

100. Charles, E. D. *et al.* Clonal B cells in patients with hepatitis C virus–associated mixed cryoglobulinemia contain an expanded anergic CD21low B-cell subset. *Blood* **117**, 5425–5437 (2011).

101. Moir, S. *et al.* Evidence for HIV-associated B cell exhaustion in a dysfunctional memory B cell compartment in HIV-infected viremic individuals. *J. Exp. Med.* **205**, 1797–1805 (2008).

102. Weiss, G. E. *et al.* Atypical Memory B Cells Are Greatly Expanded in Individuals Living in a Malaria-Endemic Area. *J. Immunol.* **183**, 2176–2182 (2009).

103. Isnardi, I. *et al.* Complement receptor 2/CD21− human naive B cells contain mostly autoreactive unresponsive clones. *Blood* **115**, 5026–5036 (2010).

104. Wei, C. *et al.* A New Population of Cells Lacking Expression of CD27 Represents a Notable Component of the B Cell Memory Compartment in Systemic Lupus Erythematosus. *J. Immunol.* **178**, 6624–6633 (2007).

105. Claes, N. *et al.* Age-Associated B Cells with Proinflammatory Characteristics Are Expanded in a Proportion of Multiple Sclerosis Patients. *J. Immunol.* **197**, 4576–4583 (2016).

106. Saadoun, D. *et al.* Expansion of autoreactive unresponsive CD21-/low B cells in sjögren's syndrome-associated lymphoproliferation. *Arthritis Rheum.* **65**, 1085–1096 (2013).

107. Wilbrink, R. *et al.* CD27-CD38lowCD21low B-Cells Are Increased in Axial Spondyloarthritis. *Front. Immunol.* **12**, 2148 (2021).

108. Portugal, S. *et al.* Malaria-associated atypical memory B cells exhibit markedly reduced B cell receptor signaling and effector function. *Elife* **4**, (2015).

109. Sullivan, R. T. *et al.* FCRL5 Delineates Functionally Impaired Memory B Cells Associated with Plasmodium falciparum Exposure. *PLOS Pathog.* **11**, e1004894 (2015).

110. Jenks, S. A. *et al.* Distinct Effector B Cells Induced by Unregulated Toll-like Receptor 7 Contribute to Pathogenic Responses in Systemic Lupus Erythematosus. *Immunity* **49**, 725-739.e6 (2018).

111. Muellenbeck, M. F. *et al.* Atypical and classical memory B cells produce Plasmodium

falciparum neutralizing antibodies. *J. Exp. Med.* **210**, 389–399 (2013).

112. Lau, D. *et al.* Low CD21 expression defines a population of recent germinal center graduates primed for plasma cell differentiation. *Sci. Immunol.* **2**, (2017).

113. Andrews, S. F. *et al.* Activation Dynamics and Immunoglobulin Evolution of Pre-existing and Newly Generated Human Memory B cell Responses to Influenza Hemagglutinin. *Immunity* **51**, 398-410.e5 (2019).

114. Knox, J. J. *et al.* T-bet+ B cells are induced by human viral infections and dominate the HIV gp140 response. *JCI insight* **2**, (2017).

115. Mahmood, Z. *et al.* CD27-IgD- memory B cells are modulated by in vivo interleukin-6 receptor (IL-6R) blockade in rheumatoid arthritis. *Arthritis Res. Ther.* **17**, 61 (2015).

116. Rincón-Arévalo, H. *et al.* Atypical phenotype and response of B cells in patients with seropositive rheumatoid arthritis. *Clin. Exp. Immunol.* **204**, 221–238 (2021).

117. Rivera-Correa, J. *et al.* Atypical memory B-cells are associated with Plasmodium falciparum anemia through anti-phosphatidylserine antibodies. *Elife* **8**, (2019).

118. Sachinidis, A., Xanthopoulos, K. & Garyfallos, A. Age-Associated B Cells (ABCs) in the Prognosis, Diagnosis and Therapy of Systemic Lupus Erythematosus (SLE). *Mediterr. J. Rheumatol.* **31**, 311–318 (2020).

119. Rubtsov, A. V. *et al.* Toll-like receptor 7 (TLR7)-driven accumulation of a novel $CD11c^{+}$ B-cell population is important for the development of autoimmunity. *Blood* **118**, 1305–1315 (2011).

120. Bao, W., Xie, M. & Ye, Y. Age-associated B cells indicate disease activity in rheumatoid arthritis. *Cell. Immunol.* **377**, (2022).

121. Zhang, F. *et al.* Defining Inflammatory Cell States in Rheumatoid Arthritis Joint Synovial Tissues by Integrating Single-cell Transcriptomics and Mass Cytometry. *Accel. Med. Partnersh. RA Phase* **13**, (2018).

122. Hao, Y., O'Neill, P., Naradikian, M. S., Scholz, J. L. & Cancro, M. P. A B-cell subset uniquely responsive to innate stimuli accumulates in aged mice. *Blood* **118**, 1294–1304 (2011).

123. Qin, Y. *et al.* Age-associated B cells contribute to the pathogenesis of rheumatoid arthritis by inducing activation of fibroblast-like synoviocytes via TNF-α-mediated ERK1/2 and JAK-STAT1 pathways. *Ann. Rheum. Dis.* (2022). doi:10.1136/ARD-2022-222605

124. Mouat, I. C., Goldberg, E. & Horwitz, M. S. Age-associated B cells in autoimmune diseases. *Cell. Mol. Life Sci.* **79**, 1–11 (2022).

125. Lam, J. H. & Baumgarth, N. The Multifaceted B Cell Response to Influenza Virus. *J. Immunol.* **202**, 351–359 (2019).

126. Palm, A. K. E. & Henry, C. Remembrance of Things Past: Long-Term B Cell Memory After Infection and Vaccination. *Front. Immunol.* **10**, 1787 (2019).

127. Lee, F. E.-H., Falsey, A. R., Halliley, J. L., Sanz, I. & Walsh, E. E. Circulating antibody-secreting cells during acute respiratory syncytial virus infection in adults. *J. Infect. Dis.* **202**, 1659–1666 (2010).

128. Wrammert, J. *et al.* Rapid and massive virus-specific plasmablast responses during acute dengue virus infection in humans. *J. Virol.* **86**, 2911–2918 (2012).

129. Hargreaves, D. C. *et al.* A coordinated change in chemokine responsiveness guides plasma cell movements. *J. Exp. Med.* **194**, 45–56 (2001).

