UNIVERSITÄT ZU LÜBECK

Aus dem Institut für Anatomie
Direktor: Prof. Dr. Jürgen Westermann

Analyse der T-Zellreaktion gegen komplexe zelluläre Antigene auf Basis des Rezeptorrepertoires

Inauguraldissertation zur Erlangung der Doktorwürde
der Universität zu Lübeck

-Aus der Sektion Medizin-
vorgelegt von Martin Meinhardt
aus Tübingen

Lübeck 2022

1. Berichterstatter*in:	Prof. Dr. med. Jürgen Westermann
2. Berichterstatter*in:	Prof. Dr. rer. biol. hum. Inke König
Tag der mündlichen Prüfung:	18.04.2023
Zum Druck genehmigt:	Lübeck den 18.04.2023

-Promotionskomission der Sektion Medizin-

Inhaltsverzeichnis

Abkürzungsverzeichnis

CD	cluster of differentiation
CDR	complementary determining region
CN	Kopienzahl (engl. copy number)
d	Tage (engl. days)
FDR	Falscherkennungsrate (engl. false discovery rate)
MHC	major histocompatibility complex
NGS	next generation sequencing
PBS	phosphate-buffered saline
pMHC	Kombination aus Peptidfragment und MHC
SRBC	Schaferythrozyten (engl. sheep red blood cells)
μ	Mittelwert
σ	Standardabweichung

1 Einleitung

1.1 Aufgabe und Funktion der T-Zellen

T-Zellen nehmen innerhalb des Immunsystems Schlüsselfunktionen wahr. Sie erfüllen einerseits eine Vielzahl verschiedener Steuerungsfunktionen, sind aber auch unmittelbar an der Elimination von Pathogenen beteiligt. Entscheidend für ihre Funktion ist der T-Zellrezeptor, mit dessen Hilfe eine T-Zelle spezifische Antigene (i.d.R. Peptidfragmente) erkennen kann. Aufgebaut ist der T-Zellrezeptor aus einem Heterodimer, von welchem zwei verschiedene Formen bekannt sind. Die große Mehrheit der zirkulierenden T-Zellen ist bei Menschen und Mäusen gleichermaßen dem αβ-Typ zuzurechnen [16]. Dies bedeutet, dass der Rezeptor dieser Zellen aus je einer α- und einer β-Kette aufgebaut ist. Neben dem αβ-Typ existiert ein weiterer T-Zelltyp, dessen Rezeptor aus einer γ- und einer δ- Kette besteht [31]. Da dieser Typ in der vorliegenden Arbeit keine Berücksichtigung findet, beziehen sich alle folgenden Ausführungen ausschließlich auf den αβ-Typ. Die verschiedenen funktionellen T-Zellgruppen weisen unterschiedliche Oberflächenantigene auf, welche nach dem CD-System (engl. cluster of differentiation) eingeteilt werden. Während T-Helferzellen an ihrer Zelloberfläche CD4-Moleküle exprimieren, ist das CD8-Molekül das charakteristische Merkmal der zytotoxischen T-Zellen. Obwohl zwischen dem T-Zellrezeptor und löslichen Immunglobulinen eine Vielzahl von Parallelen bestehen, existiert ein fundamentaler Unterschied: Während Immunglobuline an unprozessierte (native) Antigene binden können, kann eine T-Zelle ihr Antigen nur erkennen, wenn dieses von einer anderen Zelle präsentiert wird. Zum Zweck der Antigenpräsentation und damit der T-Zellaktivierung existieren spezielle Molekülstrukturen, welche als Haupthistokompatibilitätskomplexe (engl. major histocompatibility complex, MHC) bezeichnet werden. Die Kombination aus Haupthistokompatibilitätskomplex und Antigen (Peptidfragment) wird gemeinhin mit pMHC abgekürzt. Bei der Aktivierung von T-Zellen spielen zwei verschiedene MHC-Subtypen eine besondere Rolle. Zytotoxische T-Zellen ($CD8^+$) interagieren mit MHC-Komplexen vom Typ I, welche auf allen kernhaltigen Zellen exprimiert werden. Ihre Aufgabe besteht in erster Linie in der Erkennung und Elimination von Zellen, welche mit Viren infiziert sind sowie von Tumorzellen. Klinisch kommt ihnen bei der Frage, ob ein transplantiertes Organ abgestoßen oder toleriert wird, eine besondere Bedeutung zu. Im Unterschied hierzu nehmen CD4-positive Zellen in erster Linie Hilfsfunktionen war. Sie interagieren mit

MHC-Komplexen vom Typ II, die von Zellen exprimiert werden, welche auf die Präsentation phagozytierter Antigene spezialisiert sind. Aus einer naiven $CD4^+$-T-Zelle können sich kontextabhängig verschiedene Arten von Effektor-T-Zellen entwickeln. Die Hauptaufgabe von T-Helferzellen vom Typ I (TH-1) besteht in der Aktivierung von Makrophagen und zytotoxischen T-Zellen (zelluläre Immunität). Im Gegensatz hierzu kontrollieren T-Helferzellen vom Typ II (TH-2) die Aktivierung von eosinophilen Granulozyten und Mastzellen. Lange Zeit wurde diesen Zellen auch eine entscheidende Funktion bei der Aktivierung von B-Zellen zu antikörperproduzierenden Plasmazellen zugeschrieben. Dies wurde jedoch mit der Entdeckung der follikulären T-Helferzellen, welche auf diese Aufgabe spezialisiert sind, relativiert [10, 31]. Eine naive T-Zelle wird durch Präsentation des passenden Antigens durch eine andere Zelle aktiviert. Für diesen, auch als „Priming“ bezeichneten Vorgang, sind neben der Antigenpräsentation weitere kostimulierende Signale erforderlich. Die so aktivierte T-Zelle reagiert mit intensiver Proliferation und Differenzierung zu Effektorzellen. Hierdurch entsteht eine große Anzahl identischer T-Zellklone, welche für die Bekämpfung des aktivierenden Antigens prädestiniert sind. Werden diesen Effektorzellen erneut passende Antigene präsentiert, so üben diese ihre zytotoxische bzw. Hilfsfunktion aus, ohne dass hierfür weitere kostimulierende Signale erforderlich wären. Ein kleiner Teil der ausdifferenzierten T-Zellen bleibt zur Aufrechterhaltung eines immunologischen Gedächtnisses für längere Zeit im Organismus erhalten. Für weitere Details zur Entwicklung und Funktion der T-Zellen sei der Leser auf [3], [31] und [45] verwiesen.

1.2 Diversität und Homogenität des T-Zellrezeptorrepertoires

Allen T-Zellen ist gemeinsam, dass ihre Rezeptoren auf die Erkennung einiger wenige Antigene beschränkt sind. Gleichzeitig ist der Organismus einer Vielzahl verschiedener Pathogene ausgesetzt. Die Spezialisierung der einzelnen T-Zelle einerseits und die Vielfalt potenzieller Pathogene andererseits, erfordern ein T-Zellrezeptorrepertoire von hoher Diversität. In der Maus wird die Anzahl verschiedener $\alpha\beta$- Rezeptoren auf $\sim 2{\cdot}10^6$ geschätzt [33]. Diese Diversität wird durch das Prinzip der somatischen Rekombination ermöglicht. V- und J-Segmente, welche im Erbgut mehrfach angelegt sind, werden zufällig miteinander kombiniert. Anders als bei der α-Kette, wird bei der Generierung der β-Kette ein zusätzliches D-Segment zwischen V- und J-Segment eingebaut. An

den Übergängen zwischen den einzelnen Segmenten werden zufällig Nukleotide eingefügt bzw. entfernt. Hierdurch wird die Anzahl der potenziell generierbaren Rezeptoren enorm gesteigert. Bei diesem stochastischen Kombinationsprozess entstehen regelhaft autoreaktive Klone, welche im Rahmen des Reifungsprozesses im Thymus eliminiert werden. Jede der beiden Ketten des T-Zellrezeptors beinhaltet drei hypervariable Regionen welche mit CDR1-3 (engl. complementary determining region) bezeichnet werden. Während die Regionen CDR1-2 auf dem V-Segment kodiert sind, überspannt die CDR3-Region die Verbindung des V- (D-) und J-Segmentes, was zu einem Höchstmaß an Diversität in diesem Bereich führt. Diese zusätzliche Diversität spiegelt die besondere Funktion dieses Bereiches im Falle eines Antigenkontaktes wider. Während die Regionen CDR1-2 in erster Linie mit dem MHC-Komplex interagieren, auf welchem das Antigen präsentiert wird, besteht die Aufgabe der CDR3-Region in der unmittelbaren Interaktion mit dem präsentierten Antigen [30, 31]. Die Antigenspezifität des T-Zellrezeptors wird daher in hohem Maße von diesen Bereichen determiniert. Das Prinzip der somatischen Rekombination ermöglicht es, eine unübersehbare Vielfalt verschiedener T-Zellrezeptoren zu generieren. Der zugrundeliegende Kombinationsprozess vollzieht sich unabhängig von aktuell verfügbaren Antigenen. Auf diese Weise entsteht eine umfangreiche Bibliothek an T-Zellspezifitäten, auf die bei Auftreten eines Pathogens zurückgegriffen werden kann. Selbst gegen Pathogene, mit denen ein Individuum, oder auch eine gesamte Spezies erstmalig konfrontiert wird, kann so eine adäquate T-Zellantwort gewährleistet werden. In der Maus wird die Anzahl an potentiell generierbaren αβ-Rezeptoren auf $\sim 10^{15}$ geschätzt [11, 33]. Diese Zahl übersteigt den Umfang des im einzelnen Tier tatsächlich realisierten Repertoires um Größenordnungen. Würde man annehmen, dass jede der potenziell generierbaren Sequenzen mit der gleichen Wahrscheinlichkeit realisiert würde, wären zwischen den T-Zellrezeptorrepertoires verschiedener Individuen keine relevanten Schnittmengen zu erwarten [25]. Tatsächlich können jedoch bestimmte CDR3-Sequenzen in einem Großteil der Individuen detektiert werden. In der Literatur werden solche Sequenzen als *public* charakterisiert [12]. Oftmals sind solche Sequenzen nicht nur in vielen Individuen vorhanden, sondern werden auch innerhalb desselben Individuums vielfach detektiert. Qualitativ zeichnen sie sich durch eine, gegenüber dem Gesamtrepertoire, modifizierte Verteilung der V- bzw. J-Segmente sowie eine im Mittel etwas verkürzte CDR3-Sequenz aus [25]. Eine T-Zellreaktion, welche von derartigen Klonen dominiert wird, bezeichnet man übli-

cherweise als *public response*. In Abgrenzung hierzu wird eine *private response* von individuellen Klonen dominiert.

1.3 Analyse des T-Zellrezeptorrepertoires mit Hilfe moderner Sequenzierverfahren

Wann immer eine T-Zellreaktion durch Kontakt mit einem Antigen hervorgerufen wird, führen die darauffolgenden Proliferations- und Differenzierungsvorgänge zu einer Umgestaltung des T-Zellrezeptorrepertoires. Durch Analyse des Rezeptorrepertoires ist es daher möglich, Rückschlüsse auf aktuelle immunologische Vorgänge zu ziehen. Moderne Sequenzierverfahren, welche es ermöglichen, Millionen von Sequenzen synchron auszulesen, werden gemeinhin unter dem Begriff *next generation sequencing (NGS)* zusammengefasst [43]. Im Fokus steht hierbei insbesondere die CDR3-Region, da diese unmittelbar mit dem präsentierten Antigen interagiert und daher am ehesten Rückschlüsse auf die T-Zellreaktion zulässt. In Abhängigkeit der zu untersuchenden Fragestellung kommen bei der Sequenzierung des T-Zellrezeptorrepertoires verschiedene technische Vorgehensweisen zur Anwendung. Die Datengrundlage dieser Arbeit beruht ausschließlich auf Populationsanalysen (engl. *bulk methods*). Dies bedeutet, dass die ausgelesenen α- bzw. β- Ketten lediglich innerhalb des Gesamtrepertoires quantifiziert werden. Sie werden hierbei nicht einer einzelnen Zelle zugeordnet. Dies schließt Aussagen über die Paarung von α- und β-Kette kategorisch aus. Da die Spezifität des T-Zellrezeptors erst durch die Kombination der beiden Ketten determiniert wird, ist diese Einschränkung im Hinblick auf funktionelle Aspekte des T-Zellrezeptorrepertoires durchaus relevant. Andererseits bietet diese Vorgehensweise gegenüber zellbasierten Verfahren (engl. *single cell methods*) eine Fülle von Vorteilen. Da für Populationsanalysen keine lebenden Zellen benötigt werden, stellen sie deutlich geringere Anforderungen an das biologische Ausgangsmaterial und sind zudem kostengünstiger. Vor allem aber sind sie zur Analyse großer Zellmengen prädestiniert, während zellbasierte Methoden hier oftmals an ihre Grenzen stoßen. Sämtliche, in dieser Arbeit analysierte Datensätze beruhen auf mRNA-Sequenzen. Dies bietet gegenüber der Analyse genomischer DNA den entscheidenden Vorteil, dass neben der Zellzahl auch das Aktivitätsniveau der jeweiligen T-Zellen (aufgrund des erhöhten Expressionslevels der mRNA-Sequenzen) in das quantitative Ergebnis miteinfließt. Andererseits ist es so nicht möglich, Rückschlüsse auf

die Anzahl der T-Zellen eines Klons zu ziehen. Eine umfassende Übersicht über die verschiedenen Methoden zur Analyse des T-Zellrezeptorrepertoires findet sich in [38]. Unabhängig von der genauen technischen Vorgehensweise fallen bei der NGS-basierten Analyse des T-Zellrezeptorrepertoires umfangreiche Datenmengen an, welche die hohe Diversität des T-Zellrezeptorrepertoires zumindest in Teilen widerspiegeln. Gleichzeitig ist im Fall einer Immunreaktion meist nur ein Bruchteil der T-Zellen eines Organismus in die aktuelle T-Zellantwort involviert. Selbst bei T-Zellreaktionen gegen komplexe Antigene (beispielsweise Bakterien) wird der Anteil auf 0,01%-0,1% geschätzt [23]. Es liegt daher auf der Hand, dass eine Immunreaktion in der Regel nur zu dezenten Veränderungen innerhalb des sequenzierten Repertoires führen kann. Die entscheidenden Informationen müssen hierbei meist aus mehreren biologischen Parametern extrahiert werden. Hierfür hat sich die Anwendung statistischer Kennzahlen bewährt, welche ursprünglich für Fragestellungen aus den Bereichen der Ökologie und der Wirtschaftswissenschaften entwickelt wurden [20, 38, 40]. In der vorliegenden Arbeit wird eine hochkomplexe T-Zellreaktion anhand eines etablierten Mausmodells untersucht. Die Tiere wurden hierbei mit Schaferythrozyten (engl. sheep red blood cells, SRBC) immunisiert. Es handelt sich hierbei um ein großes zelluläres Antigen, aus welchem eine große Anzahl verschiedener Epitope extrahiert werden kann. Dementsprechend ist an der resultierenden T-Zellreaktion eine Vielzahl verschiedener T-Zellklone beteiligt. Um eine solche Reaktion beobachten und analysieren zu können, wurden zwei etablierte statistische Kennzahlen zu einem flexiblen Kennzahlsystem verallgemeinert. Dieses System ermöglicht es, das T-Zellrezeptorrepertoire im Hinblick auf nahezu beliebige biologische Parameter zu analysieren. Im Folgenden werden die elementaren Eigenschaften des angewendeten Mausmodells kurz zusammengefasst.

1.4 Die T-Zellreaktion im SRBC-Modell

Durch intravenöse Applikation von Schaferythrozyten (engl. sheep red blood cells, SRBC) kann in Mäusen gezielt eine T-Zellreaktion ausgelöst werden, ohne die Tiere hierbei besonderem Stress auszusetzen [44, 51]. Die applizierten Schaferythrozyten werden innerhalb weniger Stunden aus dem Blutstrom eliminiert [44, 52], es ist daher naheliegend, dass die nachfolgende T-Zellreaktion einer zeitlich relativ genau festgelegten Dynamik folgt. Ausgangspunkt dieser Dissertation sind zwei Publikationen, welche

in jüngerer Zeit am Institut für Anatomie der Universität zu Lübeck entstanden sind. In diesen wurde die SRBC-induzierte T-Zellantwort auf Basis von NGS-Daten untersucht. In [47] wurde gezeigt, dass es in diesem Modell, drei Tage nach Applikation des Antigens, zu einer starken Expansion einzelner Klonotypen (Menge an T-Zellen mit gleicher CDR3-Aminosäuresequenz, siehe Abschnitt 2.2) in den T-Zellzonen der Milz kommt. Es stellte sich hierbei heraus, dass in zwei verschiedenen T-Zellzonen meist verschiedene Klonotypen expandieren. Dieses, als klonale Segregation bezeichnete, Phänomen ist selbst dann noch gegeben, wenn zwei T-Zellzonen aus derselben Milz miteinander verglichen werden. Ein analoger Effekt wurde in [51] zwischen den T-Zellrezeptorrepertoires beobachtet, welche aus Milzschnitten unterschiedlicher Mäuse extrahiert wurden. Die SRBC-induzierte T-Zellreaktion wird also von solchen Klonotypen dominiert, welche lediglich in einem (oder sehr wenigen) Individuen vorkommen (*private response*). Daher führt die SRBC-induzierte T-Zellantwort gleichermaßen zu einer Divergenz zwischen den Repertoires unterschiedlicher Tiere, als auch zu einem analogen Effekt zwischen verschiedenen histologischen Kompartimenten innerhalb desselben Individuums. Es stellte sich heraus, dass dieser Divergenzeffekt nur für einen kurzen Zeitraum besteht. Er ist am dritten Tag nach Antigenapplikation stark ausgeprägt, jedoch bereits einen Tag später deutlich abgeschwächt. Die T-Zellantwort dauert zu diesem Zeitpunkt jedoch unverändert an. Dies zeigt sich an hohen Proliferationsraten der T-Zellen sowie der Ausbildung von Keimzentren [44, 51]. Auch innerhalb des Rezeptorrepertoires finden sich Hinweise auf ein Fortdauern der T-Zellantwort. Mit Hilfe einer Genexpressionsanalyse konnte eine kleine Population an Klonotypen identifiziert werden, welche in der Mehrzahl der immunisierten Tiere expandiert. Neben den reagiblen *private*-Klonotypen besteht also noch eine nachgeordnete *public*-Komponente. Die Expansion dieser Klonotypen erreicht ihren Höhepunkt am 4. Tag nach Antigenapplikation [47, 51]. Zu diesem Zeitpunkt sind die Divergenzeffekte zwischen den Repertoires der einzelnen Tiere bereits deutlich rückläufig. Die naheliegende Erklärung, dass die wenigen *public*-Klonotypen einer langsameren Kinetik folgen, greift zu kurz. Dies zeigt sich insbesondere an der Verteilung der detektierten V- bzw. J-Segmente, welche im Lauf der Immunreaktion erheblichen Verschiebungen unterworfen ist. Während sich solche Verschiebungen zu Beginn der Immunreaktion nur innerhalb der häufigsten Klonotypen detektieren lassen, ist das Gesamtrepertoire am 7. Tag nach Antigenapplikation signifikant verändert, was auf eine umfangreiche Durchsetzung des Repertoires

mit reagiblen Klonotypen schließen lässt [51]. Zusammenfassend lässt sich sagen, dass es nach Immunisierung mit SRBC gleichermaßen zu hochflüchtigen Divergenzeffekten innerhalb der häufigen Klonotypen als auch zu einer längerfristigen Umgestaltung des T-Zellrezeptorrepertoires kommt.

1.5 Vorgehensweise und Ziele dieser Arbeit

Anhand des SRBC-Models wird in dieser Arbeit ein flexibles System statistischer Kennzahlen zur systematischen Analyse einer hochkomplexen T-Zellreaktion entwickelt. Dieses System ermöglicht es, sowohl die Homogenität der Klonotypen eines Datensatzes[1] als auch Ähnlichkeiten zwischen verschiedenen Datensätzen zu quantifizieren. Ausgangspunkt sind hierbei drei etablierte statistische Kennzahlen, welche standardmäßig zur Analyse des T-Zellrezeptorrepertoires eingesetzt werden. Die entscheidende Verallgemeinerung besteht darin, dass die einzelnen Sequenzen nicht mehr im Hinblick auf Übereinstimmung, etwa der CDR3-Region, sondern auf ein variables Ähnlichkeitskriterium hin evaluiert werden. Dieses Kriterium kann vom Anwender frei gewählt werden und richtet sich insbesondere nach der konkreten Fragestellung. Im Hinblick auf die bestehenden Ergebnisse sollen hierbei insbesondere drei biologische Parameter im SRBC-Modell untersucht werden, wobei die Datengrundlage ausschließlich auf β-Ketten des T-Zellrezeptorrepertoires beruht. Der erste Parameter betrachtet die CDR3-Region, welche für die Spezifität einer T-Zelle maßgeblich ist. Verschiedene CDR3-Sequenzen, welche sich nur um einzelne Aminosäuren unterscheiden, wurden sowohl in naiven als auch in antigenspezifischen T-Zellpopulationen vorbeschrieben [15, 24]. Es wäre daher gleichermaßen denkbar, dass es nach Antigenexposition zu einem vermehrten, als auch zu einem verminderten Auftreten derart ähnlicher Sequenzen kommt. In jedem Fall können solche Homogenisierungs- bzw. Heterogenisierungseffekte Hinweise auf die Anwesenheit reagibler Klonotypen liefern. Als weiterer Parameter werden die im T-Zellrezeptorrepertoire detektierten V- bzw. J-Segmente analysiert. Mitunter können diese Segmente Hinweise auf die Funktion bestimmter T-Zellpopulationen liefern [40]. Ihre Verteilung wird vielfach zur Gruppierung verschiedener T-Zellrezeptorrepertoires genutzt [20]. Die bereits in [51] gezeigte Heterogenisierung der detektierten V- bzw. J-Segmente, nach SRBC-Applikation, wird in dieser Arbeit systematisch analysiert.

[1] Der Begriff *Datensatz* bezieht sich in dieser Arbeit stets auf das T-Zellrezeptorrepertoire, welches aus einer einzelnen Probe extrahiert wurde.

Zur Generierung der β-Kette stehen in der Maus 22 V- und 13 J-Segmente zur Verfügung [1]. Verglichen mit der Anzahl möglicher CDR3-Sequenzen, ist die Anzahl der Kombinationsmöglichkeiten der Segmente somit vergleichsweise überschaubar, was sowohl die Auswertung der Daten als auch deren biologische Interpretation stark vereinfacht. Zudem legt [51] nahe, dass dieser Parameter Immunisierungseffekte in weiten Teilen des Repertoires zu detektieren vermag und hierbei ein hohes Maß an Sensitivität aufweist. Als dritter Parameter wird die Diversität der Nukleotidkodierungen der CDR3-Aminosäuresequenzen analysiert. Aufgrund der Degeneration des genetischen Codes kann eine einzelne CDR3-Aminosäuresequenz von einer Vielzahl verschiedener Nukleotidsequenzen kodiert werden. Während häufige *public*-Klonotypen oft von einer Vielzahl verschiedener Nukleotidsequenzen kodiert werden [25, 53, 54] stimmen reagible T-Zellklone, welche jeweils auf eine einzige aktivierte T-Zelle zurückgehen, sowohl in der Aminosäuresequenz also auch in der Nukleotidsequenz ihrer CDR3-Regionen überein. Die Diversität der Nukleotidkodierungen einer einzelnen CDR3-Aminosäuresequenzen ermöglicht so eine grobe Einschätzung im Hinblick auf Herkunft und Funktionalität der zugehörigen T-Zellen.

Die Datengrundlage dieser Arbeit besteht aus T-Zellrezeptorrepertoires, welche mittels NGS aus Milzschnitten extrahiert wurden. Die hierfür verwendeten Versuchstiere wurden in vier experimentelle Gruppen aufgeteilt. Drei dieser Gruppen wurden mit SRBC immunisiert und zu verschiedenen Zeitpunkten (3d, 4d, 7d nach Antigenapplikation) getötet. Die verbliebene Gruppe dient als Kontrolle. Innerhalb eines Datensatzes wird ein T-Zellklon durch seine CDR3-Sequenz, die zugeordneten V- bzw. J-Segmente sowie die Kopienzahl repräsentiert. Letztere gibt an, wie oft der Klon in der jeweiligen Probe detektiert wurde. Die identifizierten CDR3-Sequenzen überspannen im Mittel einen Bereich von $\sim 12 \pm 1,4$ ($\mu \pm \sigma$) Aminosäuren. Tabelle 1 zeigt beispielhaft einen kleinen Ausschnitt eines solchen Datensatzes. Die drei eingangs genannten Parameter (CDR3-Sequenzen, V- bzw. J-Segmente, Nukleotidkodierung) werden zunächst in den vier experimentellen Gruppen analysiert und verglichen. Hierbei werden neben dem Gesamtrepertoire auch Teilbereiche analysiert. Diese werden auf Basis der Häufigkeit (Kopienzahlen) der jeweiligen Sequenzen systematisch voneinander abgegrenzt. Auf diese Weise lassen sich nicht nur eine Vielzahl von Immunisierungseffekten innerhalb des T-Zellrezeptorrepertoires detektieren, sondern auch konkrete Teilbereiche innerhalb der Datensätze eingrenzen, in denen ein Großteil der reagiblen Klone vermutet

CDR3-Nukleotidsequenz	CDR3-Aminosäuresequenz	V-Segment	J-Segment	Kopienzahl
GCCAGC...CAGTAC	ASSQEWTGYEQY	V5	J2-7	6245
GCCAGC...TTGTAC	ASAGGRQNTLY	V19	J2-4	5892
GCCAGC...CAGTAC	ASSLAGREQY	V15	J2-7	4233
GCCAGC...CAGTAC	ASSSGDEQY	V3	J2-7	3546
GCCAGC...CAGTAC	ASSLALGYEQY	V15	J2-7	2706
⋮	⋮	⋮	⋮	⋮

Tabelle 1: Datenstruktur der extrahierten T-Zellrezeptorrepertoires. Jeder T-Zellklon wird durch die zugehörige CDR3-Sequenz, die zugeordneten Segmente sowie seine Kopienzahl repräsentiert. Letztere gibt an, wie oft der Klon innerhalb der jeweiligen Probe detektiert wurde. Die Tabelle zeigt exemplarisch die fünf häufigsten Sequenzen eines Tieres aus der Kontrollgruppe.

werden kann. Sofern nichts anderes angegeben, bezieht sich der Begriff *Teilrepertoire* in dieser Arbeit stets auf den Teilbereich eines Datensatzes, in welchem die Kopienzahlen der Klonotypen in einem gegebenen Intervall liegen. Um festzustellen, welche der Parameter eine hinreichende Trennschärfe aufweisen, um die einzelnen Tieren im Hinblick auf den Immunisierungsstatus zu klassifizieren, kommen in dieser Arbeit zwei statistische Lernverfahren zur Anwendung, welches auf den neuen statistischen Kennzahlen beruhen. Mit diesen Verfahren wird untersucht, inwiefern sich die Repertoires der immunisierten Tiere von denen der Kontrollgruppe abgrenzen lassen. Im Hinblick auf das SRBC-Modell sollen mit Hilfe dieses statistischen Werkzeugkastens Antworten auf folgende Fragen gefunden werden:

1. Wie ist die schnelle Rückbildung der immunisierungsbedingten Divergenzeffekte [47] vor dem Hintergrund einer fortdauernden T-Zellantwort zu erklären?

2. Wie können Teilbereiche im sequenzierten Repertoire identifiziert werden, in denen sich reagible Klone konzentrieren?

3. Wie verteilen sich die reagierenden *public*-Klonotypen innerhalb des Rezeptorrepertoires und besteht hierbei ein Unterschied zur vorherrschenden *private response*?

Letztlich steht das SRBC-Modell in dieser Arbeit jedoch nur beispielhaft für eine hochkomplexe T-Zellreaktion. Bei einer solchen Reaktion kann, aufgrund der Vielfalt

und Inhomogenität der resultierenden Epitope, nicht davon ausgegangen werden, dass die reagiblen T-Zellklone anhand gemeinsamer Merkmale, beispielsweise homogener Muster innerhalb der CDR3-Region, identifiziert werden können. Die hier vorgeschlagene Methodik wurde daher primär mit dem Ziel entwickelt, unabhängig von gemeinsamen Strukturen innerhalb der Rezeptoren reagibler Klonotypen angewendet werden zu können. Die Kennzahlen und Algorithmen sollen dabei auch auf andere Modelle und Fragestellungen anwendbar sein. Auch die im Rahmen dieser Promotion entwickelten Computerprogramme sind darauf ausgelegt, mit wenig Aufwand, an verschiedene experimentelle Gegebenheiten angepasst werden zu können. Die wesentlichen Ergebnisse dieser Arbeit wurden vorab in der Zeitschrift *PLOS ONE* veröffentlicht [29]. Da diese Publikation vollständig auf der vorliegenden Dissertation beruht, wird aus Gründen der Übersichtlichkeit stellenweise auf eine gesonderte Zitierung verzichtet. Die in dieser Arbeit verwendete Notation, insbesondere im Hinblick auf Abkürzungen der neu eingeführten Kennzahlen, orientiert sich so weit wie möglich an der (englischsprachigen) Publikation.

2 Material und Methoden

2.1 Mausmodell und Probenentnahme[1]

Um eine kontrollierte T-Zellreaktion auszulösen, wurden weiblichen C57BL/6J Mäusen im Alter zwischen 8 und 12 Wochen Schaferythrozyten (engl. sheep red blood cells, SRBC) injiziert. Hierzu wurden jeweils 10^9 Erythrozyten in 200 µl phosphatgepufferter Salzlösung (engl. phosphate-buffered saline, PBS) gelöst und den Tieren in die Schwanzvene injiziert. Eine Kontrollgruppe erhielt lediglich eine PBS-Injektion ohne zusätzlichen Wirkstoff. Die immunisierten Tiere wurden 3d (n=10), 4d (n=10), und 7d (n=10) nach SRBC-Injektion getötet. Die Kontrolltiere (n=20) wurden unabhängig vom Todeszeitpunkt (3d/4d) zu einer gemeinsamen Kontrollgruppe zusammengefasst (n=20). Da diese Versuche im Rahmen eines Forschungsprojektes über den Einfluss von Schlaf auf funktionelle Aspekte des Immunsystems durchgeführt wurden, wurde jeweils die Hälfte der Tiere aus jeder Gruppe, nach Applikation des Antigens, einem stressarmen Schlafentzug unterzogen. Da diese Form des Schlafentzuges keine signifikanten Auswirkungen auf die T-Zellreaktion entfaltet [51], wurden in dieser Arbeit Tiere mit bzw. ohne Schlafentzug, für jeden der drei Zeitpunkte, zu einer Gruppe zusammengefasst. Nach dem Töten der Tiere wurden Gefrierschnitte der Milz gewonnen und hieraus sämtliche RNA extrahiert. Die β-Ketten des T-Zellrezeptorrepertoires wurden mittels reverser Transkription in DNA überführt und anschließend mittels PCR amplifiziert. Hierbei kamen etablierte Standardverfahren zum Einsatz. Die α-Ketten des T-Zellrezeptors werden in dieser Arbeit nicht berücksichtigt. Weitere technische Details sowie rechtliche Aspekte der Versuchsdurchführung sind in [51] zu finden.