130. Good-Jacobson, K. L., O'Donnell, K., Belz, G. T., Nutt, S. L. & Tarlinton, D. M. c-Myb is required for plasma cell migration to bone marrow after immunization or infection. *J. Exp. Med.* **212**, 1001–1009 (2015).

131. Biajoux, V. *et al.* Efficient Plasma Cell Differentiation and Trafficking Require Cxcr4 Desensitization. *Cell Rep.* **17**, 193–205 (2016).

132. Watson, C. T. *et al.* Complete Haplotype Sequence of the Human Immunoglobulin Heavy-Chain Variable, Diversity, and Joining Genes and Characterization of Allelic and Copy-Number Variation. *Am. J. Hum. Genet.* **92**, 530–546 (2013).

133. Lefranc, M.-P. Nomenclature of the Human Immunoglobulin Genes. in *Current Protocols in Immunology* **40**, A.1P.1--A.1P.37 (John Wiley & Sons, Inc., 2001).

134. Williams, S. C. *et al.* Sequence and Evolution of the Human Germline V$λ$Repertoire. *J. Mol. Biol.* **264**, 220–232 (1996).

135. Giudicelli, V. *et al.* IMGT, the international ImMunoGeneTics database. *Nucleic Acids Res.* **25**, 206–211 (1997).

136. Wardemann, H. *et al.* Predominant autoantibody production by early human B cell precursors. *Science* **301**, 1374–1377 (2003).

137. Watanabe, A. *et al.* Self-tolerance curtails the B cell repertoire to microbial epitopes. *JCI insight* **4**, (2019).

138. Harwood, N. E. & Batista, F. D. Early Events in B Cell Activation. *http://dx.doi.org.proxy.kib.ki.se/10.1146/annurev-immunol-030409-101216* **28**, 185–210 (2010).

139. Bouillet, P., Metcalf, D. & Huang, D. C. S. Proapoptotic Bcl-2 relative Bim required for certain apoptotic responses, leukocyte homeostasis, and to preclude autoimmunity. *Science (80-.).* **286**, 1735–1738 (1999).

140. Enders, A. *et al.* Loss of the Pro-Apoptotic BH3-only Bcl-2 Family Member Bim Inhibits BCR Stimulation–induced Apoptosis and Deletion of Autoreactive B Cells. *J. Exp. Med.* **198**, 1119–1126 (2003).

141. Brink, R. & Phan, T. G. Self-Reactive B Cells in the Germinal Center Reaction. *https://doi-org.proxy.kib.ki.se/10.1146/annurev-immunol-051116-052510* **36**, 339–357 (2018).

142. Chan, T. D. *et al.* Elimination of Germinal-Center-Derived Self-Reactive B Cells Is Governed by the Location and Concentration of Self-Antigen. *Immunity* **37**, 893–904 (2012).

143. Brink, R. The imperfect control of self-reactive germinal center B cells. *Curr. Opin. Immunol.* **28**, 97–101 (2014).

144. Lee, S., Ko, Y. & Kim, T. J. Homeostasis and regulation of autoreactive B cells. *Cell. Mol. Immunol.* **17**, 561–569 (2020).

145. Pascual, V. *et al.* VH restriction among human cold agglutinins. The VH4-21 gene segment is required to encode anti-I and anti-i specificities. *J. Immunol.* **149**, 2337–2344 (1992).

146. Samuels, J., Ng, Y.-S., Coupillaud, C., Paget, D. & Meffre, E. Impaired early B cell tolerance in patients with rheumatoid arthritis. *J. Exp. Med.* **201**, 1659–1667 (2005).

147. Baka, Z. *et al.* Citrullination under physiological and pathological conditions. *Jt. bone spine* **79**, 431–436 (2012).

148. Wang, S. & Wang, Y. Peptidylarginine deiminases in citrullination, gene regulation, health and pathogenesis. *Biochim. Biophys. Acta* **1829**, 1126–1135 (2013).

149. Rosenstein, E. D., Greenwald, R. A., Kushner, L. J. & Weissmann, G. Hypothesis: the humoral immune response to oral bacteria provides a stimulus for the development of rheumatoid arthritis. *Inflammation* **28**, 311–318 (2004).

150. Masson-Bessière, C. *et al.* The major synovial targets of the rheumatoid arthritis-specific antifilaggrin autoantibodies are deiminated forms of the alpha- and beta-chains of fibrin. *J. Immunol.* **166**, 4177–4184 (2001).

151. Vossenaar, E. R., Zendman, A. J. W. & Van Venrooij, W. J. Citrullination, a possible functional link between susceptibility genes and rheumatoid arthritis. *Arthritis research & therapy* **6**, 1–5 (2004).

152. Burkhardt, H. *et al.* Humoral immune response to citrullinated collagen type II determinants in early rheumatoid arthritis. *Eur. J. Immunol.* **35**, 1643–1652 (2005).

153. Kinloch, A. *et al.* Identification of citrullinated alpha-enolase as a candidate autoantigen in rheumatoid arthritis. *Arthritis Res. Ther.* **7**, R1421-9 (2005).

154. Pratesi, F. *et al.* Antibodies from patients with rheumatoid arthritis target citrullinated histone 4 contained in neutrophils extracellular traps. *Ann. Rheum. Dis.* **73**, 1414–1422 (2014).

155. Steen, J. *et al.* Recognition of Amino Acid Motifs, Rather Than Specific Proteins, by Human Plasma Cell–Derived Monoclonal Antibodies to Posttranslationally Modified Proteins in Rheumatoid Arthritis. *Arthritis Rheumatol.* **71**, 196–209 (2019).

156. Pruijn, G. J., Wiik, A. & van Venrooij, W. J. The use of citrullinated peptides and proteins for the diagnosis of rheumatoid arthritis. *Arthritis Res. Ther.* **12**, 203 (2010).

157. van Gaalen, F. A. *et al.* Autoantibodies to cyclic citrullinated peptides predict progression to rheumatoid arthritis in patients with undifferentiated arthritis: a prospective cohort study. *Arthritis Rheum.* **50**, 709–715 (2004).

158. Meyer, O. *et al.* Anticitrullinated protein/peptide antibody assays in early rheumatoid arthritis for predicting five year radiographic damage. *Ann. Rheum. Dis.* **62**, 120–126 (2003).

159. van der Linden, M. P. M. *et al.* Value of anti-modified citrullinated vimentin and third-generation anti-cyclic citrullinated peptide compared with second-generation anti-cyclic citrullinated peptide and rheumatoid factor in predicting disease outcome in undifferentiated arthritis and r. *Arthritis Rheum.* **60**, 2232–2241 (2009).

160. Jørgensen, K. T. *et al.* Cytokines, autoantibodies and viral antibodies in premorbid and

postdiagnostic sera from patients with rheumatoid arthritis: case-control study nested in a cohort of Norwegian blood donors. *Ann. Rheum. Dis.* **67**, 860–866 (2008).

161. Demoruelle, M. K. *et al.* Performance of anti-cyclic citrullinated Peptide assays differs in subjects at increased risk of rheumatoid arthritis and subjects with established disease. *Arthritis Rheum.* **65**, 2243–2252 (2013).

162. Goeldner, I. *et al.* Anti-cyclic citrullinated peptide antibodies and rheumatoid factor in rheumatoid arthritis patients and relatives from Brazil. *Rheumatology (Oxford).* **49**, 1590–1593 (2010).

163. van Zanten, A. *et al.* Presence of anticitrullinated protein antibodies in a large population-based cohort from the Netherlands. *Ann. Rheum. Dis.* **76**, 1184–1190 (2017).

164. Terao, C. *et al.* Effects of smoking and shared epitope on the production of anti-citrullinated peptide antibody in a Japanese adult population. *Arthritis Care Res. (Hoboken).* **66**, 1818–1827 (2014).

165. Tasliyurt, T. *et al.* The frequency of antibodies against cyclic citrullinated peptides and rheumatoid factor in healthy population: a field study of rheumatoid arthritis from northern Turkey. *Rheumatol. Int.* **33**, 939–942 (2013).

166. van Steenbergen, H. W., Mangnus, L., Reijnierse, M., Huizinga, T. W. J. & van der Helm-van Mil, A. H. M. Clinical factors, anticitrullinated peptide antibodies and MRI-detected subclinical inflammation in relation to progression from clinically suspect arthralgia to arthritis. *Ann. Rheum. Dis.* **75**, 1824–1830 (2016).

167. Nam, J. L. *et al.* Ultrasound findings predict progression to inflammatory arthritis in anti-CCP antibody-positive patients without clinical synovitis. *Ann. Rheum. Dis.* **75**, 2060–2067 (2016).

168. Bos, W. H. *et al.* Arthritis development in patients with arthralgia is strongly associated with anti-citrullinated protein antibody status: a prospective cohort study. *Ann. Rheum. Dis.* **69**, 490–4 (2010).

169. Harre, U. *et al.* Induction of osteoclastogenesis and bone loss by human autoantibodies against citrullinated vimentin. *J. Clin. Invest.* **122**, 1791–1802 (2012).

170. Khandpur, R. *et al.* NETs are a source of citrullinated autoantigens and stimulate inflammatory responses in rheumatoid arthritis. *Sci. Transl. Med.* **5**, 178ra40 (2013).

171. Trouw, L. A. *et al.* Anti-cyclic citrullinated peptide antibodies from rheumatoid arthritis patients activate complement via both the classical and alternative pathways. *Arthritis Rheum.* **60**, 1923–1931 (2009).

172. Sokolove, J., Zhao, X., Chandra, P. E. & Robinson, W. H. Immune complexes containing citrullinated fibrinogen costimulate macrophages via Toll-like receptor 4 and Fcγ receptor. *Arthritis Rheum.* **63**, 53–62 (2011).

173. Suurmond, J. *et al.* Toll-like receptor triggering augments activation of human mast cells by anti-citrullinated protein antibodies. *Ann. Rheum. Dis.* **74**, 1915–1923 (2015).

174. Laurent, L. *et al.* Fcγ receptor profile of monocytes and macrophages from rheumatoid arthritis patients and their response to immune complexes formed with autoantibodies to citrullinated proteins. *Ann. Rheum. Dis.* **70**, 1052–1059 (2011).

175. Wigerblad, G. *et al.* Autoantibodies to citrullinated proteins induce joint pain independent of inflammation via a chemokine-dependent mechanism. *Ann. Rheum. Dis.* **75**, 730–8 (2016).

176. Krishnamurthy, A. *et al.* Identification of a novel chemokine-dependent molecular mechanism underlying rheumatoid arthritis-associated autoantibody-mediated bone loss. *Ann. Rheum. Dis.* **75**, 721–9 (2016).

177. van der Woude, D. *et al.* Epitope spreading of the anti-citrullinated protein antibody response occurs before disease onset and is associated with the disease course of early arthritis. *Ann. Rheum. Dis.* **69**, 1554–1561 (2010).

178. Verpoort, K. N. *et al.* Isotype distribution of anti-cyclic citrullinated peptide antibodies in undifferentiated arthritis and rheumatoid arthritis reflects an ongoing immune response. *Arthritis Rheum.* **54**, 3799–3808 (2006).

179. Willemze, A. *et al.* The ACPA recognition profile and subgrouping of ACPA-positive RA patients. *Ann. Rheum. Dis.* **71**, 268–274 (2012).

180. van de Stadt, L. A. *et al.* The extent of the anti-citrullinated protein antibody repertoire is associated with arthritis development in patients with seropositive arthralgia. *Ann. Rheum. Dis.* **70**, 128–133 (2011).

181. van de Stadt, L. A. *et al.* Development of the anti-citrullinated protein antibody repertoire prior to the onset of rheumatoid arthritis. *Arthritis Rheum.* **63**, 3226–3233 (2011).

182. Sokolove, J. *et al.* Autoantibody epitope spreading in the pre-clinical phase predicts progression to rheumatoid arthritis. *PLoS One* **7**, e35296 (2012).

183. Pelzek, A. J. *et al.* Persistence of Disease-Associated Anti–Citrullinated Protein Antibody–Expressing Memory B Cells in Rheumatoid Arthritis in Clinical Remission. *Arthritis Rheumatol.* **69**, 1176–1186 (2017).

184. Kerkman, P. F. *et al.* Circulating plasmablasts/plasmacells as a source of anticitrullinated protein antibodies in patients with rheumatoid arthritis. *Ann. Rheum. Dis.* **72**, 1259–1263 (2013).