2.2 Extraktion des T-Zellrezeptorrepertoires

Zur Identifikation der CDR3β-Sequenzen, Zusammenfassung der Sequenzen zu Clustern sowie zur Korrektur von Lesefehlern wurde die Plattform MiTCR [7] genutzt. Hierbei wurden die Standardparameter der Benutzeroberfläche ClonoCalc [13] verwen-

[1] Die Tierversuche, die nachfolgende Probenentnahme sowie die Extraktion des T-Zellrezeptorrepertoires mittels NGS wurde von Prof. Dr. Jürgen Westermann, Dr. Cornelia Tune, Dr. Kathrin Kalies, Dr. Andrea Schampel, Lisa-Kristin Schierloh und Dr. René Pagel am Institut für Anatomie der Universität zu Lübeck geplant und durchgeführt. Dies erfolgte in Kooperation mit dem Institut für Neurobiologie der Universität zu Lübeck sowie der Abteilung für Innere Medizin II der Universität Tübingen. Die Beiträge der einzelnen Personen sind in [51] detailliert aufgelistet.

det. Nach Ausschluss funktionsloser Sequenzen wurden aus jedem Milzschnitt im Mittel $\sim 1,9$ Mio. Nukleotidsequenzen extrahiert. Hierbei wurden im Mittel $\sim$ 100.000 verschiedene CDR3-Sequenzen detektiert. Aufgrund von Messartefakten war die Anzahl ausgelesener Sequenzen in der 7d Gruppe signifikant erhöht (im Mittel $\sim 2,7$ Mio. im Vergleich zu $\sim 1,6$ Mio. in den übrigen Gruppen). Um die Vergleichbarkeit der Datensätze wiederherzustellen, wurde daher aus den Datensätzen der 7d Gruppe jeweils eine zufällige Stichprobe von $\sim 1,6$ Mio. Sequenzen gezogen. Da die Anzahl der DNA-Moleküle in einer analysierten Probe (nach PCR-Amplifikation) die Zahl der ausgelesenen Sequenzen um ein Vielfaches übersteigt [6], kann der Sequenzierungsvorgang nährungsweise als eine zufällige Stichprobe mit Zurücklegen interpretiert werden. Um diesen zu simulieren, wurde die Stichprobe aus den 7d-Daten *mit* Zurücklegen gezogen. Die Anzahl der ausgelesenen Sequenzen, welche sich einer bestimmten CDR3 Region zuordnen lassen, wird im Folgenden mit Kopienzahl (engl. copy number, CN) bezeichnet. Nukleotidsequenzen, welche gleiche CDR3β-Aminosäuresequenzen kodieren, wurden zu „hypothetischen Klonen“ , welche als *Klonotypen* bezeichnet wurden, zusammengefasst. Der Begriff *Klonotyp* bezeichnet hierbei eine Menge an T-Zellen, welche identische CDR3β-Aminosäuresequenzen aufweisen. Jedem der so definierten Klonotypen wurden die V- und J-Segmente desjenigen Repräsentanten (Nukleotidsequenz) zugeordnet, welcher im jeweiligen Datensatz die höchste Kopienzahl aufweist. Die Kopienzahl der jeweiligen Aminosäuresequenz ergibt sich aus den aufsummierten Werten der zugehörigen Nukleotidsequenzen. Klonotypen welche lediglich einmalig im jeweiligen Datensatz detektiert wurden (CN = 1) wurden entfernt. Bei diesem Vorgehen ergeben sich, auf Aminosäureebene, für jeden Milzschnitt zwischen $\sim$ 50.000 und $\sim$ 100.000 verschiedene Klonotypen. Insgesamt beinhalten die Datensätze $\sim 1,5 \cdot 10^6$ verschiedene CDR3-Sequenzen. Man beachte, dass sich die Rezeptoren der T-Zellen, welche auf diese Weise zu Klonotypen zusammengefasst werden, durchaus unterscheiden können. Dies ist sowohl im Bereich der α- Kette, als auch innerhalb der β- Kette (außerhalb der CDR3-Region) möglich. Auch ist die hier getroffene Annahme, dass Sequenzen mit gleicher CDR3 Region identische V- bzw. J-Segmente aufweisen, eine starke Vereinfachung, welche die Variabilität des T-Zellrezeptorrepertoires nicht vollumfänglich widerspiegelt [54]. Dennoch finden sich ähnliche Vorgehensweisen in etablierten Softwarepaketen zur Analyse des T-Zellrezeptorrepertoires (siehe z.B. [13] einschließlich des zugehörigen Quellcodes).

2.3 Statistik

Für die Analyse des T-Zellrezeptorrepertoires werden in dieser Arbeit verallgemeinerte Varianten etablierter statistischer Kennzahlen genutzt. Der Einfachheit halber beschränken sich sämtliche der folgenden Betrachtungen auf β-Ketten des T-Zellrezeptorrepertoires auf Aminosäureebene. Grundsätzlich könnte die in diesem Abschnitt eingeführte Methodik jedoch analog auf andere Daten (z.B. die α-Ketten des T-Zellrezeptorrepertoires) angewendet werden. Zunächst muss eine einheitliche Notation eingeführt werden. Im Folgenden bezeichne Ω die Menge aller möglichen β-Ketten des T-Zellrezeptorrepertoires. Ein Datensatz X lässt sich als endliche Menge von Wertepaaren $\big(x_i, \nu_X(x_i)\big)_{i=1,\dots,m}$ auffassen, wobei $\nu_X(x)$ die Kopienzahl der Sequenz $x \in \Omega$ in X angibt. Die Variable X wird im Folgenden gleichermaßen für einen Datensatz als auch für die Menge der darin enthaltenen Klonotypen genutzt.

2.3.1 Quantifizierung der Homogenität des T-Zellrezeptorrepertoires

Die in diesem Abschnitt eingeführten statistischen Kennzahlen dienen dazu, die Homogenität bzw. Diversität der Sequenzen innerhalb eines Datensatzes zu quantifizieren. Grundlage ist hierfür eine Kennzahl, welche ursprünglich dazu entwickelt wurde, um die Diversität von Ökosystemen zu analysieren. Sie gehört heute zu den Standardgrößen, welche bei der Analyse des T-Zellrezeptorrepertoires eingesetzt werden.

Definition 1. *Sei* $X = \big(x_i, \nu_X(x_i)\big)_{i=1,\dots,m}$ *ein beliebiger Datensatz, bestehend aus β-Ketten des T-Zellrezeptorrepertoires, wobei die Anzahl der ausgelesenen Sequenzen (incl. Wiederholungen) mit* $M = \sum_{i=1}^{m} \nu_X(x_i)$ *gegeben ist. Der* Simpson Index *[8, 42] ist definiert als*

$$D(X) = 1 - \frac{1}{M(M-1)} \sum_{i=1}^{m} \nu_X(x_i) \cdot (\nu_X(x_i) - 1)$$

.

Die Interpretation des Simpson-Index lässt sich leicht anhand eines Urnenmodells veranschaulichen. Werden aus einem Datensatz X zwei Sequenzen, unter Berücksichtigung der Kopienzahl, ohne Zurücklegen entnommen, so entspricht $D(X)$ der Wahrscheinlichkeit, dass sich beide Sequenzen unterscheiden. $D(X)$ kann daher als Maß für

die Diversität des Datensatzes X angesehen werden. Die Grundidee dieses Vorgehens kann zur Definition zweier universell einsetzbarer Kennzahlen verwendet werden. Der entscheidende Unterschied zum Simpson-Index besteht darin, dass nicht die Wahrscheinlichkeit berechnet wird, dass sich die zwei gezogenen Sequenzen unterscheiden bzw. nicht unterscheiden, sondern solche Sequenzen zu ziehen, welche sich im Hinblick auf ein variables Kriterium ähneln. Die Möglichkeit, dieses Kriterium entsprechend anzupassen, erlaubt es, diese Kennzahlen auf eine Vielzahl verschiedener Fragestellungen anzuwenden. Im Folgenden bezeichne

- $R \subset \Omega \times \Omega$ eine beliebige symmetrische und reflexive Relation (d.h. ein binäres Kriterium wann zwei beliebige Sequenzen als *ähnlich* angesehen werden),
- $\mathbb{1}(\cdot)$ die Indikatorfunktion.

Zwei mögliche Ähnlichkeitskriterien, welche in dieser Arbeit zur Anwendung kommen, definieren beispielsweise zwei β-Ketten des T-Zellrezeptorrepertoires als ähnlich, falls sich ihre CDR3-Regionen um maximal eine Aminosäure unterscheiden oder ihnen identische V- bzw. J-Segmente zugeordnet wurden.

Definition 2. *Sei* $X = \big(x_i, \nu_X(x_i)\big)_{i=1,\dots,m}$ *ein beliebiger Datensatz, bestehend aus β-Ketten des T-Zellrezeptorrepertoires, wobei die Anzahl der ausgelesenen Sequenzen (incl. Wiederholungen) mit* $M = \sum_{i=1}^{m} \nu_X(x_i)$ *gegen ist. In Abhängigkeit der Relation* R *definieren wir*

- *den* gewichteten Repertoire-Homogenitätsindex *(engl. weighted Repertoire Homogeneity Index) als*

$$\mathrm{wRHI}_R(X) = \frac{1}{M\,(M-1)} \sum_{i,j=1,\dots,m} \nu_X\,(x_i)\,\big(\nu_X\,(x_j) - \mathbb{1}(i=j)\big)\mathbb{1}(x_i R x_j)$$

- *sowie den (ungewichteten)* Repertoire-Homogenitätsindex *(engl. Repertoire Homogeneity Index) als*

$$\mathrm{RHI}_R(X) = \frac{\displaystyle\sum_{\substack{i=1,\dots,m \\ j<i}} \mathbb{1}(x_i R x_j)}{\binom{m}{2}}$$

Bemerkung 1. *Werden aus einem Datensatz X zufällig zwei Sequenzen, unter Berücksichtigung der Kopienzahl, ohne Zurücklegen entnommen, so gibt* wRHI_R *die Wahrscheinlichkeit an, dass diese im Hinblick auf R in Relation stehen.* RHI_R *ergibt sich unmittelbar aus* wRHI_R *durch Vernachlässigung der Kopienzahl (d.h. jedem Klonotyp wird eine Kopienzahl von* 1 *zugeordnet). Hieraus ergeben sich unmittelbar die folgenden Eigenschaften:*

- $\mathrm{wRHI},\ \mathrm{RHI} \in [0,1]$,
- $\mathrm{RHI}(X) = 0 \Leftrightarrow \neg(x_i R x_j) \qquad \forall i,j = 1,\ldots,m,\ i \neq j$,
- $\mathrm{wRHI}(X) = 1 \Leftrightarrow \mathrm{RHI}(X) = 1 \Leftrightarrow x_i R x_j \qquad \forall i,j = 1,\ldots,m$,
- *Im Spezialfall, dass R die Identitätsrelation bezeichnet, ergibt sich*

$$\mathrm{wRHI}(X) = 1 - D(X).$$

In dieser Arbeit wurden die beiden Kennzahlen auf zwei verschiedene Ähnlichkeitskriterien angewendet. Das erste bezieht sich unmittelbar auf die CDR3-Aminosäuresequenz. Um dieses Kriterium aufstellen zu können ist zunächst eine weitere Definition erforderlich.

Definition 3. *Für beliebige Zeichenketten s_1 und s_2 gibt die* Levenshtein-Distanz *die Anzahl der Schreiboperationen an (Zeichen einfügen, entfernen, ersetzten), welche erforderlich ist, um s_1 in s_2 zu überführen [32].*

Als erstes Kriterium werden nun zwei Klonotypen als ähnlich definiert, falls die Levenshtein-Distanz zwischen den jeweiligen CDR3-Sequenzen einen Wert von 1 nicht überschreitet. Dies bedeutet, dass die CDR3-Sequenzen entweder übereinstimmen oder durch eine Schreiboperation ineinander überführt werden können. Das zweite Kriterium definiert, wie oben bereits erwähnt, diejenigen Sequenzen als ähnlich, denen sowohl identische V- als auch identische J-Segmente zugeordnet wurden. Im Folgenden werden die so konstruierten Kennzahlen mit $\mathrm{wRHI}_{\mathrm{LD}}$ und $\mathrm{RHI}_{\mathrm{LD}}$ bzw. $\mathrm{wRHI}_{\mathrm{VJ}}$ und $\mathrm{RHI}_{\mathrm{VJ}}$ bezeichnet.

2.3.2 Quantifizierung der Diversität multipler Nukleotidkodierungen für einzelne CDR3-Aminosäuresequenzen

Aufgrund der Degeneration des genetischen Codes kann die Aminosäuresequenz einer CDR3-Region, eines gegebenen Klonotyps, in einem Datensatz durch eine Vielzahl verschiedener Nukleotidsequenzen codiert sein. Die Diversität der Nukleotidkodierungen für einzelne CDR3-Aminosäuresequenzen kann mit Hilfe der folgenden Kennzahlen quantifiziert werden:

Definition 4. *Gegeben sei ein Datensatz* $X = \big(x_i, \nu_X(x_i)\big)_{i=1,\dots,m}$ *bestehend aus β-Ketten des T-Zellrezeptorrepertoires. Für jeden Klonotyp* $x \in X$ *bezeichnen* $x^{(1)}, \dots, x^{(m_x)}$ *die Nukleotidsequenzen aus der zugrundeliegenden Probe, welche für die CDR3-Aminosäuresequenz von* x *kodieren.*

- *Für einen gegebenen Klonotyp* $x \in X$ *definieren wir den Nukleotid-Simpson-Index als*

$$D_{Nuk}(x) = 1 - \sum_{i=1}^{m_x} \left(\frac{\nu_X(x^{(i)})}{\nu_X(x)} \right)^2 \quad \textit{sowie}$$

- *den* Diversitätsindex der Nukleotidkodierung *(engl. Coding Diversity Index) als*

$$\mathrm{CDI}(X) = \frac{1}{m} \sum_{i=1}^{m} D_{Nuk}(x_i).$$

Bemerkung 2.

- *Die Bedeutung von* D_{Nuk} *kann leicht mit Hilfe eines Urnenmodells veranschaulicht werden. Hierfür stelle man sich vor, dass die Nukleotidsequenzen, welche in einem gegebenen Datensatz* X *die CDR3-Aminosäuresequenz* $x \in X$ *kodieren, in einer Urne enthalten sind. Mehrfachdetektionen werden hierbei berücksichtigt. Werden nun zwei Sequenzen mit Zurücklegen entnommen, so bezeichnet* $D_{Nuk}(x)$ *die Wahrscheinlichkeit, dass sich diese Sequenzen unterscheiden.*

- *Für die Definition von* D_{Nuk} *wurde die Formel des Simpson-Index dahingehend modifiziert, dass die erste der gezogenen Sequenzen wieder in die Urne zurückge-*

legt wird. Hierdurch ist es möglich, diese Kennzahl auch auf Sequenzen anzuwenden, welche nur einmalig detektiert wurden ($CN = 1$). In dieser Arbeit werden solche Sequenzen allerdings nicht berücksichtigt (siehe Abschnitt 2.2).

- *Der Diversitätsindex der Nukleotidkodierung* CDI *entspricht einem ungewichteten Mittelwert der Nukleotid-Simpson-Indices. Die Kennzahl kann anhand eines mehrstufigen Zufallsexperimentes veranschaulicht werden. In einem ersten Schritt wird zufällig ein Klonotyp des Datensatzes X ausgewählt, wobei die Wahrscheinlichkeit ausgewählt zu werden für alle Klonotypen gleich ist. In einem zweiten Schritt werden aus den detektierten Nukleotidsequenzen, welche die zugehörige CDR3-Aminosäuresequenz kodieren zwei Sequenzen zufällig (mit Zurücklegen und unter Berücksichtigung von Mehrfachdetektionen, s.o.) entnommen.* $\mathrm{CDI}(X)$ *entspricht der Wahrscheinlichkeit bei diesem Experiment zwei verschiedene Nukleotidsequenzen zu erhalten. Die Kennzahl liefert somit ein Maß für die Diversität der Nukleotidkodierungen der CDR3-Aminosäuresequenzen innerhalb eines Datensatzes.*

2.3.3 Quantifizierung der Ähnlichkeit zweier verschiedener Repertoires

Die in Abschnitt 2.3.1 definierten Kennzahlen ermöglichen es, die Homogenität (bzw. die Ähnlichkeit) der Sequenzen innerhalb eines Datensatzes zu analysieren. Im Gegensatz hierzu dienen die in diesem Abschnitt definierten Indizes dazu, Ähnlichkeitsaspekte zwischen verschiedenen Datensätzen zu quantifizieren. Ausgangspunkt hierfür stellen etablierte Kennzahlen dar, mit denen das Auftreten identischer Klonotypen innerhalb zweier Datensätze quantifiziert werden kann.

Definition 5. *Es seien $X = \big(x_i, \nu_X(x_i)\big)_{i=1,\dots,m}$ und $Y = \big(y_i, \nu_Y(y_i)\big)_{i=1,\dots,n}$ beliebige Datensätze bestehend aus β-Ketten des T-Zellrezeptorrepertoires, wobei die Anzahl der ausgelesenen Sequenzen (incl. Wiederholungen) mit $M = \sum_{i=1}^{m} \nu_X(x_i)$ bzw. $N = \sum_{i=1}^{n} \nu_Y(y_i)$ gegen ist. Wir bezeichnen mit*

(i)

$$\mathrm{MH}(X,Y) = \frac{2 \sum\limits_{z \in X \cap Y} \nu_X(z)\nu_Y(z)}{N \cdot M\Big(\sum\limits_{x \in X} \frac{\nu_X(x)^2}{M^2} + \sum\limits_{y \in Y} \frac{\nu_Y(y)^2}{N^2}\Big)}$$

den Morisita-Horn Index *[19], mit*

(ii)

$$J(X,Y) = \frac{|X \cap Y|}{|X \cup Y|}$$

den Jaccard-Index [21] sowie mit

(iii)

$$S(X,Y) = \frac{2\,|X \cap Y|}{|X| + |Y|}$$

den Sørensen-Index *[21].*

Bemerkung 3.

- *Der Morisita-Horn Index* $\mathrm{MH}(X,Y)$ *stellt eine Maßzahl zur Quantifizierung von Ähnlichkeiten zweier Datensätze* X *und* Y *dar. Er ist genau dann gleich* 0 *wenn die Mengen der jeweils detektierten Klonotypen disjunkt sind und genau dann gleich* 1 *wenn es sich um identische Mengen handelt und der Anteil der einzelnen Klonotypen an der Gesamtzahl der ausgelesenen Sequenzen jeweils übereinstimmt [14], d.h. falls*

$$\frac{\nu_X(z)}{M} = \frac{\nu_Y(z)}{N} \qquad \forall z \in X \cup Y$$

- *Der Jaccard-Index stellt ebenfalls eine Maßzahl für die Ähnlichkeit zweier Datensätze dar. Er ist genau dann gleich* 0*, wenn die Mengen der jeweils detektierten Klonotypen disjunkt sind und genau dann gleich* 1 *wenn diese Mengen übereinstimmen. Im Gegensatz zum Morisita-Horn Index, werden die Kopienzahlen bei der Berechnung nicht berücksichtigt.*

- *Der Sørensen-Index entspricht in Anwendung und Interpretation weitgehend dem Jaccard-Index. Er wird hier nur erwähnt, da er als Spezialfall eines allgemeineren Index aufgefasst werden kann.*

Analog zu den in Abschnitt 2.3.1 definierten Homogenitätsindizes werden im Folgenden der Morisita-Horn Index bzw. der Sørensen-Index dahingehend verallgemeinert, dass nicht mehr das Auftreten identischer, sondern ähnlicher Klonotypen quantifiziert wird. Unter welchen Voraussetzungen zwei Klonotypen als ähnlich angesehen werden, wird formell über eine reflexive, symmetrische Relation R definiert.

Definition 6. *Es seien* $X = \big(x_i, \nu_X(x_i)\big)_{i=1,\dots,m}$ *und* $Y = \big(y_i, \nu_Y(y_i)\big)_{i=1,\dots,n}$ *zwei beliebige Datensätze, bestehend aus β-Ketten des T-Zellrezeptorrepertoires, wobei die Anzahl der ausgelesenen Sequenzen (incl. Wiederholungen) mit* $M = \sum_{i=1}^{m} \nu_X(x_i)$ *bzw.* $N = \sum_{i=1}^{n} \nu_Y(y_i)$ *gegen ist. In Abhängigkeit der Relation R definieren wir*

- *den* gewichteten Repertoire-Ähnlichkeitsindex *(engl. weighted Repertoire Similarity Index) zwischen X und Y als*

$$\mathrm{wRSI}_R(X,Y) = \frac{2 \sum\limits_{\substack{i=1,\dots,m \\ j=1,\dots,n}} \nu_X(x_i)\nu_Y(y_j)\mathbb{1}(x_i R y_j)}{NM\Big(\sum\limits_{i,j=1,\dots,m} \frac{\mathbb{1}(x_i R x_j)\nu_X(x_i)\nu_X(x_j)}{M^2} + \sum\limits_{i,j=1,\dots,n} \frac{\mathbb{1}(y_i R y_j)\nu_Y(y_i)\nu_Y(y_j)}{N^2} \Big)}$$

- *sowie den (ungewichteten)* Repertoire-Ähnlichkeitsindex *(engl. Repertoire Similarity Index) zwischen X und Y als*

$$\mathrm{RSI}_R(X,Y) = \frac{2 \sum\limits_{\substack{i=1,\dots,m \\ j=1,\dots,n}} \mathbb{1}(x_i R y_j)}{nm\Big(\sum\limits_{i,j=1,\dots,m} \frac{\mathbb{1}(x_i R x_j)}{m^2} + \sum\limits_{i,j=1,\dots,n} \frac{\mathbb{1}(y_i R y_j)}{n^2} \Big)}.$$

Proposition 1.

(i) $\mathrm{wRSI}_R,\ \mathrm{RSI}_R \geq 0.$

(ii) Im Spezialfall, dass R die Identitätsrelation bezeichnet, ergibt sich unmittelbar

$$\mathrm{wRSI}_R(X,Y) = \mathrm{MH}(X,Y) \quad \textit{und} \quad \mathrm{RSI}_R(X,Y) = S(X,Y).$$

(iii) Die Relation R ist genau dann transitiv (und damit eine Äquivalenzrelation), falls 1 *eine obere Schranke für einen (und damit für beide) der beiden Indizes darstellt, d.h. falls*

$$\mathrm{wRSI}_R\,,\ \mathrm{RSI}_R \leq 1$$

für beliebige Datensätze X und Y.

(iv) Falls

$$\frac{\sum_{i=1}^{m}\nu_X(x_i)\mathbb{1}(x_iRz)}{M} = \frac{\sum_{i=1}^{n}\nu_Y(y_i)\mathbb{1}(y_iRz)}{N} \qquad \forall z \in X \cup Y \tag{1}$$

so gilt $\mathrm{wRSI}_R(X,Y) = 1$. *Falls R transitiv ist, so gilt die Gegenrichtung ebenfalls.*

(v) Analog zu (iv) ergibt sich für den ungewichteten Repertoire-Ähnlichkeitsindex: Falls

$$\frac{\sum_{i=1}^{m}\mathbb{1}(x_iRz)}{m} = \frac{\sum_{i=1}^{n}\mathbb{1}(y_iRz)}{n} \qquad \forall z \in X \cup Y$$

so gilt $\mathrm{RSI}_R(X,Y) = 1$. *Falls R transitiv ist, so gilt die Gegenrichtung ebenfalls.*

Beweis. Es genügt die Aussagen für wRSI_R zu zeigen, für RSI_R kann der Beweis analog durchgeführt werden. Hierfür wird jedem Klonotyp eine willkürliche Kopienzahl von 1 zugeordnet. Die Aussagen (i) und (ii) sind trivial. Um (iii) zu zeigen, nehmen wir zunächst an, dass R transitiv und somit eine Äquivalenzrelation ist. Dies bedeutet, dass zwei Systeme von Äquivalenzklassen $\left(C(x_i')\right)_{i=1,\dots,m'}$ bzw. $\left(C(y_i')\right)_{i=1,\dots,n'}$ existieren, welche den folgenden Bedingungen genügen:

$$X = \bigcup_{i=1}^{m'} C(x_i'), \qquad Y = \bigcup_{i=1}^{n'} C(y_i')$$

und

$$\begin{aligned} C(x_i') \cap C(x_j') &= \emptyset \quad \forall i,j = 1,\dots,m',\ i \neq j, \\ C(y_i') \cap C(y_j') &= \emptyset \quad \forall i,j = 1,\dots,n',\ i \neq j, \end{aligned}$$

wobei $(x_i')_{i=1,\dots m'}$ und $(y_i')_{i=1,\dots n'}$ beliebige Repräsentanten der zugehörigen Äquivalenzklassen darstellen. Wir bezeichnen mit

$$M(x_k') = \sum_{i=1}^{m}\nu_X(x_i)\mathbb{1}\left(x_iRx_k'\right), \quad k = 1,\dots m' \quad \text{und}$$

$$N(y'_k) = \sum_{i=1}^{n} \nu_Y(y_i) \mathbb{1}\left(y_i R y'_k\right), \quad k = 1, \dots n'$$

die aufsummierten Kopienzahlen der detektierten Klonotypen aus der jeweiligen Äquivalenzklasse. Falls keine der Sequenzen in X in Relation zu einer Sequenz in Y steht, ist die Behauptung trivial. Andernfalls darf ohne Beschränkung der Allgemeinheit angenommen werden, dass ein Index $s \in \mathbb{N}$ existiert, welcher die folgende Bedingung erfüllt:

$$x'_i R y'_j \Leftrightarrow i, j \leq s \wedge i = j.$$

Andernfalls kann die Indizierung der Repräsentanten entsprechend modifiziert werden. Wir erhalten

$$\begin{aligned} 0 &\leq \sum_{i=1}^{s} \left(\frac{M(x'_i)}{M} - \frac{N(y'_i)}{N} \right)^2 \\ &= \sum_{i=1}^{s} \frac{M(x'_i)^2}{M^2} - \sum_{i=1}^{s} \frac{2M(x'_i)N(y'_i)}{MN} + \sum_{i=1}^{s} \frac{N(y'_i)^2}{N^2}. \end{aligned}$$

Addiert man den negativen Summanden auf beiden Seiten der Ungleichung, so ergibt sich

$$\frac{2}{NM} \sum_{i=1}^{s} M(x'_i)N(y'_i) \leq \sum_{i=1}^{s} \frac{M(x'_i)^2}{M^2} + \sum_{i=1}^{s} \frac{N(y'_i)^2}{N^2}.$$

Daraus folgt

$$\begin{aligned} \frac{2}{MN} \sum_{i=1}^{s} M(x'_i)N(y'_i) &= \frac{2}{MN} \sum_{k=1}^{s} \left(\sum_{i=1}^{m} \nu_X(x_i) \mathbb{1}\left(x_i R x'_k\right) \sum_{i=1}^{n} \nu_Y(y_i) \mathbb{1}\left(y_i R y'_k\right) \right) \\ &= \frac{2}{MN} \sum_{k=1}^{s} \sum_{\substack{i=1,\dots,m \\ j=1,\dots,n}} \nu_X(x_i)\nu_Y(y_j) \mathbb{1}(x_i R x'_k \wedge y_j R y'_k) \end{aligned}$$

$$
\begin{aligned}
&= \frac{2}{MN} \sum_{\substack{i=1,\dots,m \\ j=1,\dots,n}} \nu_X(x_i)\nu_Y(y_j)\mathbb{1}(x_i R y_j) \\
&\leq \sum_{i=1}^{s} \frac{M(x'_i)^2}{M^2} + \sum_{i=1}^{s} \frac{N(y'_i)^2}{N^2} \\
&\leq \sum_{i=1}^{m'} \frac{M(x'_i)^2}{M^2} + \sum_{i=1}^{n'} \frac{N(y'_i)^2}{N^2} \\
&= \sum_{k=1}^{m'} \frac{1}{M^2} \Big(\sum_{i=1}^{m} \nu_X(x_i)\mathbb{1}(x_i R x'_k) \Big)^2 \\
&+ \sum_{k=1}^{n'} \frac{1}{N^2} \Big(\sum_{i=1}^{n} \nu_Y(y_i)\mathbb{1}(y_i R y'_k) \Big)^2 \\
&= \sum_{k=1}^{m'} \frac{1}{M^2} \sum_{i,j=1,\dots,m} \nu_X(x_i)\nu_X(x_j)\mathbb{1}(x_i R x'_k \wedge x_j R x'_k) \\
&+ \sum_{k=1}^{n'} \frac{1}{N^2} \sum_{i,j=1,\dots,n} \nu_Y(y_i)\nu_Y(y_j)\mathbb{1}(y_i R y'_k \wedge y_j R y'_k) \\
&= \sum_{i,j=1,\dots,m} \frac{\nu_X(x_i)\nu_X(x_j)\mathbb{1}(x_i R x_j)}{M^2} \\
&+ \sum_{i,j=1,\dots,n} \frac{\nu_Y(y_i)\nu_Y(y_j)\mathbb{1}(y_i R y_j)}{N^2}.
\end{aligned}
$$

Hieraus ergibt sich

$$\frac{2}{MN}\sum_{\substack{i=1,\dots,m\\ j=1,\dots,n}} \nu_X(x_i)\nu_Y(y_j)\mathbb{1}(x_i R y_j) \leq \sum_{i,j=1,\dots,m} \frac{\nu_X(x_i)\nu_X(x_j)\mathbb{1}(x_i R x_j)}{M^2} + \sum_{i,j=1,\dots,n} \frac{\nu_Y(y_i)\nu_Y(y_j)\mathbb{1}(y_i R y_j)}{N^2}.$$

Dividiert man diese Ungleichung durch den Term auf der rechten Seite, so ergibt sich die Behauptung.