185. Kerkman, P. F. *et al.* Identification and characterisation of citrullinated antigen-specific B cells in peripheral blood of patients with rheumatoid arthritis. *Ann. Rheum. Dis.* **75**, 1170–1176 (2016).

186. Kerkman, P. F. *et al.* Synovial fluid mononuclear cells provide an environment for long-term survival of antibody-secreting cells and promote the spontaneous production of anti-citrullinated protein antibodies. *Ann. Rheum. Dis.* **75**, 2201–2207 (2016).

187. Scherer, H. U. *et al.* Immunoglobulin 1 (IgG1) Fc-glycosylation profiling of anti-citrullinated peptide antibodies from human serum. *Proteomics. Clin. Appl.* **3**, 106–115 (2009).

188. Lloyd, K. A. *et al.* Variable domain N-linked glycosylation and negative surface charge are key features of monoclonal ACPA: Implications for B-cell selection. *Eur. J. Immunol.* **48**, 1030–1045 (2018).

189. Rombouts, Y. *et al.* Extensive glycosylation of ACPA-IgG variable domains

modulates binding to citrullinated antigens in rheumatoid arthritis. *Ann. Rheum. Dis.* **75**, 578–585 (2016).

190. Hafkenscheid, L. *et al.* Structural Analysis of Variable Domain Glycosylation of Anti-Citrullinated Protein Antibodies in Rheumatoid Arthritis Reveals the Presence of Highly Sialylated Glycans. *Mol. Cell. Proteomics* **16**, 278–287 (2017).

191. Sahlström, P. *et al.* Different Hierarchies of Anti–Modified Protein Autoantibody Reactivities in Rheumatoid Arthritis. *Arthritis Rheumatol.* **72**, 1643–1657 (2020).

192. Vergroesen, R. D. *et al.* B-cell receptor sequencing of anti-citrullinated protein antibody (ACPA) IgG-expressing B cells indicates a selective advantage for the introduction of N-glycosylation sites during somatic hypermutation. *Ann. Rheum. Dis.* **77**, 956 LP – 958 (2018).

193. van de Bovenkamp, F. S., Hafkenscheid, L., Rispens, T. & Rombouts, Y. The Emerging Importance of IgG Fab Glycosylation in Immunity. *J. Immunol.* **196**, 1435–1441 (2016).

194. Kempers, A. C. *et al.* The extensive glycosylation of the ACPA variable domain observed for ACPA-IgG is absent from ACPA-IgM. *Annals of the rheumatic diseases* **77**, 1087–1088 (2018).

195. Suwannalai, P. *et al.* Anti-citrullinated protein antibodies have a low avidity compared with antibodies against recall antigens. *Ann. Rheum. Dis.* **70**, 373–379 (2011).

196. Suwannalai, P. *et al.* Avidity maturation of anti-citrullinated protein antibodies in rheumatoid arthritis. *Arthritis Rheum.* **64**, 1323–1328 (2012).

197. Raychaudhuri, S. *et al.* Five amino acids in three HLA proteins explain most of the association between MHC and seropositive rheumatoid arthritis. *Nat. Genet.* **44**, 291–296 (2012).

198. Scally, S. W. *et al.* A molecular basis for the association of the HLA-DRB1 locus, citrullination, and rheumatoid arthritis. *J. Exp. Med.* **210**, 2569–2582 (2013).

199. Pieper, J. *et al.* Memory T cells specific to citrullinated α-enolase are enriched in the rheumatic joint. *J. Autoimmun.* **92**, 47–56 (2018).

200. Gonzalez-Gay, M. A., Garcia-Porrua, C. & Hajeer, A. H. Influence of human leukocyte antigen-DRB1 on the susceptibility and severity of rheumatoid arthritis. *Seminars in arthritis and rheumatism* **31**, 355–360 (2002).

201. Weyand, C. M., Hicok, K. C., Conn, D. L. & Goronzy, J. J. The influence of HLA-DRB1 genes on disease severity in rheumatoid arthritis. *Ann. Intern. Med.* **117**, 801–806 (1992).

202. Mewar, D. *et al.* Association between radiographic severity of rheumatoid arthritis and shared epitope alleles: differing mechanisms of susceptibility and protection. *Ann. Rheum. Dis.* **67**, 980–983 (2008).

203. van Heemst, J. *et al.* Protective effect of HLA-DRB1*13 alleles during specific phases in the development of ACPA-positive RA. *Ann. Rheum. Dis.* **75**, 1891–1898 (2016).

204. van der Woude, D. *et al.* Protection against anti-citrullinated protein antibody-positive rheumatoid arthritis is predominantly associated with HLA-DRB1*1301: a meta-

analysis of HLA-DRB1 associations with anti-citrullinated protein antibody-positive and anti-citrullinated protein. *Arthritis Rheum.* **62**, 1236–1245 (2010).

205. Lundström, E. *et al.* Opposing effects of HLA-DRB1*13 alleles on the risk of developing anti-citrullinated protein antibody-positive and anti-citrullinated protein antibody-negative rheumatoid arthritis. *Arthritis Rheum.* **60**, 924–930 (2009).

206. Dekkers, J. S. *et al.* Breach of autoreactive B cell tolerance by post-translationally modified proteins. *Ann. Rheum. Dis.* **76**, 1449–1457 (2017).

207. van Heemst, J. *et al.* Crossreactivity to vinculin and microbes provides a molecular basis for HLA-based protection against rheumatoid arthritis. *Nat. Commun.* **6**, 6681 (2015).

208. Hand, T. W. *et al.* Acute gastrointestinal infection induces long-lived microbiota-specific T cell responses. *Science* **337**, 1553–1556 (2012).

209. Tsuda, R. *et al.* Monoclonal antibody against citrullinated peptides obtained from rheumatoid arthritis patients reacts with numerous citrullinated microbial and food proteins. *Arthritis Rheumatol. (Hoboken, N.J.)* **67**, 2020–2031 (2015).

210. James, E. A. *et al.* Citrulline-specific Th1 cells are increased in rheumatoid arthritis and their frequency is influenced by disease duration and therapy. *Arthritis Rheumatol. (Hoboken, N.J.)* **66**, 1712–1722 (2014).

211. Chemin, K. *et al.* A Novel HLA-DRB1*10:01-Restricted T Cell Epitope From Citrullinated Type II Collagen Relevant to Rheumatoid Arthritis. *Arthritis Rheumatol. (Hoboken, N.J.)* **68**, 1124–1135 (2016).

212. Cornillet, M., Verrouil, E., Cantagrel, A., Serre, G. & Nogueira, L. In ACPA-positive RA patients, antibodies to EBNA35-58Cit, a citrullinated peptide from the Epstein-Barr nuclear antigen-1, strongly cross-react with the peptide β60-74Cit which bears the immunodominant epitope of citrullinated fibrin. *Immunol. Res.* **61**, 117–125 (2015).

213. Lundberg, K. *et al.* Antibodies to citrullinated alpha-enolase peptide 1 are specific for rheumatoid arthritis and cross-react with bacterial enolase. *Arthritis Rheum.* **58**, 3009–3019 (2008).

214. Li, S. *et al.* Autoantibodies from Single Circulating Plasmablasts React with Citrullinated Antigens and Porphyromonas gingivalis in Rheumatoid Arthritis. *Arthritis Rheumatol.* **68**, 614–626 (2016).

215. Trouw, L. A., Rispens, T. & Toes, R. E. M. Beyond citrullination: other post-translational protein modifications in rheumatoid arthritis. *Nat. Rev. Rheumatol.* **13**, 331–339 (2017).

216. Eberwine, J. *et al.* Analysis of gene expression in single live neurons. *Proc. Natl. Acad. Sci. U. S. A.* **89**, 3010–4 (1992).

217. Tyagi, S. & Kramer, F. R. Molecular Beacons: Probes that Fluoresce upon Hybridization. *Nat. Biotechnol.* **14**, 303–308 (1996).

218. Kamme, F. *et al.* Single-Cell Microarray Analysis in Hippocampus CA1: Demonstration and Validation of Cellular Heterogeneity. *J. Neurosci.* **23**, 3607–3615 (2003).

219. Tang, F. *et al.* mRNA-Seq whole-transcriptome analysis of a single cell. *Nat. Methods* **6**, 377–382 (2009).

220. Deng, Q., Ramsköld, D., Reinius, B. & Sandberg, R. Single-cell RNA-seq reveals dynamic, random monoallelic gene expression in mammalian cells. *Science* **343**, 193–6 (2014).

221. Patel, A. P. *et al.* Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science (80-.).* **344**, 1396–1401 (2014).

222. Orange, D. E. *et al.* Identification of Three Rheumatoid Arthritis Disease Subtypes by Machine Learning Integration of Synovial Histologic Features and RNA Sequencing Data. *Arthritis Rheumatol.* **70**, 690–701 (2018).

223. Zhang, F. *et al.* Defining inflammatory cell states in rheumatoid arthritis joint synovial tissues by integrating single-cell transcriptomics and mass cytometry. *Nat. Immunol.* **20**, 928–942 (2019).

224. Zhang, F. *et al.* Cellular deconstruction of inflamed synovium defines diverse inflammatory phenotypes in rheumatoid arthritis. *bioRxiv* 2022.02.25.481990 (2022). doi:10.1101/2022.02.25.481990

225. Stubbington, M. J. T. *et al.* T cell fate and clonality inference from single-cell transcriptomes. *Nat. Methods* **13**, 329–332 (2016).

226. Lindeman, I. *et al.* BraCeR: B-cell-receptor reconstruction and clonality inference from single-cell RNA-seq. *Nat. Methods* **15**, 563–565 (2018).

227. Roberts, A. & Pachter, L. Streaming fragment assignment for real-time analysis of sequencing experiments. *Nat. Methods* **10**, 71–73 (2013).

228. Soneson, C., Love, M. I. & Robinson, M. D. Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. *F1000Research* **4**, 1521 (2015).

229. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**, 10 (2011).

230. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).

231. Grabherr, M. G. *et al.* Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* **29**, 644–652 (2011).

232. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).

233. Ye, J., Ma, N., Madden, T. L. & Ostell, J. M. IgBLAST: an immunoglobulin variable domain sequence analysis tool. *Nucleic Acids Res.* **41**, W34--W40 (2013).

234. Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.* **34**, 525–527 (2016).

235. Matsuda, F. *et al.* The complete nucleotide sequence of the human immunoglobulin heavy chain variable region locus. *J. Exp. Med.* **188**, 2151–62 (1998).

236. Rodriguez, O. L. *et al.* A Novel Framework for Characterizing Genomic Haplotype

Diversity in the Human Immunoglobulin Heavy Chain Locus. *Front. Immunol.* **11**, 2136 (2020).

237. Rodriguez, O. L. *et al.* Antibody repertoire gene usage is explained by common genetic variants in the immunoglobulin heavy chain locus. *bioRxiv* 2022.07.04.498729 (2022). doi:10.1101/2022.07.04.498729

238. Tsai, F.-J. *et al.* Identification of novel susceptibility Loci for kawasaki disease in a Han chinese population by a genome-wide association study. *PLoS One* **6**, e16853 (2011).

239. Georgiou, G. *et al.* The promise and challenge of high-throughput sequencing of the antibody repertoire. *Nat. Biotechnol.* **32**, 158–168 (2014).

240. Corcoran, M. M. *et al.* Production of individualized V gene databases reveals high levels of immunoglobulin genetic diversity. *Nat. Commun.* **7**, 13642 (2016).

241. Gadala-Maria, D., Yaari, G., Uduman, M. & Kleinstein, S. H. Automated analysis of high-throughput B-cell sequencing data reveals a high frequency of novel immunoglobulin V gene segment alleles. *Proc. Natl. Acad. Sci. U. S. A.* **112**, E862-70 (2015).

242. Gadala-Maria, D. *et al.* Identification of subject-specific immunoglobulin alleles from expressed repertoire sequencing data. *Front. Immunol.* **10**, 129 (2019).

243. Mikocziova, I. *et al.* Polymorphisms in human immunoglobulin heavy chain variable genes and their upstream regions. *Nucleic Acids Res.* **48**, 5499–5510 (2020).

244. Gidoni, M. *et al.* Mosaic deletion patterns of the human antibody heavy chain gene locus shown by Bayesian haplotyping. *Nat. Commun.* **10**, 628 (2019).

245. Romo-González, T., Morales-Montor, J., Rodríguez-Dorantes, M. & Vargas-Madrazo, E. Novel Substitution Polymorphisms of Human Immunoglobulin VH Genes in Mexicans. *Hum. Immunol.* **66**, 731–739 (2005).