Falls R nicht transitiv ist, so existiert mindestens ein Tripel von Sequenzen $a, b, c \in \Omega$, welches die folgende Bedingung erfüllt.

$$aRb, \quad bRc \quad \text{und} \quad \neg(aRc). \tag{2}$$

Wir definieren die beiden Datensätze

$$X = \{(a,1),\ (c,1)\} \quad \text{und} \quad Y = \{(b,1)\} \tag{3}$$

und erhalten unmittelbar

$$\mathrm{wRSI}_R(X,Y) = \frac{4}{3} > 1.$$

Um (iv) zu zeigen, nehmen wir zunächst an, dass (1) erfüllt ist. Die Formel für den gewichteten Repertoire-Ähnlichkeitsindex kann folgendermaßen dargestellt werden:

$$\mathrm{wRSI}_R(X,Y) = \frac{\frac{2}{MN}\sum_{\substack{i=1,\dots,m\\ j=1,\dots,n}} \nu_X(x_i)\nu_Y(y_j)\mathbb{1}(x_i R y_j)}{\sum_{i=1}^{m}\sum_{j=1}^{m}\frac{\nu_X(x_i)\nu_X(x_j)\mathbb{1}(x_i R x_j)}{M^2} + \sum_{i=1}^{n}\sum_{j=1}^{n}\frac{\nu_Y(y_i)\nu_Y(y_j)\mathbb{1}(y_i R y_j)}{N^2}}$$

$$= \frac{\frac{2}{MN} \sum\limits_{\substack{i=1,\dots,m \\ j=1,\dots,n}} \nu_X(x_i)\nu_Y(y_j)\mathbb{1}(x_i R y_j)}{\sum\limits_{i=1}^{m}\left(\frac{\nu_X(x_i)}{M}\sum\limits_{j=1}^{m}\frac{\nu_X(x_j)\mathbb{1}(x_i R x_j)}{M}\right) + \sum\limits_{i=1}^{n}\left(\frac{\nu_Y(y_i)}{N}\sum\limits_{j=1}^{n}\frac{\nu_Y(y_j)\mathbb{1}(y_i R y_j)}{N}\right)}.$$

Setzt man (1) in beide Summanden des Nenners ein, so ergibt sich unmittelbar

$$\mathrm{wRSI}_R(X,Y) = \frac{\frac{2}{MN} \sum\limits_{\substack{i=1,\dots,m \\ j=1,\dots,n}} \nu_X(x_i)\nu_Y(y_j)\mathbb{1}(x_i R y_j)}{\frac{1}{MN}\left(\sum\limits_{i=1}^{m}\sum\limits_{j=1}^{n}\nu_X(x_i)\nu_Y(y_j)\mathbb{1}(x_i R y_j) + \sum\limits_{i=1}^{n}\sum\limits_{j=1}^{m}\nu_Y(y_i)\nu_X(x_j)\mathbb{1}(x_j R y_i)\right)}$$

$$= 1.$$

Für die Gegenrichtung nehmen wir an, dass R transitiv ist und $\mathrm{wRSI}(X,Y) = 1$. Hieraus folgt

$$\frac{2\sum\limits_{\substack{i=1,\dots,m \\ j=1,\dots,n}} \nu_X(x_i)\nu_Y(y_i)\mathbb{1}(x_i R y_j)}{MN} = \sum_{i,j=1,\dots,m} \frac{\nu_X(x_i)\nu_X(x_j)\mathbb{1}(x_i R x_j)}{M^2} + \sum_{i,j=1,\dots,n} \frac{\nu_Y(y_i)\nu_Y(y_j)\mathbb{1}(y_i R y_j)}{N^2}.$$

Unter Verwendung der im Beweis von (iii) eingeführten Notation, lässt sich diese Gleichung folgendermaßen darstellen:

$$2\sum_{i=1}^{s}\left(\frac{M(x_i')}{M}\cdot\frac{N(y_i')}{N}\right) = \sum_{i=1}^{s}\left(\frac{M(x_i')}{M}\right)^2 + \sum_{i=1}^{s}\left(\frac{N(y_i')}{N}\right)^2$$

$$+\sum_{i=s+1}^{m'}\left(\frac{M(x'_i)}{M}\right)^2+\sum_{i=s+1}^{n'}\left(\frac{N(y'_i)}{N}\right)^2.$$

Durch Umstellen der Gleichung ergibt sich

$$-\sum_{i=1}^{s}\left(\frac{M(x'_i)}{M}-\frac{N(x'_i)}{N}\right)^2=\sum_{i=s+1}^{m'}\left(\frac{M(x'_i)}{M}\right)^2+\sum_{i=s+1}^{n'}\left(\frac{N(y'_i)}{N}\right)^2. \tag{4}$$

Da der Term auf der linken Seite von (4) negativ, der rechte dagegen positiv ist, ergibt sich, dass beide Terme gleich 0 sind. Dies bedeutet

$$M(x'_i)=0 \quad \text{und} \quad N(y'_j)=0 \quad \forall i,j>s$$

sowie

$$\frac{M(x'_i)}{M}=\frac{N(y'_i)}{N} \qquad \forall i=1,\ldots,s.$$

Da für jede Sequenz $z \in X \cup Y$ genau einen Index $i \in \mathbb{N}$ existiert, für den $x'_i R z$ gegeben ist, folgt hieraus die Behauptung. Um zu zeigen, dass die vorausgesetzte Transitivität tatsächlich erforderlich ist, nehmen wir an, R sei nicht transitiv. Für ein beliebiges Tripel $a, b, c \in \Omega$ welches (2) erfüllt, definieren wir die Datensätze

$$X=\{(a,1),\ (c,1)\} \quad \text{und} \quad Y=\{(a,1),\ (b,1)\}.$$

Wie man leicht nachprüfen kann, gilt für diese Datensätze $\mathrm{wRSI}_R(X,Y)=1$. Dennoch ist (1) für $z=a$ nicht erfüllt. □

Bemerkung 4. *Auch die Bedeutung der beiden in Definition 6 eingeführten Kennzahlen lässt sich mit Hilfe von Urnenmodellen veranschaulichen. Dies wird im Folgenden am Beispiel der gewichteten Version aufgezeigt. Die Betrachtungen lassen sich leicht auf das ungewichtete Analogon übertragen, indem jedem Klonotyp eine willkürliche Kopienzahl von 1 zugeordnet wird. Wir betrachten hierzu noch einmal die in Definition*

6 eingeführten Datensätze X und Y. Wird aus jedem dieser Datensätze jeweils eine Sequenz (unter Berücksichtigung der Kopienzahl) entnommen, so ist die Wahrscheinlichkeit, dass diese bezüglich R in Relation zueinander stehen gegeben mit

$$p_1 = \frac{1}{NM} \sum_{\substack{i=1,\dots,m \\ j=1,\dots,n}} \nu_X(x_i)\nu_Y(y_j)\mathbb{1}(x_i R y_j).$$

In einem unabhängigen Zufallsexperiment wählt man zunächst einen der beiden Datensätze durch Werfen einer fairen Münze aus. Anschießend werden diesem Datensatz zwei Sequenzen nacheinander (mit Zurücklegen) entnommen. Die Wahrscheinlichkeit, hierbei zwei Sequenzen zu ziehen, welche bezüglich R in Relation zueinander stehen, ist gegeben mit

$$p_2 = \frac{1}{2}\Bigg(\sum_{i,j=1,\dots,m} \frac{\mathbb{1}(x_i R x_j)\nu_X(x_i)\nu_X(x_j)}{M^2} + \sum_{i,j=1,\dots,n} \frac{\mathbb{1}(y_i R y_j)\nu_Y(y_i)\nu_Y(y_j)}{N^2} \Bigg).$$

Bildet man den Quotienten der beiden Wahrscheinlichkeiten, so erhält man

$$\begin{aligned} \frac{p_1}{p_2} &= \frac{\frac{1}{NM} \sum_{\substack{i=1,\dots,m \\ j=1,\dots,n}} \nu_X(x_i)\nu_Y(y_j)\mathbb{1}(x_i R y_j)}{\frac{1}{2}\Bigg(\sum_{i,j=1,\dots,m} \frac{\mathbb{1}(x_i R x_j)\nu_X(x_i)\nu_X(x_j)}{M^2} + \sum_{i,j=1,\dots,n} \frac{\mathbb{1}(y_i R y_j)\nu_Y(y_i)\nu_Y(y_j)}{N^2} \Bigg)} \\ &= \mathrm{wRSI}_R(X, Y) \end{aligned}$$

Da es sich bei dieser Kennzahl nicht um eine Wahrscheinlichkeit, sondern um den Quotienten zweier Wahrscheinlichkeiten handelt, sind prinzipiell auch Werte über 1 möglich. Der Grund, warum dies nicht möglich ist, falls es sich bei R um eine transitive Relation handelt, wird in Abbildung 1 veranschaulicht. Wir betrachten zwei hypothetische Datensätze X und Y, welche jeweils aus den beiden Sequenzen $\{x_1, x_2\}$ bzw. $\{y_1, y_2\}$ bestehen. Die Kopienzahl der einzelnen Sequenzen sei jeweils als 1 angenommen. Falls nur die mit schwarzen Pfeilen verbundenen Sequenzen hinsichtlich R in

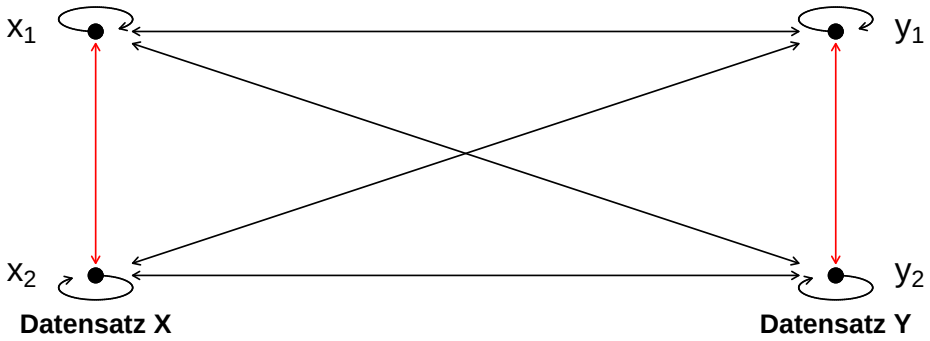

Abbildung 1: Falls dem (gewichteten) Repertoire-Ähnlichkeitsindex $\mathbf{wRSI}_R$ eine transitive Relation R zugrunde liegt, so ist 1 eine obere Schranke. Gegeben seien zwei Datensätze $X = \{x_1, x_2\}$ und $Y = \{y_1, y_2\}$, welche jeweils aus zwei Sequenzen bestehen. Die Kopienzahl der einzelnen Sequenzen sei jeweils als 1 angenommen. Falls lediglich die mit schwarzen Pfeilen verbundenen Sequenzen bzgl. R in Relation zueinander stehen, kann R nicht transitiv sein und es ergibt sich $\mathrm{wRSI}_R(X, Y) = 2$. Falls R jedoch transitiv ist, so implizieren die übrigen in Relation stehenden Sequenzen (schwarzen Pfeile) unmittelbar $x_1 R x_2$ und $y_1 R y_2$ (rote Pfeile). In diesem Fall ergibt sich $\mathrm{wRSI}_R(X, Y) = 1$.

Relation zueinander stehen, so kann R nicht transitiv sein und wir erhalten

$$p_1 = 1, \qquad p_2 = \frac{1}{2}$$

und damit

$$\mathrm{wRSI}_R(X, Y) = \frac{p_1}{p_2} = 2.$$

Wird jedoch die Transitivität von R vorausgesetzt, so wird durch die in Relation stehenden Sequenzen aus verschiedenen Datensätzen (lange schwarze Pfeile) unmittelbar impliziert, dass sich innerhalb der einzelnen Datensätze weitere Paare ähnlicher (d.h. in Relation stehender) Sequenzen befinden (rote Pfeile). Hierdurch erhöht sich die Wahrscheinlichkeit p_2 und $\mathrm{wRSI}_R(X, Y)$ sinkt auf 1 ab.

2.3.4 Partitionierung des T-Zellrezeptorrepertoires

Basierend auf den Kopienzahlen der jeweiligen Klonotypen wurde das T-Zellrezeptorrepertoire der einzelnen Tiere in Teilbereiche unterteilt, welche gesondert analysiert wurden. Hierdurch wird eine erhöhte Sensitivität gegenüber Immunisierungseffekten erzielt. Darüber hinaus ist es auf diese Weise möglich, Teilbereiche einzugrenzen, in welchen hohe Konzentrationen reagibler Klonotypen vermutet werden können. Zunächst wurde jeder Datensatz in Klonotypen von niedriger (CN = 2), mittlerer ($2 < \mathrm{CN} \leq 500$) und hoher (CN > 500) Kopienzahl unterteilt. Für jeden Zeitpunkt wurde das Repertoire der immunisierten Tiere ($n = 10$ für jeden Zeitpunkt) mit der konstanten Referenzgruppe der naiven Tiere ($n = 20$) im Hinblick auf die Anzahl der detektierten Klonotypen, die Anzahl ausgelesener Sequenzen (incl. Wiederholungen), die mittlere Länge der CDR3-Sequenzen, den Jaccard-Index sowie die neu eingeführten Kennzahlen $\mathrm{RHI_{LD}}$, $\mathrm{RHI_{VJ}}$ und CDI verglichen. Diese Kennzahlen zeichnen sich dadurch aus, dass sie alle Klonotypen des jeweiligen (Teil-)Datensatzes gleich stark gewichten. Homogenisierungs- bzw. Heterogenisierungseffekte, welche in Folge der Antigenapplikation innerhalb der einzelnen Teilbereiche auftreten, deuten auf einen hohen Anteil reagibler Klonotypen hin. Um derartige Ansammlungen innerhalb bestimmter Teilbereiche genauer eingrenzen zu können, wurden die Analysen analog mit einer feineren Unterteilung durchgeführt. Hierfür wurden die Datensätze systematisch in Untergruppen aufgeteilt. Für jeden Datensatz X wurden hierfür die Teilmengen

$$\begin{aligned} X^{(k)} &= \{x \in X : k-1 < \log_2(\mathrm{CN}) \leq k, \quad k = 1, \ldots, 9\} \quad \text{sowie} \\ X^{(10)} &= \{x \in X : \log_2(\mathrm{CN}) > 9\} \end{aligned}$$

betrachtet. Für jeden dieser 10 Teilbereiche wurden die immunisierten Tiere im Hinblick auf die oben genannten Parameter mit der Kontrollgruppe verglichen.

2.3.5 Gruppierung und Klassifikation von Datensätzen

Im Folgenden werden zwei statistische Lernverfahren definiert, welche es ermöglichen, Datensätze auf Basis variabler Kriterien zu gruppieren bzw. zu klassifizieren. Hierbei kommt ein nicht überwachter sowie ein überwachter Ansatz zur Anwendung. Während das nicht überwachte Verfahren (Clusteranalyse) dazu prädestiniert ist, unbekannte

Beziehungen zwischen den einzelnen T-Zellrezeptorrepertoires aufzudecken, teilt das überwachte Verfahren die einzelnen Datensätze in konkrete Kategorien (hier „immunisiert“ und „nicht immunisiert“) ein. Entscheidend ist hierbei, dass für eine korrekte Klassifikation die Immunisierungseffekte im einzelnen Individuum, nicht in der gesamten experimentellen Gruppe, ausschlaggebend sind. Zunächst muss ein abstrakter Abstandsbegriff (d.h. ein Maß für Verschiedenartigkeit) zwischen zwei Datensätzen definiert werden. Auf Basis der in Abschnitt 2.3.3 definierten Kennzahlen ist dies leicht möglich. Gegeben seien hierfür n symmetrische, reflexive Relationen $R_1, \ldots, R_n$ auf Ω sowie einen beliebigen Wichtungsvektor $\alpha = (\alpha_1, \ldots, \alpha_n)$ welcher die Voraussetzungen $\alpha_i \geq 0$, $i = 1, \ldots, n$ sowie $\sum_{i=1}^{n} \alpha_i = 1$ erfüllt. Für beliebige Datensätze X und Y stellen die Kennzahlen

- $d_{w,\alpha,R_1,\ldots,R_n}(X,Y) = 1 - \sum_{i=1}^{n} \alpha_i \min\big(\mathrm{wRSI}_{R_i}(X,Y), 1\big)$

- $d_{\alpha,R_1,\ldots,R_n}(X,Y) = 1 - \sum_{i=1}^{n} \alpha_i \min\big(\mathrm{RSI}_{R_i}(X,Y), 1\big)$

ein Distanzmaß [22, 28] dar. Dies bedeutet, dass die folgenden Voraussetzungen für beliebige Datensätze erfüllt sind:

- $d(X,Y) \geq 0$
- $d(X,X) = 0$
- $d(X,Y) = d(Y,X)$

Basierend auf diesen Distanzmaßen wurden die Datensätze mittels k-Medoid-Clusteranalyse [18, 35] in zwei Kategorien eingeteilt. Die Ergebnisse dieses nicht überwachten Klassifikationsverfahrens wurden mittels multidimensionaler Skalierung [17] visualisiert. Hierbei wird die abstrakte Distanz zwischen zwei Datensätzen durch Abstände von Punkten in einem zweidimensionalen Streudiagramm approximiert. Durch einfache Modifikationen kann dieses Verfahren in einen überwachten Klassifikationsalgorithmus überführt werden. Hierfür müssen die verfügbaren Datensätze zunächst in eine Trainings- und eine Testgruppe eingeteilt werden. Innerhalb der Trainingsgruppe wird jeder Datensatz mit dem zugehörigen Label „immunisiert“ bzw. „nicht immunisiert“ versehen (die drei immunisierten Gruppen werden hierbei zusammengefasst).

Anschließend wird das k-Medoid-Verfahren ($k = 1$) gesondert auf beide Teilgruppen der Trainingsdaten angewendet. Auf diese Weise wird für jede der beiden Teilgruppen genau ein Repräsentant (Medoid) bestimmt. Anschließend werden die Testdaten klassifiziert, indem jedem Datensatz das Label des nächstgelegenen Repräsentanten zugeordnet wird. Würde man die verfügbaren Datensätze in eine feste Trainings- bzw. Testgruppe einteilen, stünde ein nicht unerheblicher Teil der Datengrundlage nicht zur Anpassung des Lernverfahrens zur Verfügung, was zu einer erheblichen Verschlechterung der Trennschärfe führen kann. Um dieses Problem zu umgehen, wurde der Algorithmus im *Leave-one-out* Verfahren [55] angewendet. Dies bedeutet, dass jeder Datensatz im Hinblick auf den Immunisierungszustand klassifiziert wurde, nachdem das Lernverfahren auf die Gesamtheit der übrigen Datensätze angepasst wurde. Die Ergebnisse dieses Verfahrens wurden mit Hilfe eines exakten Fisher-Testes evaluiert. Hierbei wurde die Nullhypothese getestet, dass zwischen dem tatsächlichen Immunisierungsstatus und den vom Algorithmus zugeordneten Labeln kein Zusammenhang besteht. Dies wäre im Falle einer rein zufälligen Klassifikation gegeben. Die Grundlagen der hier angewendeten Klassifikationsverfahren sind ausführlich in [18] beschrieben.

Die Klassifikationsverfahren wurden zunächst auf Teilrepertoires angewendet, in denen, aufgrund signifikanter Diversifizierungs- bzw. Homogenisierungseffekte, hohe Konzentrationen reagibler Klonotypen vermutet werden können. Hierfür kamen zwei verschiedene Versionen von d zur Anwendung. Für die erste wurden die Übereinstimmung der detektierten V- und J-Segmente als unabhängige Kriterien gleicher Wichtung genutzt. Für die zweite Version wurden Sequenzen x und y als ähnlich (d.h. in Relation stehend) definiert, falls die Anzahl der Nukleotidkodierungen der CDR3 Sequenzen im jeweiligen Datensatz entweder übereinstimmen oder für beide Sequenzen einen Wert von 5 übersteigen. Ganze Datensätze wurden sowohl auf Basis von d (gleiche Wichtung der Klonotypen) als auch von d_w (Wichtung der Klonotypen nach Kopienzahl) klassifiziert. Auch hier wurde die Übereinstimmung der detektierten V- und J-Segmente als unabhängige Kriterien gleicher Wichtung genutzt.

2.3.6 Identifizierung reagibler *public*-Klonotypen

Obwohl die T-Zellreaktion gegen SRBC klar von individuellen Klonotypen dominiert wird (*private response)*, existieren daneben einige spezifische Klonotypen, welche in

den meisten der immunisierten Tiere expandieren. Zur Identifikation dieser Klonotypen, welche eine nachgeordnete *public*-Komponente der T-Zellreaktion repräsentieren, bietet sich eine Genexpressionsanalyse an [47, 51]. Bei dieser Analyse wurden lediglich solche Klonotypen berücksichtigt, welche in mindestens dreiviertel der immunisierten Tiere detektiert wurden. Die gesamte Genexpressionsanalyse wurde mit Hilfe des *edgeR*-Paketes [36] durchgeführt. Die mathematischen Grundlagen zu diesem Verfahren sind ausführlich in [9] beschrieben, daher können sich die Erklärungen an dieser Stelle auf eine kurze Zusammenfassung beschränken. Für jeden Datensatz X und jeden Klonotyp $x \in X$ wird angenommen, dass die Kopienzahl einer negativen Binomialverteilung folgt. Um Verdrängungseffekte durch einzelne hochexpandierte Klonotypen soweit wie möglich zu reduzieren, wurden in in dieser Arbeit die Gesamtzahlen der in den einzelnen Datensätzen ausgelesenen Sequenzen mit Hilfe des TMM-Verfahrens [37] (engl. *trimmed mean of M-values*) adjustiert. Mit Hilfe eines Likelihood-Quotienten-Testes wurde anschließend die Nullhypothese getestet, dass die zu erwartende Kopienzahl nicht durch die experimentellen Gegebenheiten (SRBC-Applikation, Zeitpunkt der Immunisierung) beeinflusst wird. Kann diese verworfen werden, ist davon auszugehen, dass der Klonotyp zumindest zu einem der drei Zeitpunkte auf die SRBC-Applikation reagiert hat. Die resultierenden p-Werte wurden mit Hilfe der FDR-Methode nach Benjamini-Hochberg [5] korrigiert. Um den Einfluss unspezifischer Proliferationsphänomene sowie PCR-Artefakte so weit wie möglich zu reduzieren, wurde die maximale Falscherkennungsrate (engl. false discovery rate, FDR) auf $0,005$ festgelegt [47]. Klonotypen, welche in allen immunisierten Gruppen unterexprimiert sind, wurden für die weiteren Analysen nicht berücksichtigt. Die auf diese Weise identifizierten Klonotypen stellen lediglich eine verschwindende Minderheit innerhalb der Gesamtmenge der expandierenden T-Zellen dar. Da ihr Expansionsverhalten in einer Vielzahl von Individuen parallel analysiert und verglichen werden kann, können sie jedoch bei der Analyse der T-Zellreaktion von besonderem Nutzen sein. In dieser Arbeit wurde die Verteilung dieser Klonotypen auf die 10, auf Basis der Kopienzahl definierten, Teilbereiche analysiert. Das Ergebnis wurde vor dem Hintergrund der Immunisierungseffekte, welche auf die Aktivität von *private*-Klonotypen schließen lassen, interpretiert.

2.3.7 Statistische Auswertung

Die T-Zellrezeptorrepertoires naiver und immunisierter Tiere wurden im Hinblick auf alle Parameter mit Hilfe des Mann-Whitney-U-Tests verglichen. Sofern nichts anderes angegeben, wurden zweiseitige Tests durchgeführt. Die Korrektur der errechneten p-Werte, im Hinblick auf Alphafehler-Kumulierung, erfolgte unter Verwendung der Holms-Prozedur. p-Werte unterhalb von $0,05$ (nach Korrektur) wurden als signifikant angesehen. Die Berechnung der Kennzahlen wRHI, RHI, wRSI und RSI wurde in Java durchgeführt. Die erforderlichen Programme wurden mit Hilfe der Entwicklungsumgebung *Eclipse IDE for Java Developers* erstellt. Zur Berechnung der Levenshtein-Distanz wurde die Bibliothek *Apache Commons Text* [2] genutzt. Alle übrigen Analysen sowie die Visualisierung der Ergebnisse wurden mit Hilfe der Statistikplattform R [34] durchgeführt. Das für die statistischen Lernverfahren erforderliche k-Medoid Verfahren wurde mit Hilfe Funktion *pam* aus dem R-Paket *cluster* [26] durchgeführt. Hierbei wurden die Standardparameter der Funktion verwendet. Man beachte, dass das gesamte Verfahren, einschließlich der initialen Bestimmung der Start-*medoids* (sog. *BUILD*-Phase), rein deterministischer Natur ist [39]. Da keine zufällige Initialisierung erforderlich ist, sind die Ergebnisse vollumfänglich reproduzierbar.

3 Ergebnisse

3.1 Innerhalb der dominanten Klonotypen können flüchtige Immunisierungseffekte nach SRBC-Applikation mit einfachen statistischen Parametern nachgewiesen werden

Alle Datensätze wurden in Klonotypen mit niedriger (CN = 2), mittlerer ($2 < \text{CN} \leq 500$) und hoher (CN > 500) Kopienzahl unterteilt. In der oberen Kategorie sind im Mittel ~ 100 Klonotypen enthalten. Diese Kategorie ist daher mit der gängigen Analyse der 100 Klonotypen mit höchster Kopienzahl vergleichbar [46, 50, 51]. Die Anzahl der insgesamt detektierten Klonotypen blieb nach Immunisierung mit SRBC unverändert. Die Anzahl der Klonotypen mit hoher Kopienzahl war am 3. Tag nach Antigenapplikation signifikant erhöht. Bereits einen Tag später war dieser Effekt deutlich abgeschwächt und nicht mehr signifikant. An Tag 7 waren, im Hinblick auf die Anzahl der detektierten Klonotypen, in keiner der Kategorien Abweichungen gegenüber den naiven Tieren feststellbar (Abbildung 2a). Ein analoger Verlauf zeigte sich bei der Betrachtung der detektierten Sequenzen (d.h. der aufsummierten Kopienzahlen) in den einzelnen Kategorien. Auch hier zeigte sich an Tag 3 eine signifikante Zunahme, welche bereits einen Tag später deutlich abgeschwächt und nicht mehr signifikant war (Abbildung 2b). Die mittlere Länge der detektierten CDR3-Sequenzen zeigte sich innerhalb der Klonotypen mit hoher Kopienzahl am 3. Tag nach Antigenapplikation leicht verkürzt. Auch dieser Effekt war bereits einen Tag später nicht mehr signifikant. Bei Betrachtung des Gesamtrepertoires waren, im Hinblick auf diesen Parameter keine Immunisierungseffekte feststellbar (Abbildung 2c). Der Jaccard-Index gibt die anteilige Schnittmenge der CDR3-Sequenzen aus zwei verschiedenen Datensätzen an. Zwischen den Gesamtrepertoires zeigte sich dieser Wert nach Immunisierung unverändert (Abbildung 2d). Zwischen den Klonotypen mit hoher Kopienzahl war dieser Wert an Tag 3 nach Antigenapplikation deutlich reduziert (in Abbildung 2d mit Pfeil markiert). Dies entspricht einer Divergenz der entsprechenden Teilrepertoires. In den einzelnen Tieren expandieren also in erster Linie individuelle Klonotypen (*private-response*). Auch dieser Effekt war einen Tag später bereits deutlich abgeschwächt und bildete sich bis Tag 7 vollständig zurück. Zu diesem Zeitpunkt war hingegen ein Abfall des Jaccard-Index zwischen Klonotypen mit niedriger Kopienzahl festzustellen (in Abbildung 2d mit Pfeil markiert). Der Jaccard-Index wurde hierbei jeweils für alle möglichen Paa-

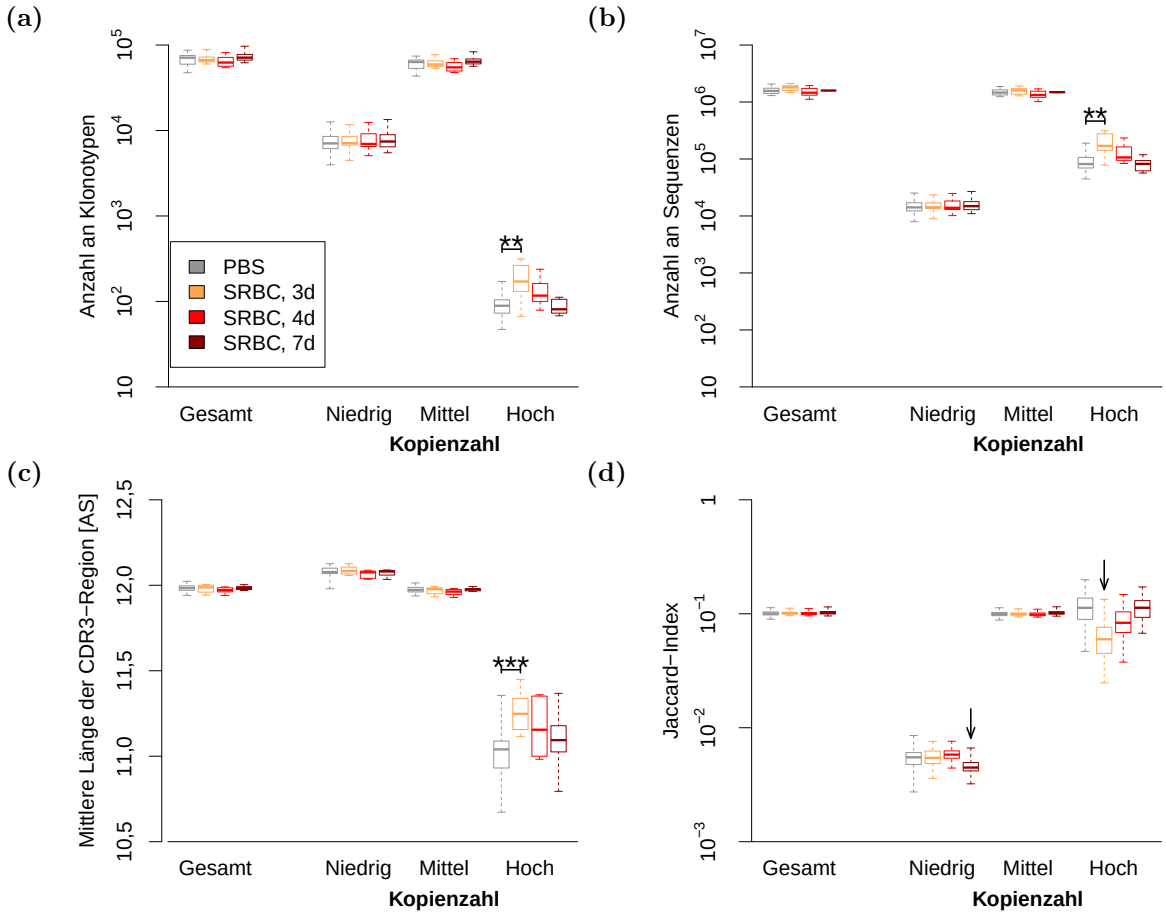

Abbildung 2: Nach SRBC-Applikation können flüchtige Expansions- und Divergenzeffekte innerhalb des T-Zellrezeptorrepertoires mit einfachen statistischen Parametern nachgewiesen werden. Jeder Datensatz wurde in Klonotypen mit niedriger (CN = 2), mittlerer ($2 < \text{CN} \leq 500$) und hoher (CN > 500) Kopienzahl unterteilt. Für jedes der so definierten Teilrepertoires, sowie für die gesamten Datensätze, wurden die immunisierten Tiere ($n = 10$ für jeden der drei Zeitpunkte) mit der Kontrollgruppe (PBS-Injektion, $n = 20$) verglichen. (a) Anzahl der detektierten Klonotypen, (b) Anzahl der ausgelesenen Sequenzen (incl. Mehrfachdetektionen), (c) mittlere Länge der CDR3-Region,(d) Jaccard-Index. Die Boxplot-Diagramme zeigen Median, Quartilsabstand sowie minimale/maximale Messwerte. Für (a), (b) und (c) wurden die drei immunisierten Gruppen mittels Mann-Whitney-U-Test mit der Kontrollgruppe verglichen. Die Adjustierung der p-Werte wurde mit Hilfe der Holm-Prozedur vorgenommen. Die adjustierten p-Werte sind folgendermaßen dargestellt: * $p < 0,05$, ** $p < 0,01$,*** $p < 0,001$. Für die Betrachtung des Gesamtrepertoires wurde diese Korrektur jeweils unabhängig von den drei Teilbetrachtungen durchgeführt. Mit Rücksicht auf multiple statistische Abhängigkeiten zwischen den einzelnen Werten wurde für (d) kein Test durchgeführt. Die Pfeile verweisen auf einen markanten Abfall des Jaccard-Index am 3. und 7. Tag nach Antigenapplikation. Abbildungen (a), (c) und (d) modifiziert nach [29].

rungen berechnet (jeder Datensatz wurde mit allen anderen Datensätzen der jeweiligen experimentellen Gruppe verglichen). Hieraus ergeben sich multiple statistische Abhängigkeiten zwischen den einzelnen Werten. Die Anwendung induktiver statistischer Methoden (Konfidenzintervalle, statistische Tests) wird hierdurch erheblich erschwert. Aus diesem Grund beschränkt sich die Analyse des Jaccard-Index in dieser Arbeit auf rein deskriptive Betrachtungen.

Zusammenfassend lässt sich feststellen, dass sich mit diesen einfachen Parametern (abgesehen von einem leichten Absinken des Jaccard-Index zwischen den Teilrepertoires mit niedriger Kopienzahl) lediglich Immunisierungseffekte detektieren lassen, welche zu Beginn der T-Zellreaktion im Bereich der dominanten Klonotypen auftreten. Bereits 4 Tage nach Antigenapplikation sind keine signifikanten Immunisierungseffekte mehr feststellbar. Da die Immunreaktion zu diesem Zeitpunkt jedoch fortdauert [44, 51] müssen sich im T-Zellrezeptorrepertoire weiterhin Immunisierungseffekte detektieren lassen. Hierfür sind offensichtlich sensitivere Parameter erforderlich.

3.2 Die SRBC-spezifische T-Zellantwort führt über einen Zeitraum von 7 Tagen zu einer Diversifizierung des Rezeptorrepertoires sowie zu einer Homogenisierung der Nukleotidkodierung der CDR3-Sequenzen

Die in Abschnitt 2.3.1 eingeführten Kennzahlen ermöglichen es, die Diversität bzw. Homogenität des T-Zellrezeptorrepertoires im Hinblick auf variable biologische Parameter zu quantifizieren. Der Repertoire-Homogenitätsindex im Hinblick auf die Levenshtein-Distanz $\mathrm{RHI}_{\mathrm{LD}}$ bleibt bei Betrachtung des Gesamtrepertoires unverändert (Abbildung 3a). Betrachtet man lediglich Klonotypen mit hoher Kopienzahl, zeigt sich an Tag 3 nach Antigenapplikation ein signifikanter Abfall dieses Wertes. Dies bedeutet, dass unter diesen Klonotypen weniger Paare mit ähnlichen CDR3-Sequenzen (d.h. solche deren Levenshtein-Distanz einen Wert von 1 nicht überschreitet) zu finden sind. Dies entspricht einem Divergenzeffekt zwischen den CDR3-Sequenzen innerhalb der Repertoires der einzelnen Tiere. Im gewissen Sinne kann dieser Effekt als ein mikroskopisches Analogon zur makroskopischen Divergenz der Repertoires unterschiedlicher Tiere aufgefasst werden, welcher sich am Abfall des Jaccard-Index manifestiert. 4 Tage nach Antigenapplikation ist dieser Divergenzeffekt nur noch andeutungsweise vorhanden und nicht mehr signifikant. An Tag 7 nach Antigenapplikation zeigt sich ein signifikanter Abfall

von RHI_{LD} im Bereich der Klonotypen mit niedriger Kopienzahl (hier entspricht die zeitliche Dynamik ebenfalls der des Jaccard-Index). Der Repertoire-Homogenitätsindex im Hinblick auf die detektierten V- bzw. J-Segmente RHI_{VJ} zeigt an Tag 7 einen signifikanten Abfall innerhalb des Gesamtrepertoires (Abbildung 3b), was auf eine Diversifizierung der Segmente schließen lässt. Bei Betrachtung der drei Teilbereiche zeigen sich im Bereich der hohen Kopienzahlen an Tag 3 und 4 signifikante Diversifizierungseffekte. Selbiges gilt an Tag 7 für Klonotypen mit mittlerer und niedriger Kopienzahl. Der Diversitätsindex der Nukleotidkodierung CDI dient dazu, die Diversität der Nukleotidkodierungen der CDR3-Aminosäuresequenzen innerhalb eines Datensatzes zu quantifizieren. Bei Betrachtung des Gesamtrepertoires zeigt diese Kennzahl keine signifikanten Immunisierungseffekte an (Abbildung 3c). Angewendet auf Klonotypen mit hoher Kopienzahl zeigt sich ein signifikanter Abfall von CDI an Tag 3 und 4 nach Antigenapplikation. Dies zeigt eine Homogenisierung der Nukleotidkodierungen der detektierten CDR3-Aminosäuresequenzen an. Ein analoger, hochsignifikanter Effekt zeigt sich an Tag 7 im Bereich der Klonotypen mit niedriger Kopienzahl.

Man beachte, dass von allen der hier betrachteten Kennzahlen, lediglich mittels RHI_{VJ} Immunisierungseffekte im Bereich der mittleren Kopienzahlen sowie im Gesamtrepertoire (ohne Berücksichtigung der Kopienzahl) detektiert werden können. Durch Anwendung des gewichteten Repertoire-Homogenitätsindex wRHI kann der Fokus einer Analyse auf die dominanten (mutmaßlich expandierenden) Klonotypen gelegt werden, ohne das Repertoire in einzelne Teilbereiche zu unterteilen. Dies kann bei Betrachtung des Gesamtrepertoires zu einer deutlichen Verbesserung der Trennschärfe führen. $wRHI_{LD}$ steigt nach Antigenapplikation an (Abbildung 4a). Allerdings ist dieser Effekt nur an Tag 4 signifikant und an Tag 7 bereits nicht mehr feststellbar. Hierbei ist zu beachten, dass eine hinreichend ausgeprägte Expansion einzelner Klone immer zu einem Anstieg von wRHI (im Hinblick auf ein beliebiges Kriterium) führen kann. Aufgrund der Expansion erhöht sich die Wahrscheinlichkeit, dass zwei identische Sequenzen gezogen werden. Aufgrund der vorausgesetzten Reflexivität der zugrundeliegenden Relation erfüllen zwei identische Sequenzen per Definition das zugrundeliegende Ähnlichkeitskriterium. Auf diese Weise erklärt sich, dass $wRHI_{LD}$, ungeachtet der Divergenzeffekte zwischen den CDR3-Sequenzen, durch die Immunisierung ansteigt. $wRHI_{VJ}$ ist in jedem der drei Zeitpunkte signifikant vermindert (Abbildung 4b). Dies deckt sich mit der o.g. Diversifizierung der detektierten V- bzw. J-Segmente.

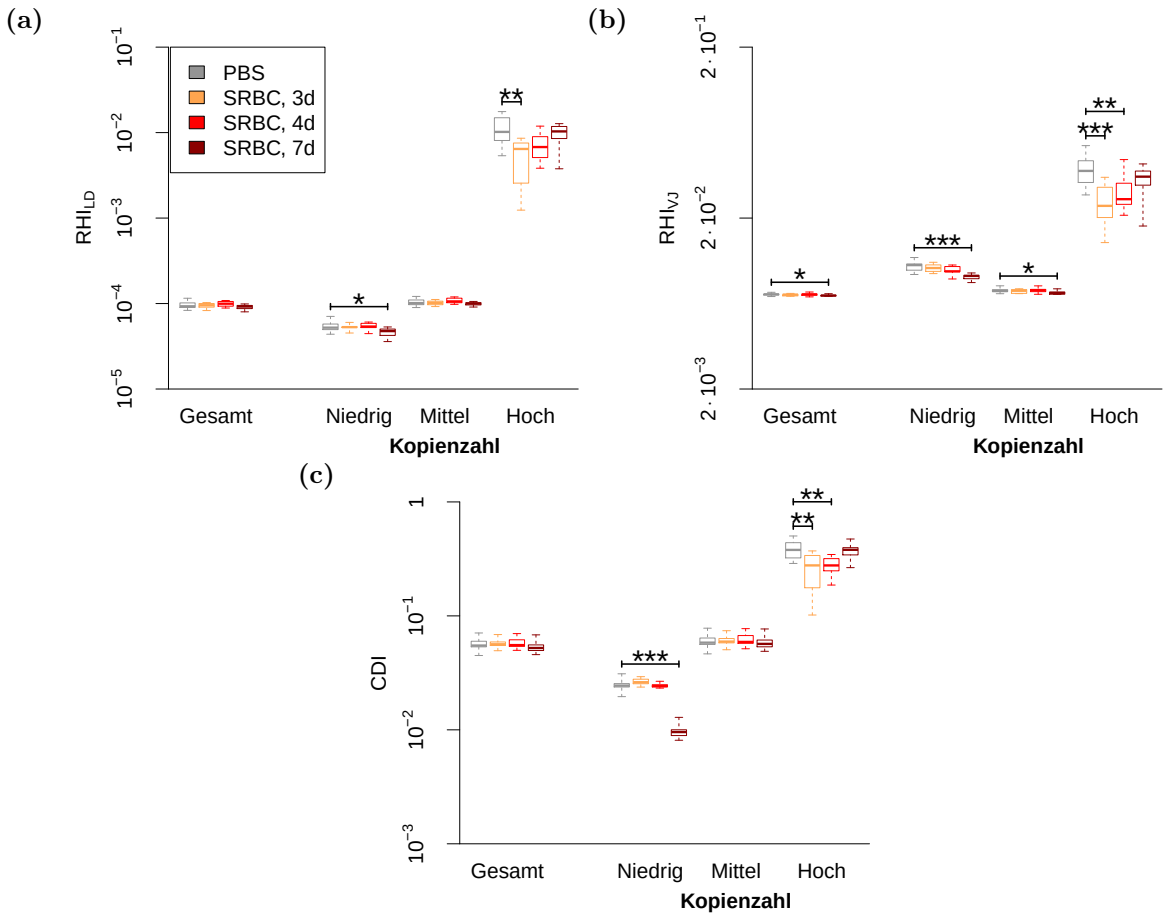

Abbildung 3: Die SRBC-spezifische T-Zellantwort führt über einen Zeitraum von 7 Tagen zu einer Diversifizierung des Rezeptorrepertoires sowie zu einer Homogenisierung der Nukleotidkodierung der CDR3-Sequenzen. Jeder Datensatz wurde in Klonotypen mit niedriger (CN $= 2$), mittlerer ($2 <$ CN ≤ 500) und hoher (CN > 500) Kopienzahl unterteilt. Für jedes der so definierten Teilrepertoires, sowie für die gesamten Datensätze wurden die immunisierten Tiere ($n = 10$ für jeden der drei Zeitpunkte) mit der Kontrollgruppe (PBS-Injektion, $n = 20$) verglichen. (a) Repertoire-Homogenitätsindex im Hinblick auf die Levenshtein-Distanz ($\mathrm{RHI_{LD}}$; zwei Klonotypen stehen definitionsgemäß in Relation, falls die Levenshtein-Distanz zwischen den zugehörigen CDR3-Aminosäuresequenzen einen Wert von 1 nicht übersteigt). (b) Repertoire-Homogenitätsindex im Hinblick auf V- und J-Segmente ($\mathrm{RHI_{VJ}}$; zwei Klonotypen stehen definitionsgemäß in Relation, falls ihnen identische V- und J-Segmente zugeordnet wurden). (c) Der Homogenitätsindex der Nukleotidkodierung CDI ermöglicht eine Quantifizierung der Diversität der Nukleotidkodierungen der einzelnen CDR3-Aminosäuresequenzen. Die Boxplot-Diagramme zeigen Median, Quartilsabstand sowie minimale/maximale Messwerte. Die drei immunisierten Gruppen wurden jeweils mittels Mann-Whitney-U-Test mit der Kontrollgruppe verglichen. Die Adjustierung der p-Werte wurde mit Hilfe der Holm-Prozedur vorgenommen. Die adjustierten p-Werte sind folgendermaßen dargestellt: * $p < 0,05$, ** $p < 0,01$,*** $p < 0,001$. Für die Betrachtung des Gesamtrepertoires wurde diese Korrektur jeweils unabhängig von den drei Teilbetrachtungen durchgeführt. Abbildungen modifiziert nach [29].

Die bisherigen Ergebnisse legen den Schluss nahe, dass an den Tagen 3 und 4 nach Antigenapplikation schwerpunktmäßig Klonotypen mit hoher Kopienzahl in die T-Zellreaktion involviert sind. Bis Tag 7 verlagert sich der Schwerpunkt in den Bereich der Klonotypen mit mittlerer und niedriger Kopienzahl. Um solche Schwerpunktbereiche genauer eingrenzen zu können, muss das gesamte T-Zellrezeptorrepertoire systematisch unterteilt und analysiert werden.

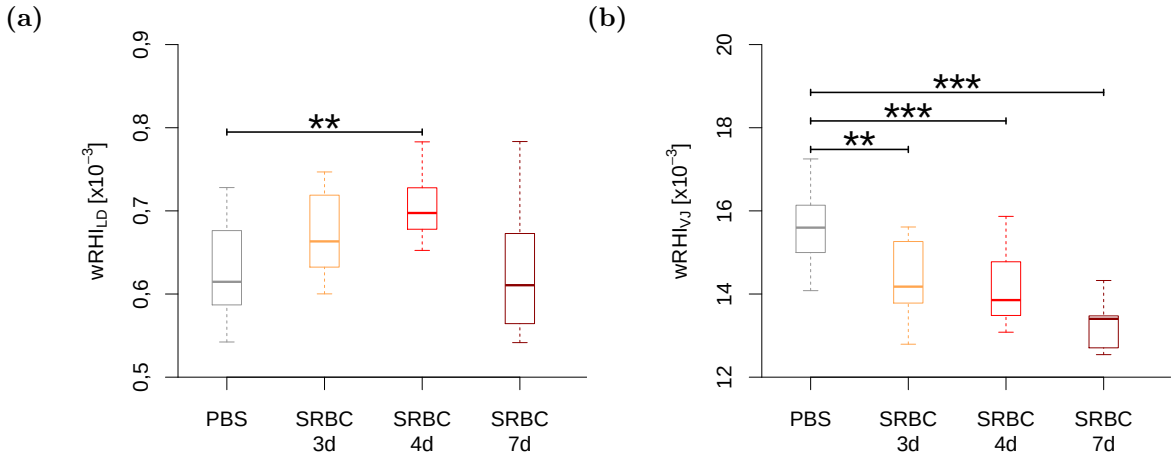

Abbildung 4: Bei der Analyse des Gesamtrepertoires kann es hilfreich sein, die Klonotypen nach Häufigkeit (Kopienzahl) zu gewichten. Hierdurch werden expandierende Klone stärker in den Fokus gerückt. Für jeden der drei Zeitpunkte wurden die immunisierten Tiere ($n = 10$ für jeden der Zeitpunkte) mit der Kontrollgruppe (PBS-Injektion, $n = 20$) verglichen. (a) Gewichteter Repertoire-Homogenitätsindex im Hinblick auf die Levenshtein-Distanz ($\mathrm{wRHI_{LD}}$; zwei Klonotypen stehen definitionsgemäß in Relation, falls die Levenshtein-Distanz zwischen den zugehörigen CDR3-Aminosäuresequenzen einen Wert von 1 nicht übersteigt). (b) Gewichteter Repertoire-Homogenitätsindex im Hinblick auf detektierte V- bzw. J-Segmente ($\mathrm{wRHI_{VJ}}$; zwei Klonotypen stehen definitionsgemäß in Relation, falls ihnen identische V- und J-Segmente zugeordnet wurden). Die Boxplot-Diagramme zeigen Median, Quartilsabstand sowie minimale/maximale Messwerte. Die drei immunisierten Gruppen wurden jeweils mittels Mann-Whitney-U-Test mit der Kontrollgruppe verglichen. Die Adjustierung der p-Werte wurde mit Hilfe der Holm-Prozedur vorgenommen. Die adjustierten p-Werte sind folgendermaßen dargestellt: * $p < 0,05$, ** $p < 0,01$,*** $p < 0,001$.

3.3 Durch systematische Aufteilung des T-Zellrezeptorrepertoires können Schwerpunktbereiche identifiziert werden, in denen sich reagible Klonotypen konzentrieren

Die T-Zellrezeptorrepertoires der einzelnen Tiere wurden in jeweils 10 Kategorien eingeteilt (Abbildung 5). Diese Einteilung basiert auf dem Zweierlogarithmus der Kopienzahl. Klontypen mit $\log_2(\mathrm{CN}) > 9$ wurden zu einer Untergruppe zusammengefasst.

Man beachte, dass die oberste der 10 Kategorien ($\log_2(\mathrm{CN}) > 9$) näherungsweise diejenigen Klonotypen beinhaltet, deren Kopienzahl, in den vorangegangenen beiden Abschnitten, als hoch bezeichnet wurde. Die unterste Kategorie ($\log_2(\mathrm{CN}) = 1$) entspricht dagegen den Klonotypen mit niedriger Kopienzahl. Der entscheidende Unterschied zu der im vorherigen Abschnitt gewählten Dreiteilung der Datensätze besteht in einer feineren Unterteilung der Klonotypen mit mittlerer Kopienzahl ($2 < \mathrm{CN} \leq 500$). Die minimale Anzahl an Klonotypen innerhalb eines einzelnen Datensatzes, welche einer der 10 Kategorien zugeordnet wurden, beträgt 45. Im Hinblick auf die Anzahl der Klonotypen und ausgelesenen Sequenzen innerhalb einer Kategorie zeigt sich lediglich an Tag 3 nach Antigenapplikation ein signifikanter Immunisierungseffekt in der obersten Kategorie ($\log_2(\mathrm{CN}) > 9$; Abbildung 5a-b). Die mittlere Länge der ausgelesenen CDR3-Sequenzen zeigt sich lediglich an Tag drei in der obersten, an Tag 4 in der zweitobersten Kategorie ($8 < \log_2(\mathrm{CN}) \leq 9$) leicht verkürzt (Abbildung 5c). Bei der Betrachtung des Jaccard-Index zeigt sich hingegen bereits ein deutlicher Vorteil der feineren Unterteilung. Die im vorherigen Abschnitt beschriebenen Divergenzeffekte zwischen den T-Zellrezeptorrepertoires sind an Tag 3 und 4 in allen Kategorien mit $\log_2(\mathrm{CN}) > 6$ feststellbar und erstrecken sich somit weit in den mittleren Bereich (Abbildung 5d).

Die Kombination der logarithmischen Unterteilung des extrahierten T-Zellrezeptorrepertoires mit den in Abschnitt 2.3.1 definierten Kennzahlen ermöglichen sowohl die Detektion weiterer Immunisierungseffekte als auch eine relativ genaue Beobachtung der zeitlichen Dynamik der T-Zellreaktion. Für jede der 10 Kategorien wurden hierfür die T-Zellrezeptorrepertoires der immunisierten Tiere im Hinblick auf $\mathrm{RHI}_{\mathrm{LD}}$, $\mathrm{RHI}_{\mathrm{VJ}}$ sowie CDI mittels Mann-Whitney-U-Tests mit denen der Kontrollgruppe verglichen und die resultierenden p-Werte in einer Heatmap visualisiert (Abbildung 6). Um die Interpretation der p-Werte zu erleichtern, wurden diese Tests einseitig durchgeführt. Ein signifikantes Testergebnis zeigt hierbei ein Absinken des zugehörigen Parameters an. Mit allen drei Kennzahlen können an Tag 3 und 4 nach Antigenapplikation Immunisierungseffekte im Bereich der Klonotypen mit hoher Kopienzahl detektiert werden (Abbildung 6a). Bis zum 7. Tag bildet sich ein weiterer Schwerpunkt im Bereich der Sequenzen mit niedriger Kopienzahl heraus. Von den drei Parametern weist $\mathrm{RHI}_{\mathrm{LD}}$ die geringste Trennschärfe auf. Daher können auf Basis dieses Wertes nur in wenigen Teilbereichen Immunisierungseffekte detektiert werden, das Signifikanzniveau ist hierbei

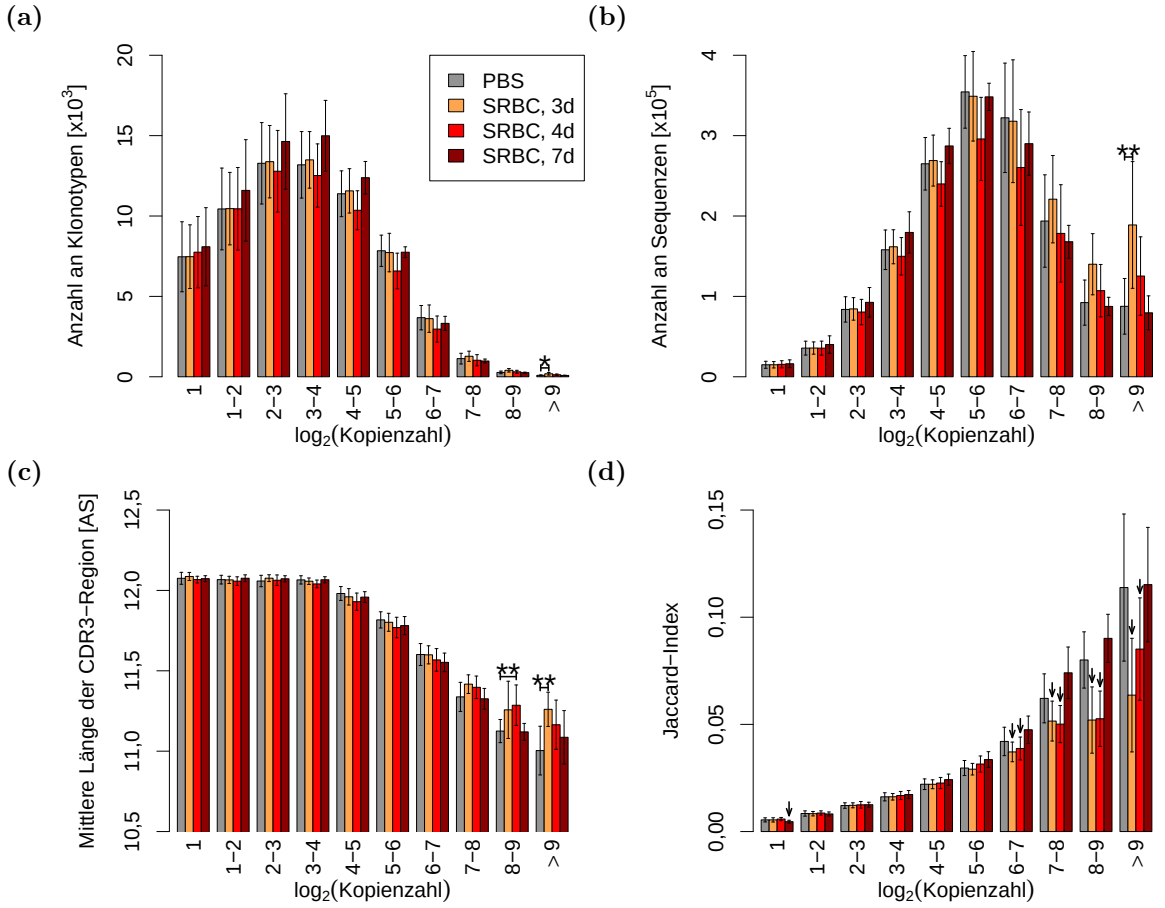

Abbildung 5: Durch systematische Unterteilung des T-Zellrezeptorrepertoires können Immunisierungseffekte in weiten Teilen detektiert werden. Anhand der Kopienzahlen der einzelnen Klonotypen wurde jeder Datensatz in 10 Kategorien eingeteilt. (a) Anzahl der detektierten Klonotypen, (b) Anzahl der ausgelesenen Sequenzen (incl. Mehrfachdetektionen), (c) mittlere Länge der CDR3-Region, (d) Jaccard-Index. Die Barplots zeigen die Mittelwerte, die Fehlerbalken die zugehörigen Standardabweichungen an. Für (a)-(c) wurden die Gruppen der immunisierten Tiere ($n = 10$ für jeden Zeitpunkt) mit der Kontrollgruppe (PBS-Injektion, $n = 20$) mittels Mann-Whitney-U-Test verglichen. Die Adjustierung der p-Werte wurde mit Hilfe der Holm-Prozedur vorgenommen. Die adjustierten p-Werte sind folgendermaßen dargestellt: * $p < 0,05$, ** $p < 0,01$,*** $p < 0,001$. Mit Rücksicht auf multiple statistische Abhängigkeiten zwischen den einzelnen Werten wurde für (d) kein Test durchgeführt. Die Pfeile verweisen auf markante Abfälle des Jaccard-Index. Abbildungen (a), (c) und (d) modifiziert nach [29].

vergleichsweise niedrig. Dagegen lassen sich mit Hilfe von RHI_{VJ} in weiten Teilen des T-Zellrezeptorrepertoires Immunisierungseffekte detektieren. Im Unterschied zu den beiden anderen Parametern zeigt dieser Index bereits an Tag 4 dezente Immunisierungseffekte in Teilbereichen mit vergleichsweise niedrigen Kopienzahlen ($1 < \log_2(\mathrm{CN}) \leq 2$) an. Zudem ist es mittels RHI_{VJ} möglich, Immunisierungseffekte unter Klonotypen mit relativ hoher Kopienzahl ($6 < \log_2(\mathrm{CN}) \leq 9$) bis einschließlich Tag 7 zu detektieren. Betrachtet man ausschließlich den oberen Teil des Repertoires, so verschiebt sich der Bereich, in welchem das höchste Signifikanzniveau erreicht wird, von der obersten Kategorie ($\log_2(\mathrm{CN}) > 9$) ins obere Mittelfeld ($7 < \log_2(\mathrm{CN}) \leq 8$). Im Bereich der Klonotypen mit $\log_2(\mathrm{CN}) > 9$ sind an Tag 7 keine signifikanten Immunisierungseffekte mehr feststellbar. Im Bereich der niedrigen Kopienzahlen erreicht CDI an Tag 7 für eine der Kategorien ($1 < \log_2(\mathrm{CN}) \leq 2$) ein höheres Signifikanzniveau als RHI_{VJ}. Dies kann ein Hinweis auf eine bessere Trennschärfe in diesem Bereich sein. Für alle Parameter verbleibt im Bereich der mittleren Kopienzahlen ein Residualbereich ($4 < \log_2(\mathrm{CN}) \leq 6$) in welchem zu keinem Zeitpunkt Immunisierungseffekte detektierbar sind. Selbst wenn die errechneten p-Werte nicht im Hinblick auf Alphafehler-Kumulierung korrigiert werden, sind in diesem Bereich keine signifikanten Unterschiede zwischen den immunisierten Tieren und denen der Kontrollgruppe feststellbar. Dies bedeutet, dass hier keine signifikanten Ergebnisse im Rahmen der Adjustierung verloren gegangen sind (Abbildung 6b). Zusammenfassend lässt sich sagen, dass in zwei Bereichen, zu bestimmten Zeitpunkten, Ansammlungen reagibler Klonotypen vermutet werden können. Im Bereich der Klonotypen deren Kopienzahl im Intervall ($6 < \log_2(\mathrm{CN}) \leq 9$) liegt, sind in allen immunisierten Gruppen, in jeweils allen Teilbereichen, signifikante Effekte feststellbar. Bis Tag 7 bildet sich ein zweiter Schwerpunkt der Immunreaktion unter Klonotypen mit relativ niedriger Kopienzahl heraus. Insbesondere zeigt sich im Bereich der Klonotypen mit $\log_2(\mathrm{CN}) \leq 2$ eine hochsignifikante Homogenisierung der Nukleotidkodierungen der CDR3-Sequenzen.