246. Scheepers, C. *et al.* Ability To Develop Broadly Neutralizing HIV-1 Antibodies Is Not Restricted by the Germline Ig Gene Repertoire. *J. Immunol.* **194**, 4371 LP – 4378 (2015).

247. Boyd, S. D. *et al.* Individual Variation in the Germline Ig Gene Repertoire Inferred from Variable Region Gene Rearrangements. *J. Immunol.* **184**, 6986 LP – 6992 (2010).

248. Peng, K. *et al.* Diversity in immunogenomics: the value and the challenge. *Nat. Methods 2021 186* **18**, 588–591 (2021).

249. Ledergor, G. *et al.* Single cell dissection of plasma cell heterogeneity in symptomatic and asymptomatic myeloma. *Nat. Med.* **24**, 1867–1876 (2018).

250. Goldstein, L. D. *et al.* Massively parallel single-cell B-cell receptor sequencing enables rapid discovery of diverse antigen-reactive antibodies. *Commun. Biol. 2019 21* **2**, 1–10 (2019).

251. Doorenspleet, M. E. *et al.* Rheumatoid arthritis synovial tissue harbours dominant B-cell and plasma-cell clones associated with autoreactivity. *Ann. Rheum. Dis.* **73**, 756–762 (2014).

252. Vencovsk, J. *et al.* Polymorphism in the immunoglobulin VH gene V1-69 affects susceptibility to rheumatoid arthritis in subjects lacking the HLA-DRB1 shared epitope. *Rheumatology (Oxford).* **41**, 401–410 (2002).

253. Walter, M. A., Gibson, W. T., Ebers, G. C. & Cox, D. W. Susceptibility to multiple sclerosis is associated with the proximal immunoglobulin heavy chain variable region. *J. Clin. Invest.* **87**, 1266–1273 (1991).

254. Hashimoto, L. L., Walter, M. A., Cox, D. W. & Ebers, G. C. Immunoglobulin heavy chain variable region polymorphisms and multiple sclerosis susceptibility. *J. Neuroimmunol.* **44**, 77–83 (1993).

255. Wood, N. W. *et al.* Susceptibility to multiple sclerosis and the immunoglobulin heavy chain variable region. *J. Neurol.* **242**, 677–682 (1995).

256. Olee, T. *et al.* Molecular basis of an autoantibody-associated restriction fragment length polymorphism that confers susceptibility to autoimmune diseases. *J. Clin. Invest.* **88**, 193–203 (1991).

257. Cho, M.-L. *et al.* Association of homozygous deletion of the Humhv3005 and the VH3-30.3 genes with renal involvement in systemic lupus erythematosus. *Lupus* **12**, 400–405 (2003).

258. Huang, D. F. *et al.* Population and family studies of three disease-related polymorphic genes in systemic lupus erythematosus. *J. Clin. Invest.* **95**, 1766–1772 (1995).

259. Mo, L. *et al.* The frequency of homozygous deletion of a developmentally regulated Vh gene (Humhv3005) is increased in patients with chronic idiopathic thrombocytopenic purpura. *Autoimmunity* **24**, 257–263 (1996).

260. Veijola R, Knip M, Puukka R, Reijonen H, Cox DW, I. J. The immunoglobulin heavy-chain variable region in insulin-dependent diabetes mellitus: affected-sib-pair analysis and association studies - PubMed. *Am. J. Hum. Genet.* **59**, 462–70 (1996).

261. Parks, T. *et al.* Association between a common immunoglobulin heavy chain allele and rheumatic heart disease risk in Oceania. *Nat. Commun.* (2017). doi:10.1038/ncomms14946

262. Nadel, B. *et al.* Decreased frequency of rearrangement due to the synergistic effect of nucleotide changes in the heptamer and nonamer of the recombination signal sequence of the V kappa gene A2b, which is associated with increased susceptibility of Navajos to Haemophilus. *J. Immunol.* **161**, 6068–6073 (1998).

263. Feeney, A. J., Atkinson, M. J., Cowan, M. J., Escuro, G. & Lugo, G. A defective Vkappa A2 allele in Navajos which may play a role in increased susceptibility to haemophilus influenzae type b disease. *J. Clin. Invest.* **97**, 2277–2282 (1996).

264. Avnir, Y. *et al.* IGHV1-69 polymorphism modulates anti-influenza antibody repertoires, correlates with IGHV utilization shifts and varies by ethnicity. *Sci. Rep.* **6**, 20842 (2016).

265. Avnir, Y. *et al.* Molecular Signatures of Hemagglutinin Stem-Directed Heterosubtypic Human Neutralizing Antibodies against Influenza A Viruses. *PLOS Pathog.* **10**, e1004103 (2014).

266. Corti, D. *et al.* A neutralizing antibody selected from plasma cells that binds to group 1

and group 2 influenza A hemagglutinins. *Science (80-.).* **333**, 850–856 (2011).

267. Wyrzucki, A. *et al.* Alternative recognition of the conserved stem epitope in influenza A virus hemagglutinin by a VH3-30-encoded heterosubtypic antibody. *J. Virol.* **88**, 7083–7092 (2014).

268. Huang, J. *et al.* Identification of a CD4-Binding-Site Antibody to HIV that Evolved Near-Pan Neutralization Breadth. *Immunity* **45**, 1108–1121 (2016).

269. Wu, X. *et al.* Rational design of envelope identifies broadly neutralizing human monoclonal antibodies to HIV-1. *Science (80-.).* **329**, 856–861 (2010).

270. Yuan, M., Liu, H., Wu, N. C. & Wilson, I. A. Recognition of the SARS-CoV-2 receptor binding domain by neutralizing antibodies. *Biochem. Biophys. Res. Commun.* **538**, 192–203 (2021).

271. Yuan, M. *et al.* Structural basis of a shared antibody response to SARS-CoV-2. *Science (80-.).* **369**, 1119–1123 (2020).

272. Fagiani, F., Catanzaro, M. & Lanni, C. Molecular features of IGHV3-53-encoded antibodies elicited by SARS-CoV-2. *Signal Transduct. Target. Ther. 2020 51* **5**, 1–3 (2020).

273. Davis, C. W. *et al.* Longitudinal Analysis of the Human B Cell Response to Ebola Virus Infection. *Cell* **177**, 1566-1582.e17 (2019).

274. Robbiani, D. F. *et al.* Recurrent Potent Human Neutralizing Antibodies to Zika Virus in Brazil and Mexico. *Cell* **169**, 597-609.e11 (2017).