Die Entstehung des zweiten Schwerpunktes im Bereich der Klonotypen mit $\mathrm{CN} \leq 4$ (d.h. $\log_2(\mathrm{CN}) \leq 2$), an Tag 7 nach Antigenapplikation, könnte sich auf verschiedene Ursachen zurückführen lassen. Denkbar wären sowohl die Wiedereinwanderung von Klonotypen, welche ursprünglich in anderen Teilen der Milz expandiert sind, als auch lokale Effekte. Zu Letzteren zählen die direkte Einwanderung von Klonotypen, welche in unmittelbarer Nähe expandiert sind, aber auch die schwache, zeitversetzte Expan-

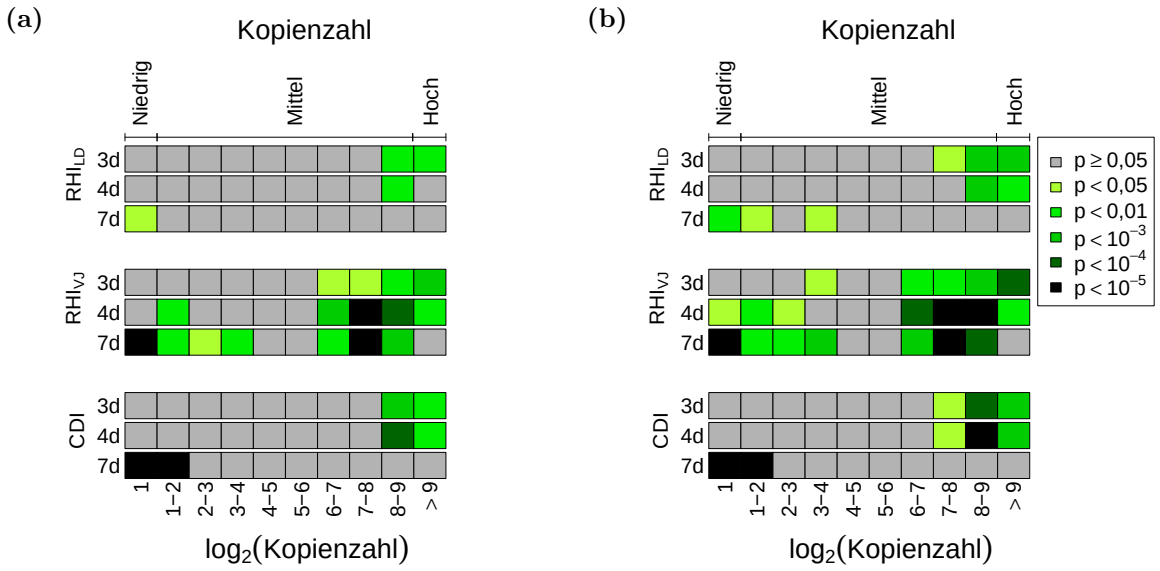

Abbildung 6: Durch systematische Aufteilung des T-Zellrezeptorrepertoires können Schwerpunktbereiche identifiziert werden, in denen sich reagible Klonotypen konzentrieren. Anhand der Kopienzahlen der einzelnen Klonotypen wurde jeder Datensatz in 10 Kategorien eingeteilt. Für jede der Kategorien wurden die Gruppen der immunisierten Tiere im Hinblick auf $\mathrm{RHI_{LD}}$, $\mathrm{RHI_{VJ}}$ sowie CDI mit der Kontrollgruppe verglichen (a). Die drei Kennzahlen zeigen zu jedem der drei Zeitpunkte signifikante Immunisierungseffekte in bestimmten Teilbereichen an. Die p-Werte beziehen sich auf einen einseitigen Mann-Whitney-U-Test. Ein signifikantes Testergebnis zeigt einen Abfall des zugehörigen Parameters an. Für jeden Zeitpunkt und jeden Parameter wurden die 10 zugehörigen p-Werte, unabhängig voneinander, mit Hilfe der Holm-Prozedur adjustiert. (b) Zeigt ein analoges Ergebnis ohne Adjustierung der p-Werte. Abbildungen modifiziert nach [29].

sion bestimmter Subpopulationen. Um der Ursache des Phänomens auf den Grund zu gehen, wurden von jeweils 6 Mäusen aus der Kontroll- sowie der 7d-Gruppe jeweils ein zusätzlicher Milzschnitt entnommen und das hierin enthaltene T-Zellrezeptorrepertoire extrahiert. Es wurde hierbei ein möglichst großer Abstand zu dem Bereich eingehalten, aus dem der bestehende Milzschnitt entnommen worden war. Hierdurch sollte der Einfluss lokaler Migrationseffekte so weit wie möglich reduziert werden. Für alle verfügbaren Milzschnitte der Kontroll- sowie der 7d-Gruppe wurden die Teilrepertoires bestehend aus Klonotypen mit CN ≤ 4 betrachtet und der Jaccard-Index, in beiden Gruppen, für alle möglichen Paare bestimmt. Die Überlegung hinter dieser Vorgehensweise ist Folgende: Wäre der Jaccard-Index zwischen den Teilrepertoires immunisierter Tiere deutlich erhöht, sofern die Schnitte aus demselben Tier stammen, wäre dies ein Hinweis dafür, dass die detektierten Klonotypen einen gemeinsamen Ursprung haben, also mit hoher Wahrscheinlichkeit aus dem Blutstrom eingewandert sind. Es zeigte sich

jedoch, dass der Jaccard-Index durch die Immunisierung zwar etwas absinkt, es hierbei aber völlig unerheblich ist, ob die Schnitte aus demselben Tier stammen oder nicht. Die erhofften Schlüsse konnten aus diesem Experiment daher nicht gezogen werden. Hierauf wird in der Diskussion noch gesondert eingegangen werden. Man beachte, dass die zusätzlichen 2×6 Milzschnitte ausschließlich für diese Betrachtungen entnommen wurden. In allen übrigen Analysen wurden sie nicht berücksichtigt.

Da die T-Zellreaktion gegen SRBC nur zu einem marginalen Anteil auf *public*-Klonotypen beruht [47, 51], sind die in diesem Abschnitt beschriebenen Effekte in erster Linie auf die Expansion von *private*-Klonotypen zurückzuführen. Dies wirft die Frage auf, inwieweit sich das in Abbildung 6 gezeigte Verteilungsmuster auf die wenigen reagiblen *public*-Klonotypen übertragen lässt.

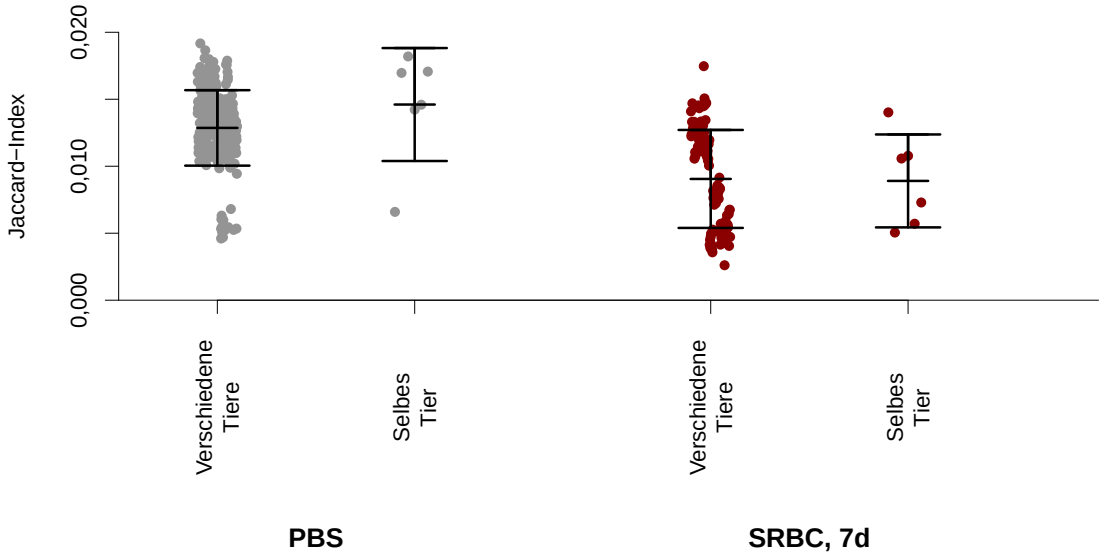

Abbildung 7: Vergleich des Jaccard-Index zwischen den T-Zellrezeptorrepertoires welche aus Milzschnitten desselben bzw. zweier verschiedener Tiere extrahiert wurden. Aus jeweils 6 Tieren der Kontrollgruppe (PBS-Injektion, $n = 20$), sowie aus der immunisierten 7d-Gruppe ($n = 10$) wurden zusätzlich zum bestehenden Milzschnitt ein Weiterer entnommen und das T-Zellrezeptorrepertoire hieraus extrahiert. Für die resultierenden 42 Datensätze wurden Teilrepertoires, bestehend aus Klonotypen mit CN ≤ 4 betrachtet und der Jaccard-Index für alle möglichen Paare innerhalb der beiden experimentellen Gruppen berechnet. Dies beinhaltete sowohl Vergleiche von Schnitten aus demselben Tier als auch aus verschiedenen Tieren. Die Fehlerbalken geben Mittelwert und Standardabweichung an. Aufgrund multipler statistischer Abhängigkeiten zwischen den einzelnen Werten wurde kein Test durchgeführt.

3.4 Die SRBC-spezifische T-Zellantwort beinhaltet eine nachgeordnete *public*-Komponente mit besonderem Verteilungsmuster

Mit Hilfe einer Genexpressionsanalyse wurden 40 Klonotypen identifiziert, welche in der Mehrzahl der immunisierten Tiere detektiert werden können und nach SRBC-Injektion expandieren. Am 3. Tag nach Antigenapplikation war der Anteil dieser Klonotypen an den insgesamt ausgelesenen Sequenzen (incl. Mehrfachdetektionen), im Mittel, auf das $3,9$-fache, einen Tag später auf das $6,2$-fache des Ausgangswertes (innerhalb der Kontrollgruppe) erhöht. Bis zum 7. Tag fiel dieser Wert wieder ab, blieb jedoch weiterhin um einen Faktor von $1,9$ gegenüber dem Ausgangswert erhöht (Abbildung 8a). Während sich die Aktivität der reagiblen *private*-Klontypen nur indirekt, beispielsweise anhand der Diversifizierung der CDR3-Sequenzen oder VJ-Segmente bzw. Homogenisierung der zugrundeliegenden Nukleotidkodierungen beobachten lässt, ist die Kopienzahl der identifizierten *public*-Sequenzen sowohl vor als auch nach der Immunisierung in jedem Datensatz bekannt. Die Expansion der *public*-Klonotypen kann daher direkt beobachtet und quantifiziert werden. Hierbei ergeben sich einige interessante Unterschiede zwischen den beiden Komponenten der T-Zellantwort. Im naiven Tier ähnelt die Verteilung der identifizierten *public*-Klonotypen auf die 10, im vorangegangenen Abschnitt definierten, Kategorien einer Glockenkurve. Die höchste Anzahl an Klonotypen war hier im Bereich $5 < \log_2(\mathrm{CN}) \leq 6$ (Abbildung 8b-d) zu finden. Zum Vergleich: Bei der Gesamtverteilung aller Klonotypen war die höchste Anzahl im Bereich $2 < \log_2(\mathrm{CN}) \leq 4$ lokalisiert (siehe Abbildung 5a). Nach Antigenapplikation kam es ab dem 3. Tag zu einer Akkumulation der SRBC-spezifischen *public*-Klonotypen in der obersten Kategorie ($\log_2(\mathrm{CN}) > 9$). Am stärksten war dieser Effekt am 4. Tag nach Immunisierung ausgeprägt. Zu diesem Zeitpunkt waren im Mittel ~ 9 der 40 identifizierten *public*-Klonotypen in diesem Bereich zu finden (Abbildung 8b). Dennoch blieb dieser Effekt bis zum 7. Tag nach Antigenapplikation signifikant (Abbildung 8c). Im Gegensatz hierzu waren an Tag 7 in der obersten Kategorie keine Effekte mehr feststellbar, welche auf eine Akkumulation reagibler *private*-Klonotypen hindeuten würden (siehe Abbildung 6). Während sich im Intervall $4 < \log_2(\mathrm{CN}) \leq 6$ zu keinem Zeitpunkt Hinweise auf Aktivität reagibler *private*-Klonotypen ergaben, waren stets $\sim 25\%$ der spezifischen *public*-Klonotypen in diesem Bereich zu finden. Am 3. und 4. Tag zeigten sich in einer bzw. zwei Kategorien im Bereich der mittleren Kopienzahlen signifikante

Immunisierungseffekte im Hinblick auf die Verteilung der *public*-Klonotypen (Abbildung 8b-c). Der markanteste Unterschied zwischen den beiden Komponenten besteht darin, dass es an Tag 7 zu keiner Akkumulation von *public*-Klonotypen im unteren Bereich ($\log_2(\mathrm{CN}) \leq 2$) kam. Allerdings deutet diese Beobachtung nicht auf ein biologisches Phänomen hin, sie lässt sich vielmehr mit der der hier vorgeschlagenen Methodik erklären. Hierauf wird in der Diskussion ausführlich eingegangen werden.

3.5 Durch statistische Klassifikationsverfahren können Immunisierungseffekte im einzelnen Tier nachgewiesen werden

Bei allen bisherigen Analysen wurden naive und immunisierte Tiere gruppenweise miteinander verglichen. Die in diesem Abschnitt dargelegten Ergebnisse sollen dagegen Aufschluss über die Frage geben, anhand welcher Kriterien die extrahierten T-Zellrezeptorrepertoires (bzw. bestimmte Teilbereiche) der einzelnen Tiere korrekt, im Hinblick auf den Immunisierungsstatus, klassifiziert werden können. Der entscheidende Unterschied zu den bisherigen Betrachtungen besteht also darin, dass die folgenden Analysen deutlich konkretere Aussagen über das Repertoire des einzelnen Tieres zulassen. Zwei Teilbereiche, in denen zuvor signifikante Immunisierungseffekte detektiert worden waren (siehe Abschnitt 3.3), wurden hierbei gesondert betrachtet. Für jeden Datensatz X wurden die Teilrepertoires

$$\begin{aligned} X^{\mathrm{hoch}} &= \{x \in X : \log_2(\mathrm{CN}) > 6, \quad \text{d.h. } \mathrm{CN} > 64\} \text{ und} \\ X^{\mathrm{niedrig}} &= \{x \in X : \log_2(\mathrm{CN}) \leq 2, \quad \text{d.h. } \mathrm{CN} \leq 4\} \end{aligned}$$

definiert. Im Hinblick auf die in Abschnitt 3.3 eingeführte Einteilung entsprechen diese Intervalle denjenigen Kategorien, in denen zu zwei verschiedenen Zeitpunkten oder im Hinblick auf zwei verschiedene biologische Parameter Immunisierungseffekte festgestellt wurden. Im Bereich $\mathrm{CN} > 64$ manifestiert sich die T-Zellreaktion an allen drei Zeitpunkten durch eine Diversifizierung der detektierten V- bzw. J-Segmente. Im Bereich $\mathrm{CN} \leq 4$ wurde neben der Diversifizierung der V- bzw. J-Segmente an Tag 7 eine hochsignifikante Homogenisierung der Nukleotidkodierungen der CDR3-Aminosäuresequenzen beobachtet (siehe Abbildung 6). Für das in Abschnitt 2.3.5 definierte Klassifikationsverfahren ist zunächst die Wahl der biologischen Parameter maßgeblich, auf deren Basis das erforderliche Distanzmaß d konstruiert werden soll. Eine zuverlässige

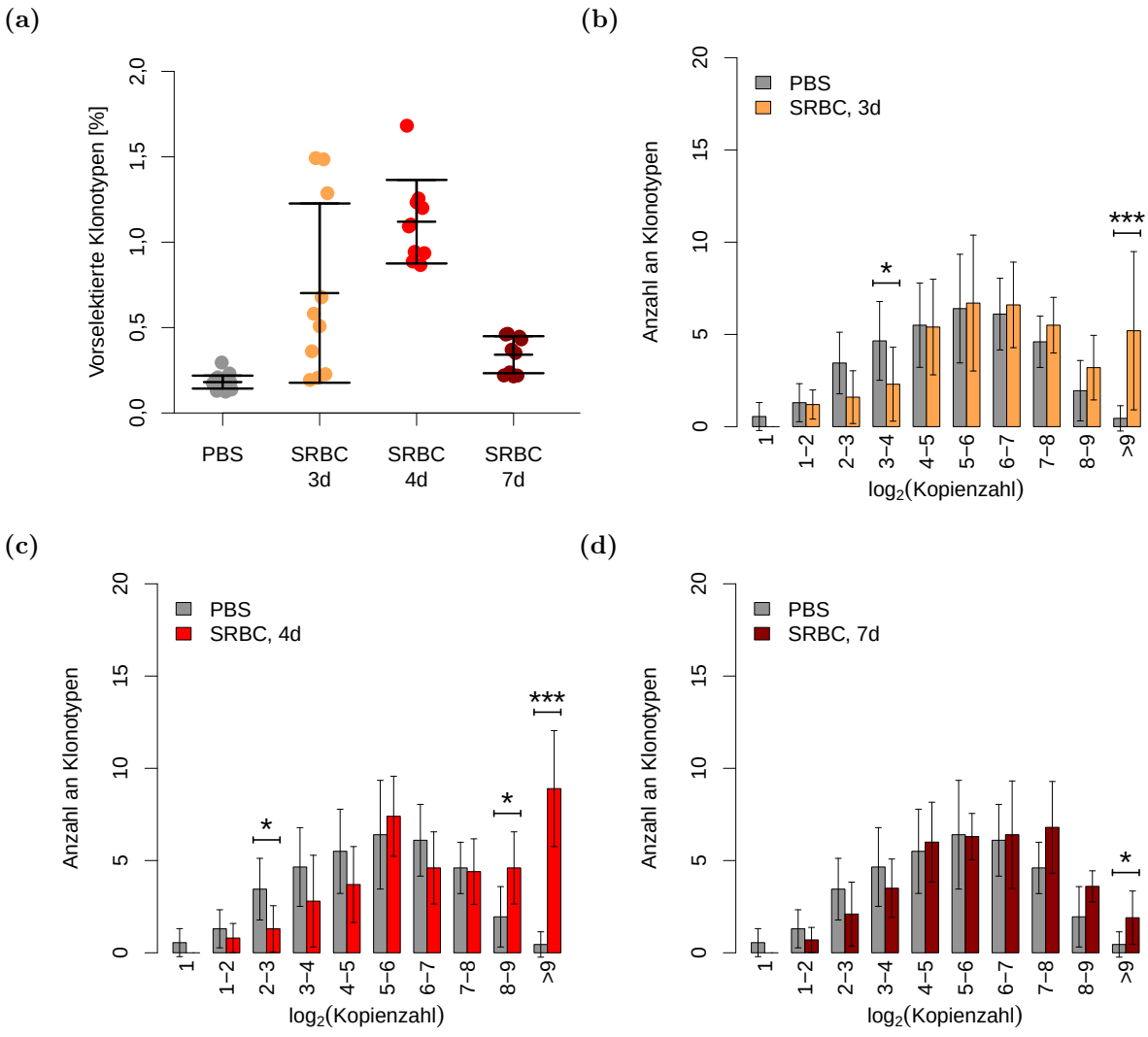

Abbildung 8: Die SRBC-spezifische T-Zellantwort beinhaltet eine nachgeordnete *public*-Komponente mit besonderem Verteilungsmuster. Mittels Genexpressionsanalyse wurden 40 Klonotypen identifiziert, welche in der Mehrheit der immunisierten Tiere expandieren. (a) Anteil der 40 vorselektierten Klonotypen an den insgesamt ausgelesenen Sequenzen (incl. Mehrfachdetektionen) in der Kontrollgruppe (PBS-Injektion, $n = 20$) sowie in den immunisierten Tieren ($n = 10$ für jeden Zeitpunkt). Die Fehlerbalken geben jeweils Mittelwert und Standardabweichung an. (b)-(d) Das extrahierte T-Zellrezeptorrepertoire der einzelnen Tiere wurde in jeweils 10 Teilbereiche unterteilt. Die Balkendiagramme zeigen die mittlere Anzahl der vorselektierten Klonotypen in den einzelnen Teilbereichen, die Fehlerbalken geben die Standardabweichung an. Für jeden Teilbereich wurden die Zahlen der immunisierten Tiere mit denen der Kontrollgruppe mittels Mann-Whitney-U-Test verglichen. Die Adjustierung der p-Werte wurde mit Hilfe der Holm-Prozedur vorgenommen. Die adjustierten p-Werte sind folgendermaßen dargestellt: * $p < 0,05$, ** $p < 0,01$,*** $p < 0,001$. Abbildungen (b)-(d) modifiziert nach [29].

Klassifikation ist nur dann möglich, wenn die Repertoires innerhalb der experimentellen Gruppen möglichst ähnlich zueinander sind (d.h. im Hinblick auf d eng beieinanderliegen) während Individuen verschiedener Gruppen größtmögliche Unterschiede zueinander aufweisen (d.h. im Hinblick auf d weit auseinanderliegen). In Abschnitt 3.3 wurden auf Basis dreier Parameter (Ähnlichkeit der CDR3-Regionen, detektierte V- bzw. J-Segmente, Nukleotidkodierung der CDR3-Regionen) Schwerpunktbereiche der T-Zellreaktion identifiziert. Es stellt sich nun die Frage, welcher dieser Parameter eine trennscharfe Klassifikation der Datensätze zulässt. Man beachte, dass die drei Parameter im Folgenden nicht dazu genutzt werden sollen, um Diversifizierungs- bzw. Homogenisierungseffekte innerhalb eines Datensatzes aufzudecken, sondern um verschiedene Datensätze in Relation zueinander zu stellen. Daher kommen in diesem Abschnitt keine Homogenitätsindizes sondern die in Abschnitt 2.3.3 definierten Ähnlichkeitsindizes zur Anwendung. Auf Basis dieser Kennzahlen wurden alle erforderlichen Distanzmaße konstruiert (siehe Abschnitt 2.3.5). Um sie anwenden zu können ist, analog zu den Homogenitätsindizes, ein Kriterium erforderlich, welches festlegt, wann zwei einzelne Sequenzen als ähnlich angesehen werden. Zwei der drei in Abschnitt 3.3 betrachteten Parameter (Ähnlichkeit der CDR3-Sequenzen im Hinblick auf die Levenshtein Distanz, zugeordnete V- bzw. J-Segmente) lassen sich unmittelbar auf die neue Fragestellung übertragen.

Aus Abschnitt 3.3 ist ersichtlich, dass der Repertoire-Homogenitätsindex im Hinblick auf die Levenshtein-Distanz $\mathrm{RHI_{LD}}$ nur eine geringe Trennschärfe zwischen naivem und immunisiertem T-Zellrezeptorrepertoire aufweist. Immunisierungseffekte sind auf Basis dieser Kennzahl nur in einem relativ kleinen Bereich detektierbar. Für die Klassifikation von Datensätzen ist das zugrundeliegende Ähnlichkeitskriterium jedoch noch aus einem weiteren Grund vollkommen ungeeignet. Dies zeigt sich bei Betrachtung der Klonotypen in den beiden obersten, der in Abschnitt 3.3 definierten Kategorien ($\log_2(\mathrm{CN}) > 8$, d.h. $\mathrm{CN} > 256$). In diesem Bereich wurden am 3. und 4. Tag nach Antigenapplikation signifikante Immunisierungseffekte im Hinblick auf $\mathrm{RHI_{LD}}$ detektiert. Berechnet man für diese Teilrepertoires den Repertoire-Ähnlichkeitsindex (mit gleichem Ähnlichkeitskriterium zwischen den Einzelsequenzen) zwischen den einzelnen Datensätzen, so stellt man an Tag 3 und 4 eine Divergenz der Repertoires (d.h. ein Absinken des Ähnlichkeitsindex) fest (Abbildung 9). Es ist hierbei offensichtlich unerheblich, ob beide oder nur eines der verglichenen Tiere immunisiert wurde. Zwei

immunisierte Tiere sind sich also (im Hinblick auf dieses Kriterium) nicht ähnlicher als ein immunisiertes Tier einem Naiven. Für eine Klassifikation der Datensätze kommt dieses Kriterium daher nicht infrage.

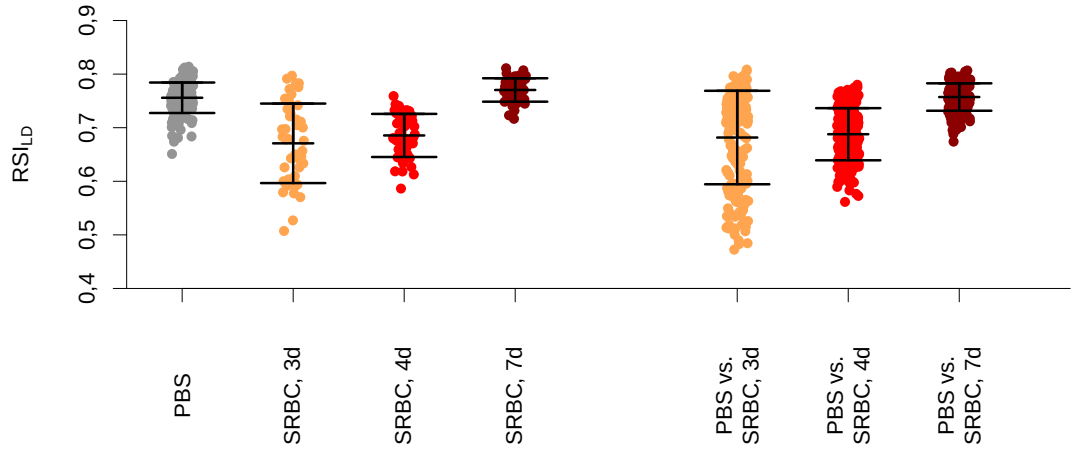

Abbildung 9: Nach Immunisierung mit SRBC werden innerhalb der häufigen Klonotypen verschiedener Datensätze weniger Paare ähnlicher CDR3-Sequenzen detektiert. Von allen Datensätzen ($n = 20$ für Kontrollgruppe (PBS-Injektion), $n = 10$ für jede der immunisierten Gruppen) wurden diejenigen Klonotypen betrachtet, deren Kopienzahl einen Wert von 256 übersteigt. Die resultierenden Teilrepertoires wurden paarweise mittels des Repertoire-Ähnlichkeitsindex im Hinblick auf die Levenshtein-Distanz analysiert (RSI_{LD}; zwei Klonotypen werden hierbei als ähnlich angesehen, falls die Levenshtein-Distanz der zugehörigen CDR3-Sequenzen einen Wert von 1 nicht übersteigt). Es zeigt sich ein transienter Divergenzeffekt zwischen den Teilrepertoires. Hierbei ist unerheblich, ob zwei immunisierte Tiere, oder ein immunisiertes Tier mit einem Naiven verglichen wird. Die Fehlerbalken geben Mittelwert und Standardabweichung an. Aufgrund multipler statistischer Abhängigkeiten zwischen den einzelnen Werten wurde kein Test durchgeführt.

Wie in Abschnitt 3.3 beschrieben, wurden in weiten Teilen des T-Zellrezeptorrepertoires Diversifizierungseffekte im Hinblick auf die detektierten V- bzw. J-Segmente festgestellt. Um die Datensätze auf Basis der Segmente zu klassifizieren, wurden die Übereinstimmung der V- bzw. J-Segmente, der einzelnen Klonotypen, als unabhängige, gleich gewichtete Ähnlichkeitskriterien genutzt. Zunächst wurden diese Kriterien auf die Teilrepertoires mit hoher Kopienzahl ($CN > 64$) angewendet und die Datensätze mit Hilfe einer Clusteranalyse in zwei Gruppen eingeteilt. Wie Tabelle 2 zu entnehmen ist, wird einer der Cluster klar von den immunisierten, der andere von den naiven Tieren dominiert. Eine (unerwünschte) Heterogenisierung innerhalb der Cluster wird vor allem

durch Tiere der Kontroll- sowie der 3d-Gruppe hervorgerufen. Die approximativen Distanzen zwischen den einzelnen Datensätzen sowie das Ergebnis der Clusteranalyse werden in Abbildung 10a veranschaulicht. Mit Hilfe des, in Abschnitt 2.3.5 definierten, überwachten Klassifikationsverfahrens konnten diese Teilrepertoires, auf Basis der V- bzw. J-Segmente, relativ zuverlässig klassifiziert werden (Abbildung 10b). Hierbei wurden 18 der 20 naiven Tiere korrekt klassifiziert. Innerhalb der 3d-Gruppe wurden 3, in der 4d-Gruppe ein Tier fehlklassifiziert. Die 7d-Gruppe wurde durchgehend fehlerfrei klassifiziert. Dieses Ergebnis unterscheidet sich signifikant von einer zufälligen Klassifikation ($p < 6 \cdot 10^{-8}$).

Das Segmentkriterium wurde ebenso auf Teilrepertoires mit niedriger Kopienzahl ($CN \leq 4$) angewendet. Hierbei zeigt sich eine Clusterstruktur, bei der lediglich die 7d-Gruppe einen leicht separierten Cluster bildet (Abbildung 10c). Dennoch konnte mit Hilfe der Clusteranalyse keine der experimentellen Gruppen zuverlässig von den übrigen Datensätzen abgegrenzt werden (Tabelle 2). Um das überwachte Klassifikationsverfahren in sinnvoller Weise auf diese Teilrepertoires anwenden zu können, wurden die Datensätze der 3d- und 4d-Gruppe entfernt. Im analysierten Bereich sind zu diesen Zeitpunkten lediglich marginale Immunisierungseffekte feststellbar (siehe Abbildung 6), was sich auch in der in Abbildung 10c dargestellten Clusterstruktur wiederspiegelt. Würden diese Teilrepertoires einem überwachten Lernverfahren als „immunisiert“ präsentiert, würde der Algorithmus hierdurch erheblich gestört werden. Insgesamt wurden drei naive und ein immunisiertes Tier fehlklassifiziert. Dennoch übertrifft dieses Ergebnis eine Zufallsklassifikation signifikant ($p < 2 \cdot 10^{-4}$).