275. Imkeller, K. *et al.* Antihomotypic affinity maturation improves human B cell responses against a repetitive epitope. *Science* **360**, 1358–1362 (2018).

276. Murugan, R. *et al.* Evolution of protective human antibodies against Plasmodium falciparum circumsporozoite protein repeat motifs. *Nat. Med. 2020 267* **26**, 1135–1145 (2020).

277. Floudas, A. *et al.* Inside the Joint of Inflammatory Arthritis Patients: Handling and Processing of Synovial Tissue Biopsies for High Throughput Analysis. *Front. Med.* **9**, 441 (2022).

278. Floudas, A. *et al.* Pathogenic, glycolytic PD-1+ B cells accumulate in the hypoxic RA joint. *JCI Insight* **5**, (2020).

279. Donlin, L. T. *et al.* Methods for high-dimensonal analysis of cells dissociated from cyropreserved synovial tissue. *Arthritis Res. Ther.* **20**, 1–15 (2018).

280. Jiang, L. *et al.* Synthetic spike-in standards for RNA-seq experiments. *Genome Res.* **21**, 1543–1551 (2011).

281. Wu, X. *et al.* Single-cell sequencing of immune cells from anticitrullinated peptide antibody positive and negative rheumatoid arthritis. *Nat. Commun. 2021 121* **12**, 1–15 (2021).

282. Scheel, T., Gursche, A., Zacher, J. & Berek, C. V-region gene analysis of locally defined synovial B and plasma cells reveals selected B cell expansion and accumulation of plasma cell clones in rheumatoid arthritis. *Arthritis Rheum.* **63**, 63–72 (2011).

283. Stephenson, W. *et al.* Single-cell RNA-seq of rheumatoid arthritis synovial tissue using low-cost microfluidic instrumentation. *Nat. Commun.* **9**, 791 (2018).

284. Lu, D. R. *et al.* T Cell-Dependent Affinity Maturation and Innate Immune Pathways Differentially Drive Autoreactive B Cell Responses in Rheumatoid Arthritis. *Arthritis Rheumatol. (Hoboken, N.J.)* **70**, 1732–1744 (2018).

285. Chatzidionysiou, K. *et al.* Highest clinical effectiveness of rituximab in autoantibody-positive patients with rheumatoid arthritis and in those for whom no more than one previous TNF antagonist has failed: Pooled data from 10 European registries. *Ann. Rheum. Dis.* **70**, 1575–1580 (2011).

286. Takahashi, N., Kojima, T., Asai, S., Terabe, K. & Ishiguro, N. FRI0107 Effectiveness of Abatacept on Clinical Disease Activity and Radiographic Progression in Rheumatoid Arthritis Patients in Daily Clinical Practice in Japan: Comparisons According to ACPA Status. *Ann. Rheum. Dis.* **79**, 633–634 (2020).

287. Wang, S. *et al.* IL-21 drives expansion and plasma cell differentiation of autoreactive CD11c(hi)T-bet(+) B cells in SLE. *Nat. Commun.* **9**, 1758 (2018).

288. Khodadadi, L., Cheng, Q., Radbruch, A. & Hiepe, F. The Maintenance of Memory Plasma Cells. *Front. Immunol.* **10**, 721–721 (2019).

289. Pitzalis, C., Kelly, S. & Humby, F. New learnings on the pathophysiology of RA from synovial biopsies. *Curr. Opin. Rheumatol.* **25**, 334–344 (2013).

290. Kempers, A. C., Hafkenscheid, L., Scherer, H. U. & Toes, R. E. M. Variable domain glycosylation of ACPA-IgG: A missing link in the maturation of the ACPA response? *Clin. Immunol.* **186**, 34–37 (2018).

291. van de Stadt, L. A. *et al.* Monoclonal anti-citrullinated protein antibodies selected on citrullinated fibrinogen have distinct targets with different cross-reactivity patterns. *Rheumatol. (United Kingdom)* **52**, 631–635 (2013).

292. Bang, H. *et al.* Mutation and citrullination modifies vimentin to a novel autoantigen for rheumatoid arthritis. *Arthritis Rheum.* **56**, 2503–2511 (2007).

293. Germar, K. *et al.* Generation and Characterization of Anti–Citrullinated Protein Antibody–Producing B Cell Clones From Rheumatoid Arthritis Patients. *Arthritis Rheumatol.* **71**, 340–350 (2019).

294. Kongpachith, S. *et al.* Affinity Maturation of the Anti–Citrullinated Protein Antibody Paratope Drives Epitope Spreading and Polyreactivity in Rheumatoid Arthritis. *Arthritis Rheumatol.* **71**, 507–517 (2019).

295. Titcombe, P. J. *et al.* Pathogenic Citrulline-Multispecific B Cell Receptor Clades in Rheumatoid Arthritis. *Arthritis Rheumatol.* **70**, 1933–1945 (2018).

296. Sahlström, P. *et al.* Different Hierarchies of Anti–Modified Protein Autoantibody Reactivities in Rheumatoid Arthritis. *Arthritis Rheumatol.* **72**, 1643–1657 (2020).

297. Mouquet, H. *et al.* Polyreactivity increases the apparent affinity of anti-HIV antibodies by heteroligation. *Nature* **467**, 591–595 (2010).

298. Guthmiller, J. J. *et al.* Polyreactive Broadly Neutralizing B cells Are Selected to Provide Defense against Pandemic Threat Influenza Viruses. *Immunity* **53**, 1230-

1244.e5 (2020).

299. Tiller, T. *et al.* Autoreactivity in Human IgG+ Memory B Cells. *Immunity* **26**, 205–213 (2007).

300. Maschmeyer, P. *et al.* Immunological memory in rheumatic inflammation - a roadblock to tolerance induction. *Nat. Rev. Rheumatol.* **17**, 291–305 (2021).

301. He, L. *et al.* Toward a more accurate view of human B-cell repertoire by next-generation sequencing, unbiased repertoire capture and single-molecule barcoding. *Sci. Rep.* **4**, 6778 (2014).

302. Jiang, N. *et al.* Lineage structure of the human antibody repertoire in response to influenza vaccination. *Sci. Transl. Med.* **5**, 171ra19 (2013).

303. Khan, T. A. *et al.* Accurate and predictive antibody repertoire profiling by molecular amplification fingerprinting. *Sci. Adv.* **2**, e1501371 (2016).

304. Travers, K. J., Chin, C.-S., Rank, D. R., Eid, J. S. & Turner, S. W. A flexible and efficient template format for circular consensus sequencing and SNP detection. *Nucleic Acids Res.* **38**, e159 (2010).

305. Rosenfeld, A. M. *et al.* Computational Evaluation of B-Cell Clone Sizes in Bulk Populations. *Front. Immunol.* **9**, 1472 (2018).

306. Vergani, S. *et al.* Novel Method for High-Throughput Full-Length IGHV-D-J Sequencing of the Immune Repertoire from Bulk B-Cells with Single-Cell Resolution. *Front. Immunol.* **8**, 1157 (2017).

307. Sasso, E. H., Buckner, J. H. & Suzuki, L. A. Ethnic differences of polymorphism of an immunoglobulin VH3 gene. *J. Clin. Invest.* **96**, 1591–1600 (1995).

308. Watson, C. T., Glanville, J. & Marasco, W. A. The Individual and Population Genetics of Antibody Immunity. *Trends Immunol.* **38**, 459–470 (2017).

309. Tam, V. *et al.* Benefits and limitations of genome-wide association studies. *Nat. Rev. Genet.* **20**, 467–484 (2019).

310. Parks, T. *et al.* Association between a common immunoglobulin heavy chain allele and rheumatic heart disease risk in Oceania. *Nat. Commun.* **8**, 14946 (2017).

311. van de Bovenkamp, F. S. *et al.* Adaptive antibody diversification through N-linked glycosylation of the immunoglobulin variable region. *Proc. Natl. Acad. Sci. U. S. A.* **115**, 1901–1906 (2018).

312. Youings, A., Chang, S. C., Dwek, R. A. & Scragg, I. G. Site-specific glycosylation of human immunoglobulin G is altered in four rheumatoid arthritis patients. *Biochem. J.* **314 (Pt 2**, 621–630 (1996).

313. Sabouri, Z. *et al.* Redemption of autoantibodies on anergic B cells by variable-region glycosylation and mutation away from self-reactivity. *Proc. Natl. Acad. Sci. U. S. A.* **111**, E2567-75 (2014).

314. Eyre, S. *et al.* High-density genetic mapping identifies new susceptibility loci for rheumatoid arthritis. *Nat. Genet.* **44**, 1336–1340 (2012).

315. Watson, C. T. & Breden, F. The immunoglobulin heavy chain locus: genetic variation,

missing data, and implications for human disease. *Genes Immun.* **13**, 363–373 (2012).

316. Mikocziova, I., Greiff, V. & Sollid, L. M. Immunoglobulin germline gene variation and its impact on human disease. *Genes Immun.* **22**, 205–217 (2021).

317. Collins, A. M., Yaari, G., Shepherd, A. J., Lees, W. & Watson, C. T. Germline immunoglobulin genes: Disease susceptibility genes hidden in plain sight? *Curr. Opin. Syst. Biol.* **24**, 100–108 (2020).

318. Ohlin, M. Poorly Expressed Alleles of Several Human Immunoglobulin Heavy Chain Variable Genes are Common in the Human Population. *Front. Immunol.* **11**, (2021).

319. Sasso, E. H., Buckner, J. H. & Suzuki, L. A. Ethnic differences of polymorphism of an immunoglobulin VH3 gene. *J. Clin. Invest.* **96**, 1591–1600 (1995).

320. Lewis, M. J. *et al.* Molecular Portraits of Early Rheumatoid Arthritis Identify Clinical and Treatment Response Phenotypes In Brief. *CellReports* **28**, 2455-2470.e5 (2019).

321. Corsiero, E., Pratesi, F., Prediletto, E., Bombardieri, M. & Migliorini, P. NETosis as source of autoantigens in rheumatoid arthritis. *Frontiers in Immunology* **7**, 485 (2016).

322. Teng, Y. K. O. *et al.* Immunohistochemical analysis as a means to predict responsiveness to rituximab treatment. *Arthritis Rheum.* **56**, 3909–3918 (2007).

323. Padyukov, L. *et al.* A genome-wide association study suggests contrasting associations in ACPA-positive versus ACPA-negative rheumatoid arthritis. *Ann. Rheum. Dis.* **70**, 259–265 (2011).

324. Triaille, C. *et al.* Common Transcriptomic Effects of Abatacept and Other DMARDs on Rheumatoid Arthritis Synovial Tissue. *Front. Immunol.* **12**, 724895 (2021).

325. Buch, M. H. *et al.* Mode of action of abatacept in rheumatoid arthritis patients having failed tumour necrosis factor blockade: a histological, gene expression and dynamic magnetic resonance imaging pilot study. *Ann. Rheum. Dis.* **68**, 1220–1227 (2009).

326. De Klerck, B. *et al.* Pro-inflammatory properties of stromal cell-derived factor-1 (CXCL12) in collagen-induced arthritis. *Arthritis Res. Ther. 2005 76* **7**, 1–13 (2005).

327. Cole, S. *et al.* Integrative analysis reveals CD38 as a therapeutic target for plasma cell-rich pre-disease and established rheumatoid arthritis and systemic lupus erythematosus. *Arthritis Res. Ther.* **20**, 1–14 (2018).

328. Van Vollenhoven, R. F., Kinnman, N., Vincent, E., Wax, S. & Bathon, J. Atacicept in patients with rheumatoid arthritis and an inadequate response to methotrexate: Results of a phase II, randomized, placebo-controlled trial. *Arthritis Rheum.* **63**, 1782–1792 (2011).

329. Genovese, M. C., Kinnman, N., De La Bourdonnaye, G., Pena Rossi, C. & Tak, P. P. Atacicept in patients with rheumatoid arthritis and an inadequate response to tumor necrosis factor antagonist therapy: Results of a phase II, randomized, placebo-controlled, dose-finding trial. *Arthritis Rheum.* **63**, 1793–1803 (2011).

330. Raposo, B. *et al.* Divergent and dominant anti-inflammatory effects of patient-derived anti-citrullinated protein antibodies (ACPA) in arthritis development Corresponding authors. doi:10.1101/2022.09.30.510377

331. Falahee, M. & Raza, K. Perspectives of at-Risk Individuals on Preventive Intervention

for Rheumatoid Arthritis: A Mini Review. *Front. Immunol.* **13**, (2022).

Several figures are created with BioRender.com.