In Abschnitt 3.3 wurde zur Quantifizierung der Diversität, der Nukleotidkodierungen, der CDR3-Aminosäuresequenzen eine Kennzahl genutzt, welche kein zusätzliches Ähnlichkeitskriterium zwischen den einzelnen Klonotypen erfordert. Um Datensätze auf Basis dieser Nukleotidkodierungen klassifizieren zu können, musste daher zunächst ein entsprechendes Kriterium definiert werden. Hierfür wurden zwei Klonotypen als ähnlich definiert, falls die Anzahl der Nukleotidkodierungen für die entsprechenden CDR3-Aminosäuresequenzen entweder übereinstimmt oder für beide Klonotypen einen Wert von 5 übersteigt. Angewendet auf die Teilrepertoires mit hoher Kopienzahl ($CN > 64$) führte dieses Kriterium zu einer chaotischen Clusterstruktur, in welcher sich keine der experimentellen Gruppen von den übrigen Datensätzen abgrenzen lässt (Tabelle 3, Abbildung 11a). Das Ergebnis des überwachten Klassifikationsverfahrens unterscheidet

	Experimentelle Gruppe	Cluster 1	Cluster 2
Hohe Kopienzahl (CN > 64)	PBS	4	16
	SRBC, 3d	8	2
	SRBC, 4d	9	1
	SRBC, 7d	10	0
Niedrige Kopienzahl (CN ≤ 4)	PBS	13	7
	SRBC, 3d	4	6
	SRBC, 4d	4	6
	SRBC, 7d	4	6

Tabelle 2: Auf Basis der detektierten V- bzw. J-Segmente wurden Teilrepertoires mit hoher bzw. niedriger Kopienzahl in zwei Gruppen (Cluster) eingeteilt. Für die einzelnen Teilrepertoires wurde eine Distanzmatrix berechnet und die Datensätze mittels k-Medoid-Verfahren in zwei Cluster eingeteilt. Teilrepertoires mit hoher/niedriger Kopienzahl wurden jeweils getrennt voneinander analysiert. Der obere und untere Teil der Tabelle bezieht sich daher jeweils auf eine gesonderte Auswertung. Man beachte, dass es sich bei dem angewandten Algorithmus um ein nicht überwachtes Lernverfahren handelt. Die Nummerierung der Cluster ist daher beliebig. Über die Güte der Einteilung gibt lediglich die Homogenität innerhalb der einzelnen Cluster Auskunft.

sich dementsprechend nicht signifikant von einer zufälligen Klassifikation ($p = 0,377$, Abbildung 11b). Wird das Nukleotidkriterium dagegen auf Teilrepertoires mit niedriger Kopienzahl angewendet, so ergibt sich eine Clusterstruktur, in welcher einer der Cluster exakt mit der 7d-Gruppe übereinstimmt. Der zweite Cluster beinhaltet gleichermaßen die Tiere der 3d- und 4d-Gruppe sowie die komplette Kontrollgruppe (Tabelle 3, Abbildung 11c). Nach Entfernung der 3d- und 4d-Gruppe führte das überwachte Verfahren zu einer fehlerfreien Klassifikation (Abbildung 11d). Dies kann nicht mit zufälligen Effekten erklärt werden ($p < 4 \cdot 10^{-8}$).

Da sich die Verteilung der detektierten V- bzw. J-Segmenten, nach SRBC-Applikation, in weiten Teilen des T-Zellrezeptorrepertoires signifikant verändert, bieten sich insbesondere diese Parameter zur Klassifikation vollständiger Datensätze (ohne Einschränkung auf bestimmte Teilrepertoires) an. Zunächst wurden hierbei alle Klonotypen (analog zu den Analysen der Teilrepertoires) gleich stark gewichtet. Dies führte zu einer chaotischen Clusterstruktur, in welcher sich keine der experimentellen Gruppen von den übrigen Datensätzen abgrenzen lässt (Tabelle 4, Abbildung 12a). Dementsprechend unterschied sich das Ergebnis des überwachten Klassifikationsverfahrens nicht signifikant von einer zufälligen Klassifikation (Abbildung 12b, $p = 0,130$). Daraufhin

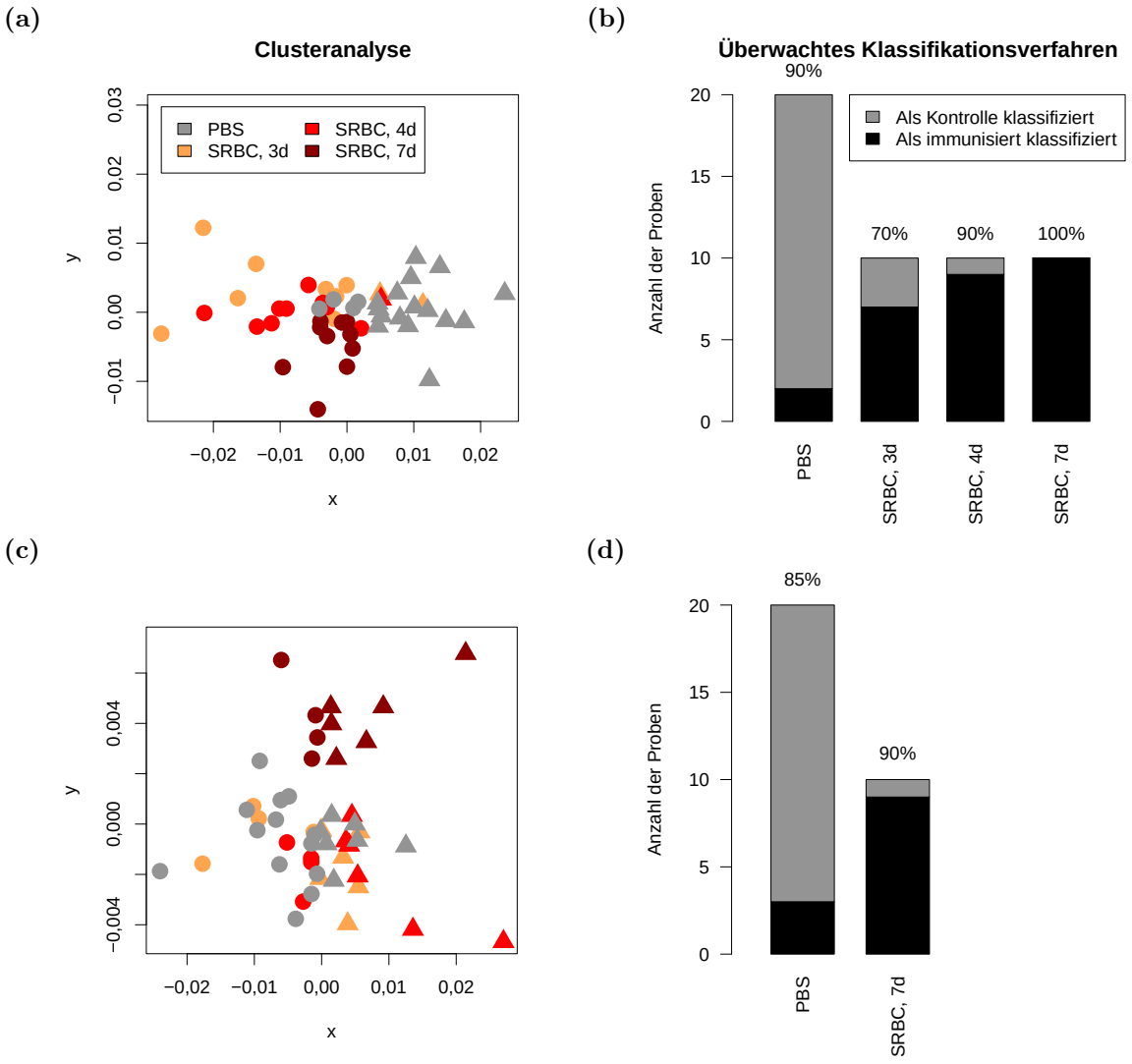

Abbildung 10: Klassifikation von Teilrepertoires auf Basis der V- bzw. J-Segmente: (a)-(b) Aus allen Datensätzen ($n = 50$) wurden die Klonotypen mit hoher Kopienzahl (CN > 64) ausgewählt. Für diese Teilrepertoires wurde (auf Basis der V- bzw. J-Segmente) eine Distanzmatrix berechnet, welche als Basis verwendet wurde, um die Teilrepertoires mittels k-Medoid-Verfahrens in zwei Gruppen (Cluster) aufzuteilen. Das Ergebnis wurde mittels multidimensionaler Skalierung visualisiert (a). Die abstrakten Distanzen zwischen den einzelnen Datensätzen werden hierbei durch die Abstände der Punkte im Streudiagramm approximiert. Die vom Algorithmus definierten Cluster sind als Kreise und Dreiecke dargestellt. (b) Auf Basis derselben Distanzmatrix wurde ein überwachtes Klassifikationsverfahren durchgeführt. Insgesamt wurden 44 der insgesamt 50 Datensätze korrekt klassifiziert. Die Zahlen oberhalb der Balken geben den Anteil der korrekt klassifizierten Teilrepertoires innerhalb der jeweiligen experimentellen Gruppe an. (c)-(d) zeigt eine analoge Klassifikation der Datensätze mit niedriger Kopienzahl (CN ≤ 4). Für das überwachte Klassifikationsverfahren wurden lediglich die Tiere der Kontroll- sowie der 7d-Gruppe verglichen. Innerhalb der Kontrollgruppe wurden 3, innerhalb der 7d-Gruppe ein Datensatz fehlklassifiziert. Abbildungen modifiziert nach [29].

	Experimentelle Gruppe	Cluster 1	Cluster 2
Hohe Kopienzahl (CN > 64)	PBS	14	6
	SRBC, 3d	9	1
	SRBC, 4d	6	4
	SRBC, 7d	6	4
Niedrige Kopienzahl (CN ≤ 4)	PBS	20	0
	SRBC, 3d	10	0
	SRBC, 4d	10	0
	SRBC, 7d	0	10

Tabelle 3: Auf Basis der Variabilität Nukleotidkodierungen der CDR3-Aminosäuresequenzen wurden Teilrepertoires mit hoher bzw. niedriger Kopienzahl in zwei Gruppen (Cluster) eingeteilt. Für die einzelnen Teilrepertoires wurde eine Distanzmatrix berechnet und die Datensätze mittels k-Medoid-Verfahren in zwei Cluster eingeteilt. Teilrepertoires mit hoher/niedriger Kopienzahl wurden jeweils getrennt voneinander analysiert. Der obere und untere Teil der Tabelle bezieht sich daher jeweils auf eine gesonderte Auswertung. Man beachte, dass es sich bei dem angewandten Algorithmus um ein nicht überwachtes Lernverfahren handelt. Die Nummerierung der Cluster ist daher beliebig. Über die Güte der Einteilung gibt lediglich die Homogenität innerhalb der einzelnen Cluster Auskunft.

wurde die Analyse dahingehend modifiziert, dass die einzelnen Klonotypen gemäß ihrer Kopienzahl gewichtet wurden. Dies bedeutet, dass das erforderliche Distanzmaß auf Basis des gewichteten Repertoire-Ähnlichkeitsindex $\mathrm{wRSI}_{\mathrm{VJ}}$ definiert wurde (siehe Abschnitt 2.3.5). Dies führte zu einer Clusterstruktur, in welcher sich die Tiere der Kontroll- sowie der 7d-Gruppe in jeweils einem Cluster konzentrieren (mit lediglich zwei Ausnahmen innerhalb der Kontrollgruppe, siehe Tabelle 4 und Abbildung 12c). Zu früheren Zeitpunkten ist dagegen keine klare Abgrenzung zwischen immunisierten und naiven Tieren möglich. Das überwachte Klassifikationsverfahren konnte die Datensätze mit moderater Genauigkeit zuordnen (Abbildung 12d). Anteilig unterlaufen hierbei die meisten Fehlklassifikationen innerhalb der 3d-Gruppe. Dennoch unterscheidet sich dieses Ergebnis signifikant von einer zufälligen Klassifikation ($p < 2 \cdot 10^{-5}$).

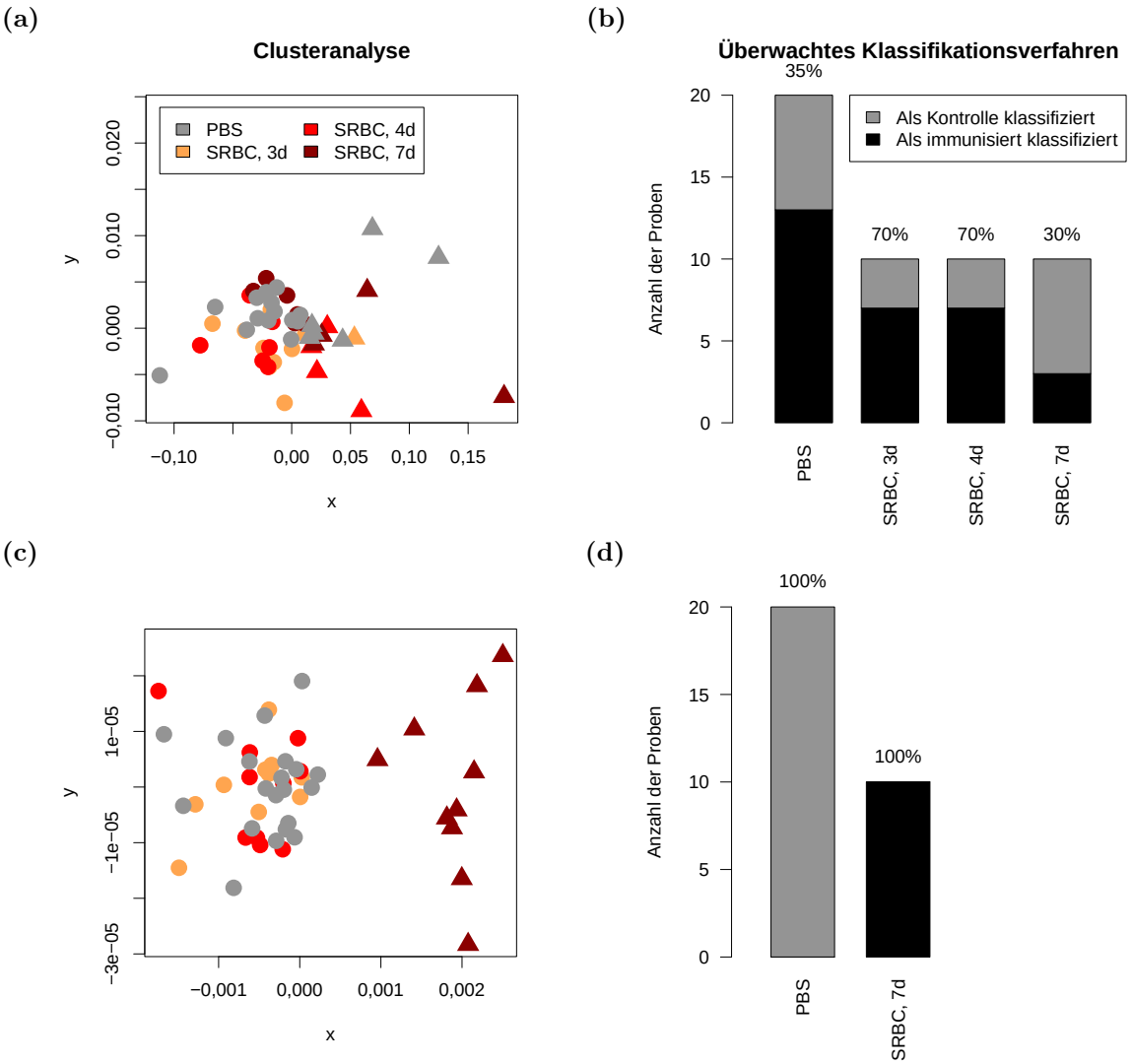

Abbildung 11: Klassifikation von Teilrepertoires auf Basis der Diversität Nukleotidkodierungen der CDR3-Aminosäuresequenzen: (a)-(b) Aus allen Datensätzen ($n = 50$) wurden die Klonotypen mit hoher Kopienzahl (CN $>$ 64) ausgewählt. Für diese Teilrepertoires wurde (auf Basis der Nukleotidkodierungen) eine Distanzmatrix berechnet, welche als Basis verwendet wurde, um die Teilrepertoires mittels k-Medoid-Verfahrens in zwei Gruppen (Cluster) aufzuteilen. Das Ergebnis wurde mittels multidimensionaler Skalierung visualisiert (a). Die abstrakten Distanzen zwischen den einzelnen Datensätzen werden hierbei durch die Abstände der Punkte im Streudiagramm approximiert. Die vom Algorithmus definierten Cluster sind als Kreise und Dreiecke dargestellt. Auf Basis dieses Kriteriums ist es nicht möglich, eine der experimentellen Gruppen von den übrigen Datensätzen abzugrenzen. Wird auf Basis der gleichen Distanzmatrix eine überwachte Klassifikation vorgenommen (b), so unterscheidet sich das Ergebnis nicht signifikant von einer Zufallsklassifikation ($p = 0,377$). Die Zahlen oberhalb der Balken geben den Anteil der korrekt klassifizierten Teilrepertoires innerhalb der jeweiligen experimentellen Gruppe an. (c)-(d) zeigt eine analoge Klassifikation der Datensätze mit niedriger Kopienzahl (CN ≤ 4). Für das überwachte Klassifikationsverfahren wurden lediglich die Tiere der Kontroll- sowie der 7d-Gruppe verglichen. Die 7d-Gruppe bildet hierbei einen deutlich abgegrenzten Cluster und kann fehlerfrei klassifiziert werden. Zu den beiden früheren Zeitpunkten können die immunisierten Teilrepertoires nicht von der Kontrollgruppe unterschieden werden. Abbildungen modifiziert nach [29].

	Experimentelle Gruppe	Cluster 1	Cluster 2
Klonotypen gleich stark gewichtet	PBS	8	12
	SRBC, 3d	3	7
	SRBC, 4d	2	8
	SRBC, 7d	7	3
Klonotypen nach Kopienzahl gewichtet	PBS	18	2
	SRBC, 3d	5	5
	SRBC, 4d	3	7
	SRBC, 7d	0	10

Tabelle 4: Gruppierung vollständiger Datensätze auf Basis der detektierten V- bzw. J-Segmente mittels Clusteranalyse. Für die T-Zellrezeptorrepertoires aller Versuchstiere ($n = 50$) wurde eine Distanzmatrix berechnet und die Datensätze mittels k-Medoid-Verfahren in zwei Cluster eingeteilt. Der obere Teil der Tabelle zeigt die Einteilung bei gleich starker Gewichtung aller Klonotypen. Der untere Teil der Tabelle zeigt das Ergebnis einer analogen Klassifikation, mit dem Unterschied, dass die Klonotypen bei der Berechnung der Distanzmatrix nach Kopienzahl gewichtet wurden. Man beachte, dass es sich bei dem angewandten Algorithmus um ein nicht überwachtes Lernverfahren handelt. Die Nummerierung der Cluster ist daher beliebig. Über die Güte der Einteilung gibt lediglich die Homogenität innerhalb der einzelnen Cluster Auskunft.

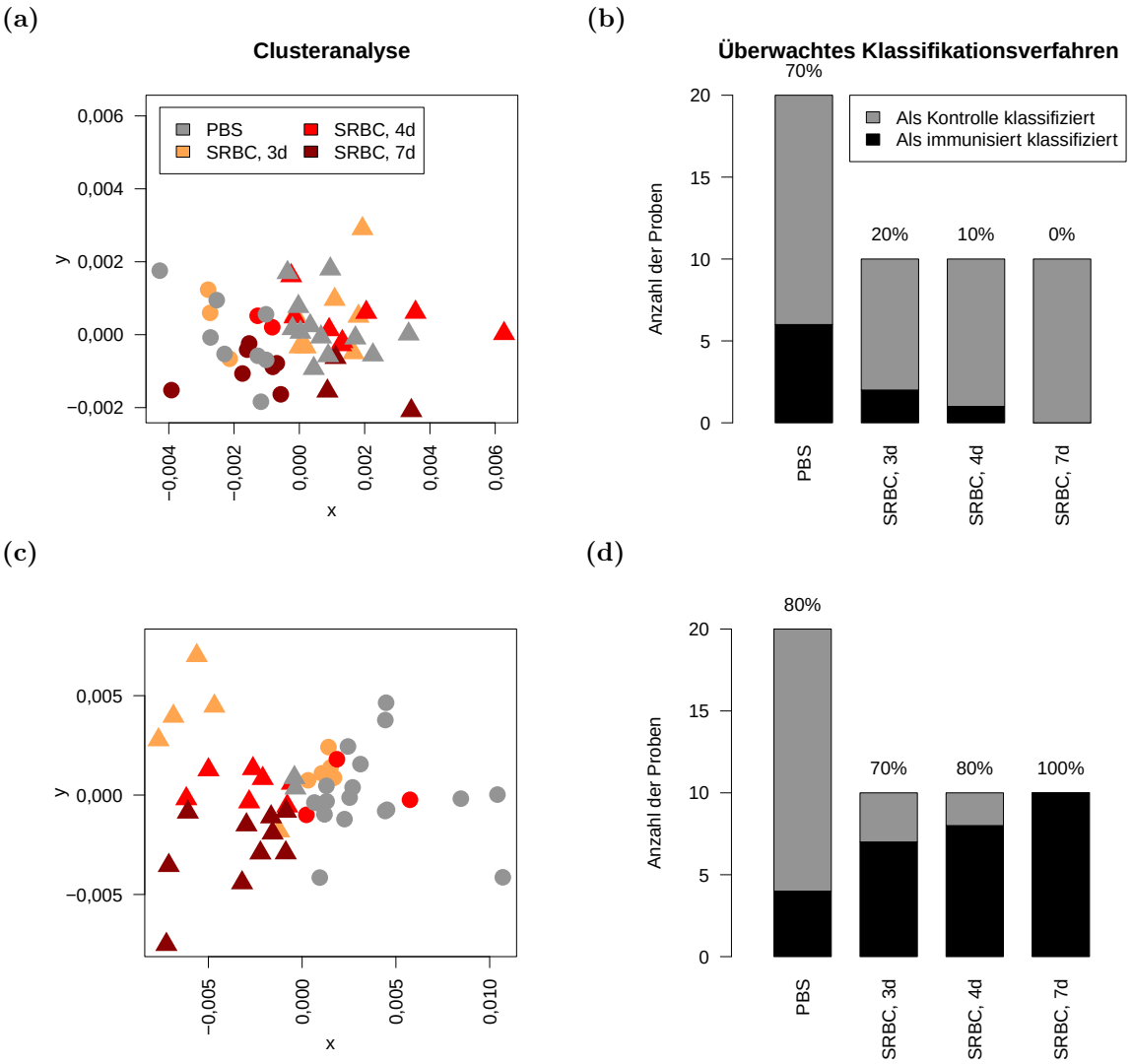

Abbildung 12: Klassifikation vollständiger Datensätze auf Basis der detektierten V- bzw. J-Segmente: (a)-(b) Für die extrahierten T-Zellrezeptorrepertoires aller Versuchstiere ($n = 50$) wurde eine Distanzmatrix berechnet, wobei alle Klonotypen gleich stark gewichtet wurden. Auf Basis dieser Distanzmatrix wurden die Datensätze mittels k-Medoid-Verfahren in zwei Cluster eingeteilt. Das Ergebnis wurde mittels multidimensionaler Skalierung visualisiert (a). Die abstrakten Distanzen zwischen den einzelnen Datensätzen werden hierbei durch die Abstände der Punkte im Streudiagramm approximiert. Die vom Algorithmus definierten Cluster sind als Kreise und Dreiecke dargestellt. Dieses Vorgehen führt zu einer chaotischen Clusterstruktur. Wird auf Basis derselben Distanzmatrix eine überwachte Klassifikation durchgeführt, wird die Mehrzahl der Proben fehlklassifiziert. Die Zahlen oberhalb der Balken geben den Anteil der korrekt klassifizierten Datensätze innerhalb der jeweiligen experimentellen Gruppe an. (c)-(d) zeigt eine analoge Klassifikation, wobei die Klonotypen bei der Berechnung der Distanzmatrix nach Kopienzahl gewichtet wurden. Dies führt zu einer deutlich verbesserten Einteilung. So können insgesamt 41 der 50 Proben korrekt klassifiziert werden. Abbildungen modifiziert nach [29].

4 Diskussion

4.1 Die SRBC-spezifische T-Zellreaktion manifestiert sich an der Verdrängung des naiven T-Zellrezeptorrepertoires

In dieser Arbeit wurde am Beispiel des SRBC-Modells eine T-Zellreaktion gegen ein komplexes zelluläres Antigen analysiert. Es stellte sich hierbei heraus, dass sich die meisten Immunisierungseffekte nur dann detektieren lassen, wenn diejenigen Teilrepertoires, in denen sich reagible Klonotypen konzentrieren, gezielt analysiert werden. Die Expansion individueller Klonotypen führt zu einem Absinken des Jaccard-Index (d.h. zu einer Reduktion der anteiligen Schnittmenge). Dieser Effekt kann als ein Auseinanderdriften der, im naiven Zustand vergleichsweise ähnlichen, Teilrepertoires angesehen werden. Zeitgleich kann ein analoger Divergenzeffekt zwischen den CDR3-Aminosäuresequenzen innerhalb der Teilrepertoires der einzelnen Tiere beobachtet werden. Dieser manifestiert sich in einem Rückgang von Sequenzpaaren mit ähnlicher CDR3-Region (gemessen an der Levenshtein-Distanz). Die detektierten V- bzw. J-Segmente werden durch die Immunreaktion diversifiziert. Dies zeigt, dass ein substanzieller Anteil der expandierenden Sequenzen auf Segmenten basiert, welche im naiven T-Zellrezeptorrepertoire vergleichsweise selten zu finden sind [51]. Im Hinblick auf die Nukleotidkodierungen der CDR3-Aminosäuresequenzen zeigte sich bei Klonotypen mit hoher Kopienzahl ein relativ hohes Maß an Diversität, welche in Folge der Immunreaktion reduziert wird. Dieser Effekt ist naheliegend, denn durch die Proliferation einer aktivierten T-Zelle entsteht stets ein Klon, dessen CDR3-Sequenzen sowohl auf Nukleotid- als auch auf Aminosäureebene übereinstimmen.

Letztlich lassen sich alle beobachteten Immunisierungseffekte durch einen einzigen Mechanismus erklären. Im naiven Zustand besteht ein substanzieller Anteil des sequenzierten T-Zellrezeptorrepertoires aus *public*-Sequenzen, welche definitionsgemäß in einer Vielzahl verschiedener Individuen vorkommen. (Bei einem paarweisen Vergleich zweier Repertoires macht die Schnittmenge $\sim$ 10% der detektierten Klonotypen aus, siehe Abbildung 2d). In der Literatur wird eine Vielzahl verschiedener Eigenschaften beschrieben, in denen sich diese besonderen Klonotypen vom übrigen T-Zellrezeptorrepertoire unterscheiden. Die große Mehrheit der *public*-Sequenzen geht auf eine stark restringierte Teilmenge an V- bzw. J-Segmenten zurück, sodass innerhalb dieser Sequenzen eine stark reduzierte Diversität der Segmente gegeben ist [25]. Die

CDR3-Aminosäuresquenz eines *public*-Klonotypen wird meist von einer Vielzahl verschiedener Nukleotidsequenzen kodiert. Das Phänomen, dass einzelne CDR3-Aminosäuresquenzen auf eine Vielzahl verschiedener Nukleotidkodierungen und damit auf verschiedene Kombinationsereignisse zurückgehen können, hat unter der Bezeichnung *konvergente Rekombination* Eingang in die Literatur gefunden. Diese wird auch als eine mögliche Erklärung für die Entstehen von *public*-Sequenzen angesehen [53, 54]. In [24] wurde anhand von Netzwerkanalysen gezeigt, dass zu einem gegebenen *public*-Klonotyp oft eine Vielzahl ähnlicher CDR3-Sequenzen (gemessen an der Levenshtein-Distanz) innerhalb des Rezeptorrepertoires gefunden werden kann. Es konnte darüber hinaus gezeigt werden, dass diese Clusterstrukturen im Falle einer Immunreaktion temporär gestört werden können und sich im Folgenden sukzessive wieder rekonstituieren.

Im hier analysierten SRBC-Modell führt die Expansion antigenspezifischer Klonotypen dazu, dass die naiven Klonotypen (einschließlich des darin enthaltenen Anteils an *public*-Sequenzen) in bestimmten Teilbereichen des sequenzierten Repertoires durch die Expansion antigenspezifischer Klonotypen ausgedünnt werden. Da es sich hierbei in erster Linie um eine *private response* handelt, fehlen den expandierenden Klonotypen in der Regel die oben genannten Eigenschaften. Es treten somit weniger Paare ähnlicher CDR3-Sequenzen auf (Absinken von $\mathrm{RHI_{LD}}$). Zudem führt der höhere Anteil an *private*-Klonotypen zu einer Diversifizierung der V- bzw. J-Segmente (Absinken von $\mathrm{RHI_{VJ}}$). Da expandierende Klonotypen meist auf eine einzige Zelle zurückgehen, führt die Immunreaktion zu einer Homogenisierung der Nukleotidkodierung (Absinken von CDI). Zusammenfassend lässt sich feststellen, dass die beobachteten Immunisierungseffekte weniger auf besondere Eigenschaften der SRBC-spezifischen Klonotypen zurückgehen, sondern auf die Verdrängung des naiven T-Zellrezeptorrepertoires, einschließlich der darin enthaltenen *public*-Komponente. Im Allgemeinen kommen Klonotypen, welche häufig in verschiedenen Individuen beobachtet werden (per Definition also als *public* klassifiziert werden) auch in den einzelnen Individuen gehäuft (d.h. in hoher Kopienzahl) vor [25]. So erklärt es sich, dass die Immunisierungseffekte im Bereich der Klonotypen mit hoher Kopienzahl (in absoluten Zahlen gemessen) sehr viel deutlicher ausfallen, als im übrigen T-Zellrezeptorrepertoire (siehe Abbildungen 2 und 3).

4.2 Zeitlich versetzt auftretende Proliferations- und Migrationseffekte bedingen die Kinetik der SRBC-spezifischen T-Zellreaktion

Durch eine systematische Aufteilung des T-Zellrezeptorrepertoires ist es möglich, die T-Zellreaktion im zeitlichen Verlauf nachzuverfolgen sowie Schwerpunktbereiche zu identifizieren, in denen sich reagible Klonotypen konzentrieren (Abbildung 6). Durch die in dieser Arbeit gewählte logarithmische Aufteilung wird sichergestellt, dass sich in allen Untergruppen genügend Klonotypen befinden, um eine valide statistische Auswertung zu gewährleisten. Ein linearer Ansatz wäre hierfür ungeeignet, da sich die große Mehrheit der Klonotypen in den unteren Kategorien (niedrige Kopienzahl) konzentrieren würde. Hierbei würden mitunter sehr heterogene Gruppen zu einer Kategorie zusammengefasst, wodurch die Aufteilung des Repertoires ihren Sinn verlieren würde. Hinzu kommt, dass die wenigen Klonotypen in den oberen Kategorien (hohe Kopienzahl) keine validen Aussagen zulassen würden.

Am 3. Tag nach Antigenapplikation kommt es zu einer massiven Expansion vergleichsweise weniger individueller Klonotypen. Diese äußerst sich in einer signifikanten Zunahme an Klonotypen mit hoher Kopienzahl sowie in einem Absinken des Jaccard-Index zwischen den entsprechenden Teilrepertoires (Abbildung 2). Der Schwerpunkt der Reaktion liegt in dieser frühen Phase im Bereich der Klonotypen mit hoher Kopienzahl. Dennoch lässt sich aus der Diversifizierung der V- bzw. J-Segmente ableiten, dass das T-Zellrezeptorrepertoire bereits an Tag 3 bis ins obere Mittelfeld ($\log_2(\mathrm{CN}) > 6$) signifikante Immunisierungseffekte aufweist. Im Folgenden werden weite Teile des sequenzierten Repertoires mit reagiblen Klonotypen durchsetzt. Der Schwerpunkt der Reaktion (d.h. der Bereich, in welchem die Immunisierungseffekte das höchste Signifikanzniveau erreichen) verlagert sich vom Bereich der Klonotypen mit sehr hoher Kopienzahl sukzessive ins obere Mittelfeld. Dies deutet darauf hin, dass fortwährend weitere Klonotypen in die T-Zellreaktion involviert werden, welche jedoch weniger stark expandieren. Intuitiv wäre es naheliegend, sowohl das abgeschwächte Expansionsverhalten als auch der spätere Zeitpunkt der Expansion mit einer geringeren Affinität der jeweiligen Klonotypen zu erklären. Allerdings wurde in [56] und [27] gezeigt, dass die Affinität eines Klonotyps zwar das finale Expansionslevel, nicht aber die initiale Expansionsgeschwindigkeit beeinflusst. Aus dem in [27] vorgeschlagenen Rechenmodell geht hervor, dass das finale Expansionslevel eines Klonotyps gleichermaßen von dessen Affinität,

der verfügbaren Antigemenge und dem Konkurrenzdruck durch andere Klonotypen determiniert wird. Dies lässt eine andere Interpretation der beobachteten Dynamik zu: Nachdem das applizierte Antigen von antigenpräsentierenden Zellen phagozytiert und prozessiert wurde, ist in der Milz zunächst ein Überangebot an pMHC-Komplexen vorhanden. Antigenspezifische Klonotypen, welche den Ort des Geschehens zufällig als erste erreichen, werden in ihrem Expansionsverhalten weder durch Antigenmangel noch durch einen hohen Konkurrenzdruck durch andere Klonotypen beeinträchtigt. Im weiteren Verlauf nimmt die Menge des verfügbaren Antigens ab, während fortwährend weitere Klonotypen eintreffen, welche um die verbleibenden pMHC-Komplexe konkurrieren. Eine Aktivierung der Klonotypen findet weiterhin statt, allerdings erreichen diese nicht mehr das Expansionslevel der frühen Phase. Während sich der Schwerpunkt der Reaktion sukzessive ins obere Mittelfeld verlagert, führt die Auswanderung reagibler, sowie die Einwanderung naiver Klonotypen dazu, dass bis zum 7. Tag in der obersten Kategorie ($\log_2(\mathrm{CN}) > 9$) keine signifikanten Immunisierungseffekte mehr feststellbar sind. Im Prinzip wäre es natürlich denkbar, dass diese Beobachtung nicht auf die Einwanderung naiver T-Zellen, sondern auf eine zeitversetzt einsetzende *public response* gegen das applizierte Antigen zurückzuführen ist. Eine solche T-Zellreaktion könnte zur Expansion antigenspezifischer Klonotypen führen, welche dem naiven Repertoire so stark ähneln, dass sich die Reaktion mit den hier vorgeschlagenen Methoden nicht detektieren lässt. Das ein solcher Effekt jedoch nicht ausschlaggebend sein kann, wird aus Abbildung 9 ersichtlich. In diesem Fall wurden Teilrepertoires mit $\log_2(\mathrm{CN}) > 8$ betrachtet. Würden diese Teilrepertoires am 7. Tag von antigenspezifischen *public*-Klonotypen dominiert, wären in zwei immunisierten Teilrepertoires mehr gleiche (insbesondere also ähnliche) CDR3-Sequenzen zu finden als bei Vergleich eines immunisierten Repertoires mit einem Naiven. Eine solche Situation wurde beispielsweise in einem Mausmodell beschrieben, in welchem die Tiere mit abgetöteten Tuberkuloseerregern immunisiert wurden [49]. Im SRBC-Model sind die Gegebenheiten jedoch anders. Wie aus Abbildung 9 hervorgeht, ist für das Auftreten gleicher (bzw. ähnlicher) Sequenzen, in zwei verschiedenen Datensätzen, lediglich entscheidend, ob einer der beiden Teilrepertoires immunisiert wurde. Die ähnlichen Sequenzen sind also dem naiven Repertoire zuzurechnen. Dieser Effekt lässt sich in analoger Weise mit Hilfe des Jaccard-Index nachweisen [51]. Vom 4. Tag an etabliert sich, im unteren Bereich des sequenzierten Repertoires ($\log_2(\mathrm{CN}) \leq 2$) ein zusätzlicher Schwerpunkt. Verschie-

dene Beobachtungen legen die Annahme nahe, dass sich dieser zweite Schwerpunkt nicht auf lokale Proliferation, sondern auf eine Wiedereinwanderung von Klonotypen aus dem Blutstrom zurückführen lässt, welche ursprünglich in anderen Teilen der Milz expandiert sind. Zunächst spricht die Verlagerung des ersten Schwerpunktes für eine beginnende Auswanderung antigenspezifischer Klonotypen aus der Milz. Zudem wurde in [47] gezeigt, dass sich SRBC-spezifische *public*-Klonotypen ab dem 4. Tag nach Antigenapplikation verstärkt im Blut detektieren lassen. Der zweite Schwerpunkt im unteren Bereich des sequenzierten Repertoires etabliert sich also genau zu dem Zeitpunkt, an dem SRBC-spezifische Klonotypen verstärkt aus der Milz auswandern und dementsprechend im Blut detektiert werden können. Auch die Tatsache, dass im Bereich zwischen den beiden Schwerpunkten ($4 < \log_2(\mathrm{CN}) \leq 6$) zu keinem der gewählten Zeitpunkte signifikante Immunisierungseffekte auftreten (Abbildung 6), kann als ein Hinweis angesehen werden, dass die antigenspezifischen Klonotypen in den beiden Schwerpunktbereichen verschiedenen Ursprungs sind. Andernfalls wäre vielmehr eine kontinuierliche Durchsetzung des gesamten Repertoires zu erwarten gewesen. In Abschnitt 3.3 wurde versucht, durch die Entnahme weiterer Milzschnitte nachzuweisen, dass sich die Immunisierungseffekte im unteren Bereich tatsächlich auf wiedereingewanderte Klonotypen zurückführen lassen. Die Idee hinter dieser Vorgehensweise besteht darin, dass Nachkommen einer expandierten T-Zelle zunächst aus der Milz auswandern und sich mit dem Blutstrom über den Organismus verteilen. Wandert eine ausreichende Anzahl dieser Zellen, im Rahmen ihrer zufälligen Irrfahrt durch das lymphatische System [48], wieder in die Milz ein, wären sie möglicherweise in unterschiedlichen Teilen der Milz zu detektieren. Würde also die Schnittmenge zwischen den T-Zellrezeptorrepertoires aus zwei verschiedenen Milzschnitten deutlich höher ausfallen, falls diese vom selben Tier stammen, wäre dies ein eindeutiger Hinweis auf Migrationseffekte. Tatsächlich konnte ein solcher Effekt jedoch nicht festgestellt werden. Dies lässt sich mit der hohen Diversität der SRBC-spezifischen T-Zellantwort erklären. Die exorbitant hohe Anzahl verschiedener antigenspezifischer Klonotypen führt dazu, dass auch in Schnitten aus derselben Milz (abgesehen von naiven *public*-Klonotypen) kaum identische Sequenzen zu finden sind.

Zusammenfassend kann festgestellt werden, dass sich die SRBC-spezifische T-Zellreaktion in zwei Phasen einteilen lässt (siehe hierzu auch [51]). In einer frühen Expansionsphase kommt es zunächst zur Aktivierung von relativ wenigen Klonotypen,

welche mit einer massiven Proliferation reagieren. Dies führt zu der [47] beschriebenen Segregation einzelner Klonotypen. Eine signifikante Veränderung der Gesamtverteilung der Kopienzahlen ist ausschließlich in dieser Phase gegeben. Der Expansionsphase folgt eine weitere Phase, in welcher weite Teile des sequenzierten T-Zellrezeptorrepertoires mit reagiblen Klonotypen durchsetzt werden. Das Auftreten von Segregationseffekten setzt eine massive Expansion einiger weniger Klonotypen voraus, welche in dieser Phase nicht mehr gegeben ist. Dies erklärt, warum sich diese Effekte, ungeachtet der histologisch andauernden T-Zellreaktion [44] ab dem 4. Tag wieder zurückbilden. Im Unterschied zur Expansionsphase, kommt es während der Durchsetzungsphase kaum noch zu quantitativen Immunisierungseffekten, wie beispielsweise einer modifizierten Gesamtverteilung der Kopienzahlen. Die fortschreitende T-Zellreaktion lässt sich nur anhand qualitativer Merkmale (z.B. einer modifizierten Verteilung der detektierten V- bzw. J-Segmente) nachverfolgen.

Die nachgeordnete *public*-Komponente der SRBC-spezifischen T-Zellantwort unterscheidet sich in mehreren Punkten von der vorherrschenden *private*-response. Zu allen drei Zeitpunkten ist ein substantieller Anteil der identifizierten *public*-Klonotypen in einem Bereich lokalisiert, in dem sich keine Hinweise auf hohe Konzentrationen reagibler *private*-Klonotypen finden (vgl. Abbildungen 6 und 8). Der wohl offensichtlichste Unterschied zwischen den beiden Komponenten der T-Zellreaktion besteht darin, dass sich am 7. Tag nach Antigenapplikation, im Bereich der Klonotypen mit niedriger Kopienzahl ($\log_2(\mathrm{CN}) \leq 2$), keine Ansammlung von *public*-Klonotypen herausbildet, während insbesondere die neu eingeführten Kennzahlen auf eine hohe Konzentration von *private*-Klonotypen hindeuten. Allerdings weißt diese Beobachtung nicht auf ein biologisches Phänomen hin, sondern ist in der hier angewendeten Methodik begründet. Aktivierte *public*-T-Zellen welche aus dem Blutstrom erneut in die Milz einwandern treffen hier meist auf lokal expandierende „Artgenossen“, d.h. auf auf T-Zellen des gleichen Klonotyps. Mit der hier verwendeten Methodik können solche T-Zellen nicht unterschieden werden. Die extrahierten Sequenzen der wiedereingewanderten T-Zellen werden daher gemeinsam mit lokal Proliferierenden unmittelbar einer höheren Kategorie zugeordnet. Man beachte, dass die Ergebnisse im Hinblick auf *public*- und *private*-Komponente der SRBC-spezifischen T-Zellreaktion nur bedingt vergleichbar sind. Die *private*-Komponente lässt sich nur indirekt anhand von Diversifizierungs- bzw. Homogenisierungseffekten beobachten. Damit diese auftreten, muss ein substantieller Anteil

im jeweiligen Teilbereich aus spezifischen Klonotypen bestehen. Im Gegensatz hierzu können die identifizierten *public*-Klonotypen individuell nachverfolgt werden. Um die Anwesenheit solcher Klonotypen nachzuweisen ist daher nur ein einziger Klonotyp erforderlich.

4.3 Statistische Klassifikationsverfahren ermöglichen eine objektive Evaluierung der Immunisierungseffekte auf der Ebene des einzelnen Individuums

Die in dieser Arbeit vorgeschlagenen Klassifikationsverfahren dienen primär dazu, die einzelnen biologischen Parameter hinsichtlich ihrer Trennschärfe zwischen naivem und immunisiertem T-Zellrezeptorrepertoire zu bewerten. Der entscheidende Unterschied zu den übrigen Betrachtungen besteht darin, dass die Immunisierungseffekte in jedem einzelnen Tier über die Güte der Klassifikation entscheiden. Sind diese Effekte in einzelnen Tieren unzureichend ausgeprägt, kann dies nicht durch stärker ausfallende Effekte in anderen Tieren ausgeglichen werden und die jeweiligen Tiere werden fehlklassifiziert. Um eine Gesamtübersicht über die Beziehungen der einzelnen Repertoires zu erhalten, wurden die errechneten Distanzmatrizen mittels multidimensionaler Skalierung visualisiert und die Datensätze mit Hilfe einer Clusteranalyse in zwei Gruppen eingeteilt. Im Gegensatz zur Clusteranalyse, werden die Datensätze bei einer überwachten Klassifikation konkreten Gruppen („nicht immunisiert“ oder „immunisiert“) zugeordnet. Da hierbei zusätzliche Informationen (tatsächlicher Immunisierungsstatus der Trainingsdatensätze) genutzt werden, können hier meist bessere Ergebnisse als bei einer Clusteranalyse erzielt werden. Zunächst wurden Teilrepertoires betrachtet, in welchen aufgrund der vorangegangenen Analysen hohe Konzentrationen antigenspezifischer Klonotypen vermutet werden können. Damit T-Zellrezeptorrepertoires auf Basis eines biologischen Parameters zuverlässig klassifiziert werden können, muss folgende Bedingung erfüllt sein: Innerhalb einer experimentellen Gruppe müssen die Datensätze im Hinblick auf den Parameter einander möglichst ähnlich sein. Dies bedeutet, dass sie bezüglich des zugehörigen Distanzmaßes eng beieinander liegen. Umgekehrt müssen sich Datensätze aus verschiedenen experimentellen Gruppen möglichst deutlich voneinander unterscheiden, bzw. weit auseinander liegen. Die Ähnlichkeit der CDR3-Sequenzen im Hinblick auf die Levenshtein-Distanz erfüllt diese Bedingung nicht, da sich die immunisierten

Proben von den Naiven nicht stärker unterscheiden, als immunisierte Proben untereinander (Abbildung 9). Als Klassifikationskriterium ist dieser Parameter daher ungeeignet. Ein mögliches Kriterium zur Klassifikation stellen dagegen die detektierten V- bzw. J-Segmente dar. Im Bereich der Klonotypen mit hoher Kopienzahl ($\log_2(\mathrm{CN}) > 6$) ließen sich hiermit relativ gute Ergebnisse erzielen. Die meisten Fehlklassifikationen unterliefen hierbei innerhalb der 3d-Gruppe (Abbildung 10b). Dies kann als ein Hinweis dahingehend interpretiert werden, dass sich die T-Zellreaktion zu diesem frühen Zeitpunkt noch nicht in allen Tieren hinreichend ausgebildet hat. Dies deckt sich mit der Beobachtung, dass die SRBC-spezifischen *public*-Klonotypen in manchen Tieren der 3d-Gruppe bereits stark expandiert sind, während ihr Anteil in anderen Tieren weiterhin auf dem Ausgangsniveau verharrt (Abbildung 8a). Zusätzlich muss jedoch noch ein weiterer Aspekt berücksichtigt werden: Wie aus Abbildung 6 hervorgeht, wird das höchste Signifikanzniveau der Immunisierungseffekte, zu diesem Zeitpunkt, im Bereich der extrem hohen Kopienzahlen erreicht. In diesem Bereich sind, in allen vier Gruppen, vergleichsweise wenige Klonotypen zu finden (Abbildung 5a). Selbst wenn beispielsweise das Teilrepertoire der Klonotypen mit $\log_2(\mathrm{CN}) > 9$ zu einem substanziellen Anteil aus antigenspezifischen Klonotypen besteht, ist deren Anzahl, gemessen an den insgesamt detektierten Sequenzen äußerst gering. So liegt der Anteil besagter Klonotypen in allen 50 Proben unter 0,5%. Zu späteren Zeitpunkten sind offensichtlich (allein im Bereich der Klonotypen mit $\log_2(\mathrm{CN}) > 6$) deutlich mehr Klonotypen in die Reaktion involviert, sodass eine zuverlässigere Klassifikation möglich ist. Betrachtet man den unteren Bereich der sequenzierten Repertoire ($\log_2(\mathrm{CN}) \leq 2$) können nur die Tiere der 7d-Gruppe von der Kontrollgruppe unterschieden werden. Dies deckt sich mit den Ergebnissen aus Abschnitt 3.3, wonach sich in diesem Bereich, (abgesehen von dezenten Effekten an Tag 4) erst am 7. Tag deutliche Immunisierungseffekte zeigen. Interessanterweise lassen sich diese Teilrepertoires auf Basis der CDR3-Nukleotidkodierungen deutlich besser klassifizieren als dies auf Basis der Segmente möglich ist. Am deutlichsten zeigt sich dies an den Ergebnissen der Clusteranalyse. Während diese bei Anwendung des Segmentkriteriums keine der experimentellen Gruppen von den übrigen Datensätzen abzugrenzen vermag (Abbildung 10c), führt die Anwendung des Nukleotidkriteriums dazu, dass die 7d-Gruppe als zusammengehöriger Cluster erkannt und von den übrigen Proben abgegrenzt wird (Abbildung 11c). Auch das überwachte Klassifikationsverfahren liefert bei Anwendung dieses Kriteriums ein besseres Ergebnis. Vor diesem Hintergrund

erscheint es zunächst überraschend, dass die Anwendung des Nukleotidkriteriums im oberen Bereich des sequenzierten Repertoires eine chaotische Clusterstruktur erzeugt, während die Anwendung des Segmentkriterium zu einer relativ guten Einteilung führt. Eine Erklärung für diese Ergebnisse kann Abbildung 13 entnommen werden. Zunächst muss festgehalten werden, dass die Definition der Teilrepertoires mit hoher Kopienzahl ($CN > 64$, d.h. $\log_2(CN) > 6$) für die Anwendung des Nukleotidkriteriums nicht optimal ist. Wie aus Abbildung 6 hervorgeht, wird die Heterogenität der Nukleotidkodierungen im Bereich $6 < \log_2(CN) \leq 8$ durch die Immunisierung nicht in signifikanter Weise verändert. Dies bedeutet, dass die im Bereich der sehr hohen Kopienzahlen ($CN > 256$, d.h. $\log_2(CN) > 8$) festgestellten Immunisierungseffekte durch Hinzunahme weiterer Klonotypen verdünnt werden. Dies führt dazu, dass die vorhandenen Immunisierungseffekte von der allgemeinen Streuung überdeckt werden (Abbildung 13a) Man beachte, dass auf Basis des Nukleotidkriteriums etwas bessere Klassifikationsergebnisse erzielt werden können, wenn Teilrepertoires mit $CN > 256$ betrachtet werden. Da die Ergebnisse jedoch weiterhin äußerst unbefriedigend ausfallen, wird hierauf nicht weiter eingegangen. Im unteren Bereich des sequenzierten Repertoires ($CN \leq 4$, d.h. $\log_2(CN) \leq 2$) ergibt sich dagegen folgendes Bild: Im naiven Zustand werden $\sim 93\%$ der detektierten CDR3-Aminosäuresequenzen von lediglich einer einzigen Nukleotidsequenz kodiert. Nahezu der gesamte Rest ($\sim 7\%$) geht auf zwei Nukleotidsequenzen zurück. Am 7. Tag nach Antigenapplikation sinkt dieser Anteil auf $\sim 3\%$ ab. Da diese Teilrepertoires sehr viele Klonotypen enthalten (siehe Abbildung 5a) sind innerhalb der einzelnen Gruppen nur noch sehr geringe Streuungen vorhanden. Daher ermöglicht der (in absoluten Zahlen) sehr dezent ausgeprägte Immunisierungseffekt eine zuverlässige Klassifikation der 7d-Gruppe.

Betrachtet man die Verteilung der V- bzw. J-Segmente stellt man fest, dass die Klonotypen mit hoher Kopienzahl ($CN > 64$) zu einem großen Teil auf wenige Segmente zurückgehen, was auf einen relativ hohen Anteil an *public*-Klonotypen schließen lässt [25]. Dieser Effekt wird durch die Immunisierung und die damit verbundene Expansion von *private*-Klonotypen deutlich abgeschwächt, bleibt jedoch weiterhin bestehen. Die Trennschärfe der Klassifikation kann somit mit der besonderen Zusammensetzung der Segmente innerhalb der *public*-Klonotypen erklärt werden. Im unteren Bereich des sequenzierten Repertoires sind auch im naiven Zustand nur wenige *public*-Klonotypen zu finden. Die Zusammensetzung der Segmente ist daher insgesamt

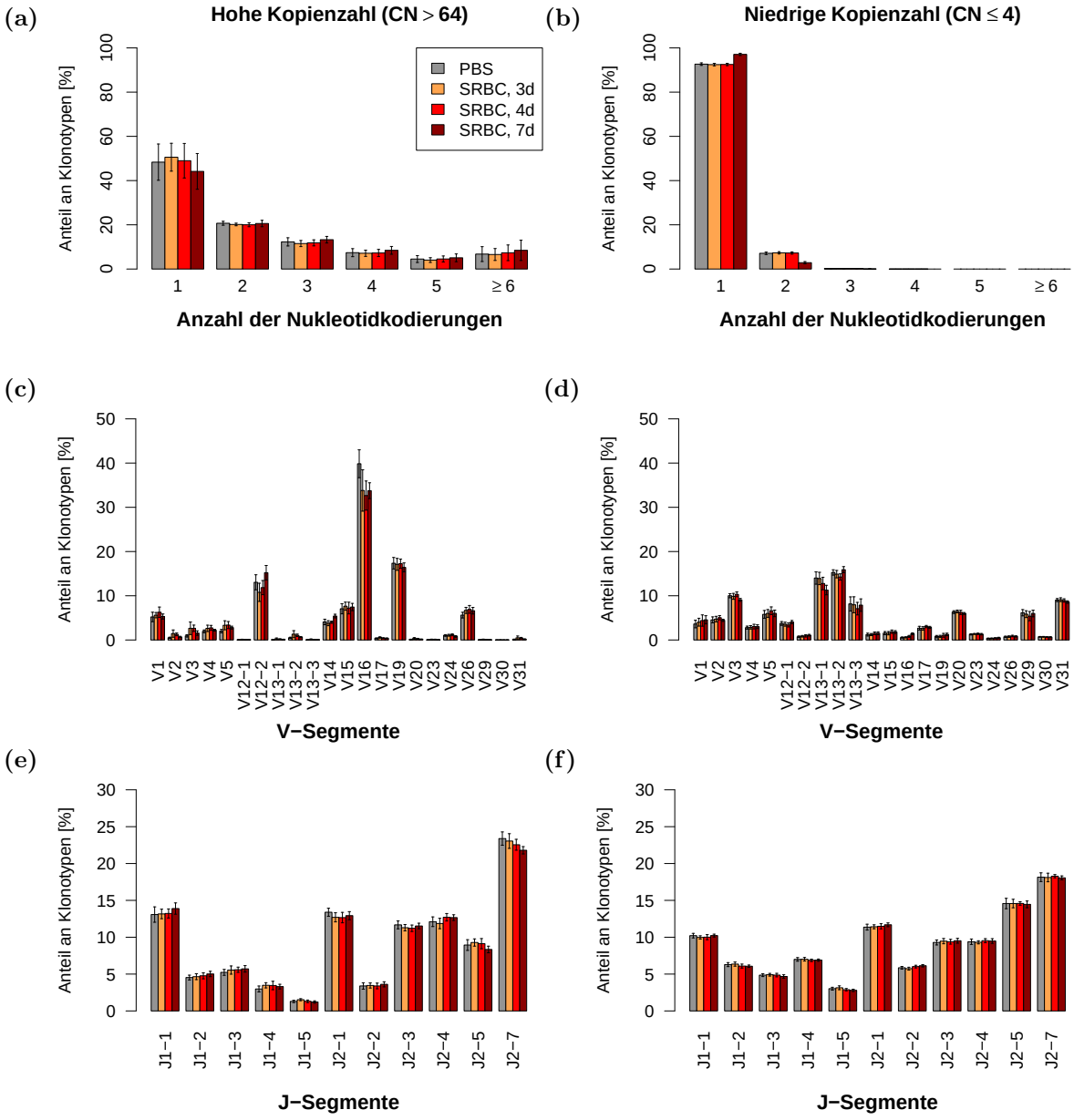

Abbildung 13: Detailbetrachtung der zur Klassifikation genutzten Parameter. Vergleich der Teilrepertoires mit hoher (CN $>$ 64, links) und niedriger Kopienzahl (CN $\leq$ 4, rechts). (a)-(b) Anzahl der Nukleotidkodierungen der CDR3-Aminosäuresequenzen. (c)-(d) Verteilung der detektierten V-Segmente. (e)-(f) Verteilung der detektierten J-Segmente. Die Barplots zeigen Mittelwerte, die Fehlerbalken die zugehörigen Standardabweichungen an ($n = 20$ für Kontrollgruppe, $n = 10$ für jede der immunisierten Gruppen).

heterogener und wird weniger von einzelnen Segmenten dominiert. Das Auftreten antigenspezifischer Klonotypen an Tag 7 führt also in erster Linie dazu, dass naive *private*-Klonotypen von antigenspezifischen *private*-Klonotypen verdrängt werden. Dies führt zu deutlich geringeren Auswirkungen auf die Zusammensetzung der Segmente und damit zu einer geringeren Trennschärfe des Klassifikationsverfahrens.

Sollen auf Basis des Segmentkriteriums ganze Datensätze, anstelle von Teilrepertoires, klassifiziert werden, müssen die Klonotypen nach Kopienzahl gewichtet werden (Abbildung 12). Hierdurch wird der Schwerpunkt der Analyse auf expandierende (und damit häufige) Klonotypen gerichtet. Die Klassifikationsverfahren liefern daher eine ähnliche Einteilung wie bei der Klassifikation der Teilrepertoires mit hoher Kopienzahl ($CN > 64$). Werden alle Klonotypen gleich stark gewichtet, führen die vorgeschlagenen Verfahren zu chaotischen Einteilungen. Dies gilt selbst für die Tiere der 7d-Gruppe, obwohl zu diesem Zeitpunkt weite Teile des Repertoires, einschließlich Klonotypen mit niedriger Kopienzahl, von reagiblen Klonotypen durchsetzt sind.

Im Hinblick auf die Trennschärfe der hier vorgeschlagenen Klassifikationsverfahren ist Folgendes zu beachten: Sowohl die Klassifikationskriterien als auch die für die Klassifikationsexperimente betrachteten Teilrepertoires wurden auf Basis der gesamten Datengrundlage (50 Milzschnitte) festgelegt. Dieselben Datensätze wurden anschließend mit Hilfe der neu entwickelten Verfahren klassifiziert. Dies bedeutet, dass Informationen über die zu klassifizierenden Daten bereits bei der Definition der Algorithmen verwendet wurden. Ein solches Vorgehen kann dazu führen, dass die Trennschärfe eines Klassifikationsverfahrens erheblich überschätzt wird. Dies ist insbesondere dann gegeben, wenn zufällige Effekte, welche sich bei einer erneuten Datenerhebung nicht reproduzieren lassen, in die Klassifikation miteinfließen. Diese Problematik wird in [41] ausführlich erörtert. Um die Trennschärfe des der Algorithmen unverfälscht zu evaluieren, müssten diese auf unabhängige Datensätze angewendet werden. Dennoch zeigen die Beispiele in dieser Arbeit, dass sich bei Anwendung eines geeigneten biologischen Kriteriums immunisierte Tiere relativ zuverlässig von Naiven unterscheiden lassen, sofern die Teilbereiche des Repertoires, in denen sich reagible Klonotypen konzentrieren, gezielt betrachtet werden.

4.4 Möglichkeiten und Grenzen der vorgeschlagenen Methodik

Ziel dieser Arbeit war es, eine T-Zellreaktion auf Rezeptorebene zu analysieren, bei der die expandierenden Klonotypen ein Höchstmaß an Diversität und Inhomogenität aufweisen. Durch die Fokussierung auf die Eigenschaften des temporär supprimierten naiven T-Zellrezeptorrepertoires, ist die hier vorgeschlagene Methodik relativ unabhängig von der Struktur der expandierenden Klonotypen. Im Kern beruht diese Methodik auf einem Kennzahlsystem, welches es ermöglicht, sowohl die Homogenität eines T-Zellrezeptorrepertoires als auch Ähnlichkeitsaspekte zwischen verschiedenen Repertoires zu quantifizieren. Dieses System kann direkt von etablierten Standardkennzahlen abgeleitet werden, welche sich vielfach bei der Analyse des T-Zellrezeptorrepertoires bewährt haben. Die entscheidende Verallgemeinerung besteht darin, dass nicht mehr das Auftreten von Klonotypen mit identischer CDR3-Region quantifiziert wird, sondern vielmehr Solche, welche im Hinblick auf ein beliebiges Kriterium als ähnlich angesehen werden. Dieses Kriterium, welches formal eine symmetrische, reflexive Relation darstellt, kann flexibel gewählt werden und richtet sich in erster Linie nach der biologischen Fragestellung. Diese zusätzliche Flexibilität stellt den entscheidenden Vorteil der vorgeschlagenen Kennzahlen dar. Ihr schwerwiegendster Nachteil besteht in einem exorbitant hohen Rechenaufwand, welcher erforderlich ist, sobald den Kennzahlen eine nicht-transitive Relation zugrunde gelegt wird. In diesem Fall müssen alle detektierten Klonotypen dahingehend überprüft werden, ob sie in Relation zueinander stehen. Falls die zugrundeliegende Relation jedoch transitiv ist, kann die Berechnung der Kennzahlen mit Hilfe der ursprünglichen Formeln (Simpson- bzw. Morisita-Horn-Index) erfolgen. In diesem Fall stellen die hier vorgeschlagenen Kennzahlen lediglich eine Anwendung einer etablierten statistischen Methodik dar. So wird beispielsweise der Simpson-Index in [15] genutzt, um die Homogenität bzw. Diversität der V-Segmente in einer Menge vorselektierter Klonotypen zu quantifizieren. Zusätzlich zur massiven Reduktion der Rechenzeit gewährleistet die Anwendung transitiver Relationen verschiedene analytische Eigenschaften der Kennzahlen, welche für manche Anwendungen wünschenswert sein können (siehe Abschnitt 2.3.3). Die einzige in dieser Arbeit verwendete nicht transitive Relation, ist die Ähnlichkeit der CDR3-Sequenzen im Hinblick auf die Levenshtein-Distanz. Ein weiteres Beispiel für eine nicht transitive Relation wäre, zwei Sequenzen als ähnlich (d.h. in Relation stehend) zu definieren, falls im zentra-

len Bereich ihrer CDR3-Region übereinstimmende Teilsequenzen mit einer gewissen Mindestlänge (z.B. drei Aminosäuren) zu finden sind. Solche lokalen Muster wurden beispielsweise in [15] zur Klassifikation von Lymphozytenpopulationen, im Hinblick auf die Antigenspezifität, genutzt. Am Institut für Anatomie der Universität zu Lübeck kommt neben dem SRBC-Modell noch ein weiteres Mausmodell zur Anwendung, bei welchem die Tiere mit Ovalbumin, in Kombination mit einem Adjuvans (Aluminiumhydroxid), immunisiert werden[1]. In diesem Mausmodell konnten die entnommenen Milzschnitte, mittels der hier vorgeschlagenen Methodik, nahezu fehlerfrei auf Basis lokaler CDR3-Muster klassifiziert werden (zwei Klonotypen wurden als ähnlich definiert, falls sich im Zentrum der CDR3-Region eine gemeinsame Sequenz mit einer Länge von mindestens 3 Aminosäuren befindet). Für die Klassifikation von Milzschnitten im SRBC-Modell erwies sich dieses Kriterium jedoch als ungeeignet, es kommt daher in dieser Arbeit nicht zur Anwendung. Diese Beobachtung verdeutlicht jedoch, wie inhomogen die SRBC-spezifischen Klonotypen im Vergleich zu anderen Mausmodellen sind.

In Kombination mit der in Abschnitt 2.3.4 vorgeschlagenen Aufteilung des extrahierten T-Zellrezeptorrepertoires, ermöglichen es die neuen Kennzahlen, Teilbereiche zu identifizieren, in welchen sich die Repertoires der immunisierten Tiere signifikant von denen der Kontrollgruppe unterscheiden. Diese können als Schwerpunktbereiche der Immunreaktion interpretiert werden, in welchen sich mutmaßlich reagible Klonotypen konzentrieren. Es stellte sich hierbei heraus, dass die verbreitete Vorstellung, nach welcher sich reagible Klonotypen, aufgrund der klonalen Expansion, grundsätzlich durch hohe Kopienzahlen auszeichnen, im Allgemeinen nicht zutrifft. Vielmehr zeigte sich, dass die Schwerpunktbereiche im Lauf der Immunreaktion einer kontinuierlichen Dynamik unterworfen sind. So können insbesondere auch Klonotypen mit extrem niedriger Kopienzahl bei der Analyse von Immunisierungseffekten von entscheidender Bedeutung sein. Beispielsweise wird in [46] ein Algorithmus zur Klassifikation von T-Zellrezeptorrepertoires vorgeschlagen. Die Autoren extrahierten CDR3-Sequenzen aus der Milz und versuchten, die Datensätze im Hinblick auf applizierte Antigene zu klassi-

[1]Die Datensätze zu diesem Mausmodell entstammen einer Kooperation des Institutes für Ernährungswissenschaften, unter Leitung von Prof. Dr. Marc Ehlers, mit dem Institut für Anatomie der Universität zu Lübeck. Details zu diesem Mausmodel sind in [4] zu finden. Die zugehörigen Datenanalysen sind nicht Teil dieser Dissertation. Sie werden hier nur erwähnt, um einen Ausblick auf mögliche weitere Anwendungen der hier vorgeschlagenen Algorithmen zu geben.

fizieren. Ihr Algorithmus erreichte hierbei eine beachtliche Trennschärfe, sofern lediglich Klonotypen mit Kopienzahl 1 in die Analyse miteinbezogen wurden. Wurden dagegen Klonotypen mit hohen Kopienzahlen betrachtet, war das Ergebnis nicht signifikant besser als eine zufällige Klassifikation. Diese Beobachtungen decken sich gut mit den Ergebnissen dieser Arbeit. Außerdem verdeutlichen sie die Vorteile einer Analyse, welche sich auf Schwerpunktbereiche der T-Zellreaktion konzentriert. Die Nachteile des in dieser Arbeit vorgeschlagenen Verfahrens zur Identifikation solcher Schwerpunktbereiche liegen auf der Hand: Treten in bestimmten Teilbereichen signifikante Abweichungen zwischen immunisierten Tieren und der Kontrollgruppe auf, so kann dies zwar ein Hinweis auf Ansammlungen reagibler Klonotypen sein, tatsächlich können solche Effekte jedoch auch andere Ursachen haben. So wäre es beispielsweise denkbar, dass die Expansion antigenspezifischer Klonotypen dazu führt, dass andere (naive) Klonotypen mit deutlich geringerer Kopienzahl sequenziert werden, als dies bei einem nicht immunisierten Tier der Fall gewesen wäre. Folglich würden diese Klonotypen in dem in Abbildung 6 dargestellten Schema einer niedrigeren Kategorie zugeteilt werden. Durch solche Verdrängungseffekte kann die Zusammensetzung in den unteren Kategorien modifiziert werden, obwohl die hinzukommenden Klonotypen nicht antigenspezifisch sind. Die Anwesenheit reagibler Klonotypen könnte durch solche Effekte gleichermaßen verdeckt als auch vorgetäuscht werden. Ein weiterer Nachteil besteht darin, dass Ansammlungen reagibler Klonotypen in einem bestimmten Teilbereich nur dann erkannt werden, wenn sich diese, im Hinblick auf den betrachteten Parameter, von den entsprechenden Teilrepertoires der Kontrollgruppe unterscheiden. Die Qualität des Ergebnisses hängt also entscheidend von der Wahl dieses Parameters ab. Der wohl gravierendste Nachteil besteht jedoch darin, dass das Verfahren lediglich Aussagen über Teilbereiche zulässt. Insbesondere können hieraus keine Informationen über die Spezifität der einzelnen Klonotypen abgeleitet werden.

Zusammenfassend lässt sich feststellen, dass es mit der in dieser Arbeit vorgeschlagenen Methodik möglich ist, eine hochgradig diverse T-Zellreaktion über einen längeren Zeitraum nachzuverfolgen. Die hier vorgeschlagene Aufteilung des T-Zellrezeptorrepertoires ermöglicht eine fokussierte Analyse einzelner Teilbereiche, ohne hierfür eine willkürliche Auswahl der Klonotypen treffen zu müssen. Die Vermeidung einer subjektiven Auswahl (wie beispielsweise die verbreitete Betrachtung der 100, 1000 oder 5000 Klonotypen mit den höchsten Kopienzahlen [20, 46, 51, 50]) führt nicht nur zu

einer Vereinfachung der Datenanalyse, sondern auch zu einer nicht zu unterschätzenden Objektivierung der Ergebnisse. Bei der Analyse von T-Zellreaktionen bietet sich eine solche Vorgehensweise insbesondere als Screeningverfahren an. Sofern es hierbei gelingt, Schwerpunktbereiche der Immunreaktion zu identifizieren, können diese anschließend mit spezialisierteren Verfahren, welche sich nach der jeweiligen Fragestellung richten, analysiert werden. Hierbei muss jedoch berücksichtigt werden, dass die hier vorgeschlagene Methodik bisher fast ausschließlich auf ein einziges Mausmodell angewendet wurde. Die Frage, inwieweit sich diese Ergebnisse auf andere biologische Gegebenheiten übertragen lassen, bedarf einer gesonderten Untersuchung.

5 Zusammenfassung

Durch große zelluläre Antigene kann eine T-Zellantwort von hoher Diversität und Komplexität hervorgerufen werden. Durch das weitgehende Fehlen gemeinsamer Strukturen innerhalb der T-Zellrezeptoren der expandierenden Klonotypen wird die Analyse einer solchen Reaktion auf Rezeptorebene erheblich erschwert. Beispielhaft wurde in dieser Arbeit ein Mausmodell betrachtet, bei welchem durch intravenöse Applikation von Schaferythrozyten eine T-Zellreaktion ausgelöst wird. Anhand dieses Modells wurden Verallgemeinerungen ökologischer Kennzahlen entwickelt, welche sich vielfach zur Analyse des T-Zellrezeptorrepertoires bewährt haben. Mit Hilfe der verallgemeinerten Kennzahlen können sowohl die Homogenität bzw. Diversität innerhalb eines Repertoires als auch Ähnlichkeitsaspekte zwischen zwei verschiedenen Repertoires quantifiziert werden. Homogenität und Ähnlichkeit können hierbei im Hinblick auf nahezu beliebige biologische Parameter analysiert werden. Exemplarisch wurden hier die Ähnlichkeit der CDR3-Region, die detektierten V- bzw. J-Segmente, sowie die Diversität der Nukleotidkodierungen der einzelnen CDR3-Aminosäuresequenzen betrachtet. Im untersuchten Mausmodell führt das applizierte Antigen, in den einzelnen Tieren, primär zur Expansion individueller Klonotypen (*private response*), welche sich im Hinblick auf diese Parameter signifikant vom naiven T-Zellrezeptorrepertoire unterscheiden. Durch eine systematische Unterteilung des sequenzierten Repertoires war es möglich, Schwerpunktbereiche zu identifizieren, in denen es zu einer transienten Verdrängung des naiven T-Zellrezeptorrepertoires kommt. In diesen Bereichen können hohe Konzentrationen antigenspezifischer Klonotypen vermutet werden können. Die T-Zellreaktion zeigte hierbei einen phasenhaften Verlauf. Während zu Beginn der Reaktion einige wenige Klonotypen stark expandieren, kommt es zu späteren Zeitpunkten zu einer kontinuierlichen Durchsetzung des Repertoires mit antigenspezifischen Klonotypen. Die einzelnen Klonotypen zeigen in dieser späteren Phase ein weniger offensives Expansionsverhalten als zu Beginn der Reaktion. Mit Hilfe einer Genexpressionsanalyse wurden einige wenige Klonotypen identifiziert, welche in der Mehrzahl der immunisierten Tiere expandieren (*public response*). Diese waren nahezu über das gesamte sequenzierte Repertoire verteilt. Insbesondere zeigten sie keine eindeutigen Konzentrationstendenzen innerhalb der identifizierten Schwerpunktbereiche. Mit Hilfe zweier statistischer Lernverfahren wurde untersucht, auf Basis welcher Parameter sich immunisierte T-Zellrezeptorrepertoires

von denen der Kontrollgruppe unterscheiden lassen. Diese Experimente zeigen beispielhaft, inwiefern die Auswahl geeigneter Teilrepertoires für die Beobachtung einer T-Zellreaktion auf Rezeptorebene ausschlaggebend sein kann.

Zusammenfassend lässt sich feststellen, dass die hier vorgeschlagene Methodik eine relativ detaillierte Analyse der untersuchten T-Zellreaktion ermöglicht. Insbesondere die Möglichkeit, Schwerpunktbereiche einer Reaktion innerhalb des sequenzierten Repertoires zu identifizieren, um diese anschließend genauer zu analysieren, könnte sich für künftige Forschungsarbeiten als hilfreich erweisen.

Literatur

[1] R. Antonacci, S. Di Tommaso, C. Lanave, E. P. Cribiu, S. Ciccarese und S. Massari. Organization, structure and evolution of 41kb of genomic DNA spanning the D-J-C region of the sheep TRB locus. *Molecular Immunology*, 45:493–509, 2008.

[2] Apache Commons Text, Version 1.4, Apache Software Foundation, 2018.

[3] M. Attaf, M. Legut, D. K. Cole und A. K. Sewell. The T cell antigen receptor: the Swiss army knife of the immune system. *Clinical and Experimental Immunology*, 181:1–18, 2015.

[4] Y. C. Bartsch, S. Eschweiler, A. Leliavski, H. B. Lunding, S. Wagt, J. Petry, G.-M. Lilienthal, J. Rahmöller, N. de Haan, A. Hölscher, R. Erapaneedi, A. D. Giannou, L. Aly, R. Sato, L. A. de Neef, A. Winkler, D. Braumann, J. Hobusch, K. Kuhnigk, V. Krémer, M. Steinhaus, V. Blanchard, T. Gemoll, J. K. Habermann, M. Collin, G. Salinas, R. A. Manz, H. Fukuyama, T. Korn, A. Waisman, N. Yogev, S. Huber, B. Rabe, S. Rose-John, H. Busch, F. Berberich-Siebelt, C. Hölscher, M. Wuhrer und M. Ehlers. IgG Fc sialylation is regulated during the germinal center reaction following immunization with different adjuvants. *The Journal of Allergy and Clinical Immunology*, 146:652–666.e11, 2020.

[5] Y. Benjamini und Y. Hochberg. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B*, 57:289–300, 1995.

[6] K. Best, T. Oakes, J. M. Heather, J. Shawe-Taylor und B. Chain. Computational analysis of stochastic heterogeneity in PCR amplification efficiency revealed by single molecule barcoding. *Scientific Reports*, 5:14629, 2015.

[7] D. A. Bolotin, M. Shugay, I. Z. Mamedov, E. V. Putintseva, M. A. Turchaninova, I. V. Zvyagin, O. V. Britanova und D. M. Chudakov. MiTCR: software for T-cell receptor sequencing data analysis. *Nature Methods*, 10:813–814, 2013.

[8] N. Chaudhary und D. R. Wesemann. Analyzing Immunoglobulin Repertoires. *Frontiers in Immunology*, 9:462, 2018.

[9] Y. Chen, A. T. L. Lun und G. K. Smyth. Differencial Expression Analysis of Complex RNA-seq Experiments Using edgeR. In S. Datta und D. Nettleton, Herausgeber, *Statistical Analysis of next Generation Sequencing Data*, S. 51–74. Springer, Heidelberg, 2014.

[10] S. Crotty. T Follicular Helper Cell Biology: A Decade of Discovery and Diseases. *Immunity*, 50:1132–1148, 2019.

[11] M. M. Davis und P. J. Bjorkman. T-cell antigen receptor genes and T-cell recognition. *Nature*, 334:395–402, 1988.

[12] Y. Elhanati, Z. Sethna, C. G. Callan Jr, T. Mora und A. M. Walczak. Predicting the spectrum of TCR repertoire sharing with a data-driven model of recombination. *Immunological Reviews*, 284:167–179, 2018.

[13] A. Fähnrich, M. Krebbel, N. Decker, M. Leucker, F. D. Lange, K. Kalies und S. Möller. ClonoCalc and ClonoPlot: immune repertoire analysis from raw files to publication figures with graphical user interface. *BMC Bioinformatics*, 18:164, 2017.

[14] M. W. Garratt und R. K. Steinhorst. Testing for Significance of Morisita's, Horn's and Related Measures of Overlap. *The American Midland Naturalist*, 96:245–251, 1976.

[15] J. Glanville, H. Huang, A. Nau, O. Hatton, L. Wagar, F. Rubelt, X. Ji, A. Han, S. Krams, C. Pettus, N. Haas, C. Arlehamn, A. Sette, S. Boyd, T. Scriba, O. Martinez und M. Davis. Identifying specificity groups in the T cell receptor repertoire. *Nature*, 547:94–98, 2017.

[16] E. Guzman, J. Hope, G. Taylor, A. L. Smith, C. Cubillos-Zapata und B. Charleston. Bovine γδ T Cells Are a Major Regulatory T Cell Subset. *The Journal of Immunology*, 193:208–222, 2014.

[17] W. Härdle und L. Simar. Multidimensional Scaling. In *Applied Multivariate Statistical Analysis*, S. 455–472. Springer, Heidelberg, 4. Aufl., 2015.

[18] T. Hastie, R. Tibshirani und J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Series in Statistics. Springer, New York, 2. Aufl., 2009.

[19] H. S. Horn. Measurement of Overlap in Comparative Ecological Studies. *The American Naturalist*, 100:419–424, 1966.

[20] M. Izraelson, T. O. Nakonechnaya, B. Moltedo, E. S. Egorov, S. A. Kasatskaya, E. V. Putintseva, I. Z. Mamedov, D. B. Staroverov, I. I. Shemiakina, M. Y. Zakharova, A. N. Davydov, D. A. Bolotin, M. Shugay, D. M. Chudakov, A. Y. Rudensky und O. V. Britanova. Comparative analysis of murine T-cell receptor repertoires. *Immunology*, 153:133–144, 2018.

[21] D. A. Jackson, K. M. Somers und H. H. Harvey. Similarity Coefficients: Measures of Co-Occurrence and Association or Simply Measures of Occurrence? *The American Naturalist*, 133:436–453, 1989.

[22] H. Kaufmann und H. Pape. Clusteranalyse. In L. Fahrmeir, H. W. Brachinger und G. Tutz, Herausgeber, *Multivariate statistische Verfahren*, S. 437–536. de Gruyter, Berlin, 2. Aufl., 1996.

[23] T. Lange, J. Born und J. Westermann. Sleep Matters: CD4+ T Cell Memory Formation and the Central Nervous System. *Trends in Immunology*, 40:674–686, 2019.

[24] A. Madi, A. Poran, E. Shifrut, S. Reich-Zeliger, E. Greenstein, I. Zaretsky, T. Arnon, F. V. Laethem, A. Singer, J. Lu, P. D. Sun, I. R. Cohen und N. Friedman. T cell receptor repertoires of mice and humans are clustered in similarity networks around conserved public CDR3 sequences. *eLife*, 6:e22057, 2017.

[25] A. Madi, E. Shifrut, S. Reich-Zeliger, H. Gal, K. Best, W. Ndifon, B. Chain, I. R. Cohen und N. Friedman. T-cell receptor repertoires share a restricted set of public and abundant CDR3 sequences that are associated with self-related immunity. *Genome Research*, 24:1603–1612, 2014.

[26] M. Maechler, P. Rousseeuw, A. Struyf, M. Hubert und K. Hornik. Cluster: Cluster Analysis Basics and Extensions, Version 2.1.3, R-Package, 2022.

[27] A. Mayer, Y. Zhang, A. S. Perelson und N. S. Wingreen. Regulation of T cell expansion by antigen presentation dynamics. *Proceedings of the National Academy of Sciences*, 116:5914–5919, 2019.

[28] A. Mead. Review of the Development of Multidimensional Scaling Methods. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 41:27–39, 1992.

[29] M. Meinhardt, C. Tune, L.-K. Schierloh, A. Schampel, R. Pagel und J. Westermann. The splenic T cell receptor repertoire during an immune response against a complex antigen: Expanding private clones accumulate in the high and low copy number region. *PLOS ONE*, 17:e0273264, 2022.

[30] A. Minervina, M. Pogorelyy und I. Mamedov. T-cell receptor and B-cell receptor repertoire profiling in adaptive immunity. *Transplant International*, 32:1111–1123, 2019.

[31] K. Murphy und C. Weaver. *Janeway's Immunobiology*. Garland Science/Taylor & Francis Group, LLC, New York, 9. Aufl., 2016.

[32] G. Navarro. A guided tour to approximate string matching. *ACM Computing Surveys*, 33:31–88, 2001.

[33] J. Nikolich-Žugich, M. K. Slifka und I. Messaoudi. The many important facets of T-cell repertoire diversity. *Nature Reviews Immunology*, 4:123–132, 2004.

[34] R: A Language and Environment for Statistical Computing, Version 4.2.1, R Foundation for Statistical Computing, 2022.

[35] A. P. Reynolds, G. Richards, B. de la Iglesia und V. J. Rayward-Smith. Clustering Rules: A Comparison of Partitioning and Hierarchical Clustering Algorithms. *Journal of Mathematical Modelling and Algorithms*, 5:475–504, 2006.

[36] M. D. Robinson, D. J. McCarthy und G. K. Smyth. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26:139–140, 2010.

[37] M. D. Robinson und A. Oshlack. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology*, 11:R25, 2010.

[38] E. Rosati, C. M. Dowds, E. Liaskou, E. K. K. Henriksen, T. H. Karlsen und A. Franke. Overview of methodologies for T-cell receptor repertoire analysis. *BMC Biotechnology*, 17:61, 2017.

[39] E. Schubert und P. J. Rousseeuw. Fast and eager k-medoids clustering: O(k) runtime improvement of the PAM, CLARA, and CLARANS algorithms. *Information Systems*, 101:101804, 2021.

[40] M. Shugay, D. V. Bagaev, M. A. Turchaninova, D. A. Bolotin, O. V. Britanova, E. V. Putintseva, M. V. Pogorelyy, V. I. Nazarov, I. V. Zvyagin, V. I. Kirgizova, K. I. Kirgizov, E. V. Skorobogatova und D. M. Chudakov. VDJtools: Unifying Post-analysis of T Cell Receptor Repertoires. *PLOS Computational Biology*, 11:e1004503, 2015.

[41] R. Simon, M. D. Radmacher, K. Dobbin und L. M. McShane. Pitfalls in the Use of DNA Microarray Data for Diagnostic and Prognostic Classification. *Journal of the National Cancer Institute*, 95:14–18, 2003.

[42] E. H. Simpson. Measurement of Diversity. *Nature*, 163:688–688, 1949.

[43] A. Six, M. E. Mariotti-Ferrandiz, W. Chaara, S. Magadan, H.-P. Pham, M.-P. Lefranc, T. Mora, V. Thomas-Vaslin, A. M. Walczak und P. Boudinot. The Past, Present, and Future of Immune Repertoire Biology – The Rise of Next-Generation Repertoire Analysis. *Frontiers in Immunology*, 4:413, 2013.

[44] C. Stamm, J. Barthelmann, N. Kunz, K.-M. Toellner, J. Westermann und K. Kalies. Dose-Dependent Induction of Murine Th1/Th2 Responses to Sheep Red Blood Cells Occurs in Two Steps: Antigen Presentation during Second Encounter Is Decisive. *PLOS ONE*, 8:e67746, 2013.

[45] S. Suerbaum und R. Blasczyk. Immunolgie. In S. Suerbaum, G. D. Burchard, S. H. E. Kaufmann und T. F. Schulz, Herausgeber, *Medizinische Mikrobiologie und Infektiologie*, Springer-Lehrbuch, S. 37–120. Springer, Berlin, 8. Aufl., 2016.

[46] Y. Sun, K. Best, M. Cinelli, J. M. Heather, S. Reich-Zeliger, E. Shifrut, N. Friedman, J. Shawe-Taylor und B. Chain. Specificity, Privacy, and Degeneracy in the CD4 T Cell Receptor Repertoire Following Immunization. *Frontiers in Immunology*, 8:430, 2017.

[47] J. Textor, A. Fähnrich, M. Meinhardt, C. Tune, S. Klein, R. Pagel, P. König, K. Kalies und J. Westermann. Deep Sequencing Reveals Transient Segregation of T Cell Repertoires in Splenic T Cell Zones during an Immune Response. *The Journal of Immunology*, 201:350–358, 2018.

[48] J. Textor, S. E. Henrickson, J. N. Mandl, U. H. von Andrian, J. Westermann, R. J. de Boer und J. B. Beltman. Random Migration and Signal Integration Promote Rapid and Robust T Cell Recruitment. *PLOS Computational Biology*, 10:e1003752, 2014.

[49] N. Thomas, K. Best, M. Cinelli, S. Reich-Zeliger, H. Gal, E. Shifrut, A. Madi, N. Friedman, J. Shawe-Taylor und B. Chain. Tracking global changes induced in the CD4 T-cell receptor repertoire by immunization with a complex antigen using short stretches of CDR3 protein sequence. *Bioinformatics*, 30:3181–3188, 2014.

[50] C. Tune, J. Hahn, S. E. Autenrieth, M. Meinhardt, R. Pagel, A. Schampel, L.-K. Schierloh, K. Kalies und J. Westermann. Sleep restriction prior to antigen exposure does not alter the T cell receptor repertoire but impairs germinal center formation during a T cell-dependent B cell response in murine spleen. *Brain, Behavior, & Immunity - Health*, 16:100312, 2021.

[51] C. Tune, M. Meinhardt, K. Kalies, R. Pagel, L.-K. Schierloh, J. Hahn, S. E. Autenrieth, C. E. Koch, H. Oster, A. Schampel und J. Westermann. Effects of sleep on the splenic milieu in mice and the T cell receptor repertoire recruited into a T cell dependent B cell response. *Brain, Behavior, & Immunity - Health*, 5:100082, 2020.

[52] A. J. Van den Eertwegh, W. J. Boersma und E. Claassen. Immunological functions and in vivo cell-cell interactions of T cells in the spleen. *Critical Reviews in Immunology*, 11:337–380, 1992.

[53] V. Venturi, K. Kedzierska, D. A. Price, P. C. Doherty, D. C. Douek, S. J. Turner und M. P. Davenport. Sharing of T cell receptors in antigen-specific responses is driven by convergent recombination. *Proceedings of the National Academy of Sciences*, 103:18691–18696, 2006.

[54] V. Venturi, D. A. Price, D. C. Douek und M. P. Davenport. The molecular basis for public T-cell responses? *Nature Reviews Immunology*, 8:231–238, 2008.

[55] I. H. Witten und E. Frank. *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann, San Francisco, 2000.

[56] D. Zehn, S. Y. Lee und M. J. Bevan. Complete but curtailed T cell response to very low affinity antigen. *Nature*, 458:211–214, 2009.

Abbildungsverzeichnis

Tabellenverzeichnis

Anhänge

Die im Rahmen dieser Dissertation erstellten Computerprogramme wurden an das Institut für Anatomie der Universität zu Lübeck übergeben. Die Quelltexte können bei berechtigtem Interesse dort eingesehen werden.

Danksagung

Zunächst möchte ich meinem Doktorvater Prof. Dr. Jürgen Westermann meinen herzlichen Dank aussprechen. Das von ihm vorgeschlagene Promotionsprojekt hat es mir ermöglicht, meine Vorkenntnisse aus dem Bereich der Datenanalyse in ein innovatives Forschungsprojekt einfließen zu lassen. Unsere zahlreichen, mitunter sehr intensiv geführten, Diskussionen haben einen nicht zu unterschätzenden Beitrag zu dieser Promotion geleistet.

Frau Prof. Dr. Silke Szymczak danke ich für die kritische Durchsicht meines Manuskriptes. Insbesondere die Abschnitte über die statistischen Klassifikationsverfahren haben sehr von ihren konstruktiven Anmerkungen profitiert.

Des Weiteren möchte ich allen Mitarbeitern des Institutes für Anatomie für die angenehme und konstruktive Arbeitsatmosphäre danken, in der ich dieses Promotionsprojekt durchführen konnte. Insbesondere möchte ich hierbei die Unterstützung von Dr. Cornelia Tune, Dr. Kathrin Kalies, Dr. Andrea Schampel, Lisa-Kristin Schierloh und Dr. René Pagel hervorheben. Diese haben die Experimente, welche den hier analysierten Datensätzen zugrunde liegen, in Zusammenarbeit mit Herrn Westermann, geplant und durchgeführt. Das hohe Maß an Professionalität, welches hierbei an den Tag gelegt wurde, hat dieses Promotionsprojekt letztlich erst möglich gemacht.

Abschließend möchte ich meiner Familie danken, welche mich in allen Lebenslagen unterstützt hat. Ganz besonders möchte ich hierbei meinen Vater Hans Meinhardt erwähnen, der die Vollendung dieser Arbeit leider nicht erlebt hat.

Lebenslauf

Daten zur Person:

Name:	Martin Meinhardt
Geburtsdatum:	17. Januar 1986
Geburtsort:	Tübingen
Nationalität:	deutsch

Schulbildung und Studium:

2003-2006:	Wirtschaftsgymnasium in Tübingen
Abschluss:	Allgemeine Hochschulreife
2006-2012:	Studium der Mathematik mit Nebenfach Informatik an der Universität Ulm
März 2012:	Diplom in Mathematik
2012-2019:	Studium der Humanmedizin an der Universität zu Lübeck
Mai 2019:	Vollendung der ärztlichen Prüfung

Promotion:

August 2016:	Beginn der wissenschaftlichen Tätigkeit in der Arbeitsgruppe von Herrn Prof. Dr. Jürgen Westermann am Institut für Anatomie der Universität zu Lübeck.
2016-2018:	Mitarbeit an einem Projekt, bei welchem der Einfluss von Immunreaktionen auf das Repertoire einzelner T-Zellzonen untersucht wurde (Ergebnisse siehe [47]).
2017-2021:	Mitarbeit an einem Projekt, bei welchem die Auswirkung von Schlafentzug auf die T-Zellreaktion untersucht wurde (Ergebnisse siehe [50, 51]).
2020-2022:	Verfassen der Dissertation.

Klinische Tätigkeit:

Seit 2021:	Arzt in Weiterbildung an der Klinik für Kinder- und Jugendmedizin des Universitätsklinikums Ulm.

Publikationen:

1. M. Meinhardt[1], S. Lück, P. Martin, T. Felka, W. Aicher, B. Rolauffs und V. Schmidt. Modeling chondrocyte patterns by elliptical cluster processes. *Journal of Structural Biology*, 177:447–458, 2012.

2. J. Textor, A. Fähnrich, M. Meinhardt, C. Tune, S. Klein, R. Pagel, P. König, K. Kalies und J. Westermann. Deep Sequencing Reveals Transient Segregation of T Cell Repertoires in Splenic T Cell Zones during an Immune Response. *The Journal of Immunology*, 201:350–358, 2018.

3. C. Tune, M. Meinhardt[1], K. Kalies, R. Pagel, L.-K. Schierloh, J. Hahn, S. E. Autenrieth, C. E. Koch, H. Oster, A. Schampel und J. Westermann. Effects of sleep on the splenic milieu in mice und the T cell receptor repertoire recruited into a T cell dependent B cell response. *Brain, Behavior, & Immunity - Health*, 5:100082, 2020.

4. C. Tune, J. Hahn, S. E. Autenrieth, M. Meinhardt, R. Pagel, A. Schampel, L.-K. Schierloh, K. Kalies und J. Westermann. Sleep restriction prior to antigen exposure does not alter the T cell receptor repertoire but impairs germinal center formation during a T cell-dependent B cell response in murine spleen. *Brain, Behavior, & Immunity - Health*, 16:100312, 2021.

5. M. Meinhardt[1], C. Tune, L.-K. Schierloh, A. Schampel, R. Pagel und J. Westermann. The splenic T cell receptor repertoire during an immune response against a complex antigen: Expanding private clones accumulate in the high and low copy number region. *PLOS ONE*, 17:e0273264, 2022.

[1]Geteilte Erstautorenschaft