Diss. ETH No. 24390

DISCOVERING FUNDAMENTAL PRINCIPLES OF ANTIBODY REPERTOIRES BY LARGE-SCALE SYSTEMS AND NETWORK ANALYSIS

A thesis submitted to attain the degree of

DOCTOR OF SCIENCES of ETH ZURICH

(Dr. sc. ETH Zurich)

presented by

ENKELEJDA MIHO

M.Sc. Pharmaceutical Chemistry and Technology, University of Bologna

DAS Pharmaceutical Medicine, ECPM, University of Basel

born on 19.02.1984

citizen of Italy, Shqipëri

accepted on the recommendation of

Prof. Dr. Sai T. Reddy
Prof. Dr. Tanja Stadler
Dr. Elisabetta Traggiai

2017

“E come quei che con lena affannata
uscito fuor del pelago a la riva
si volge a l’acqua perigliosa e guata,

così l’animo mio, ch’ancor fuggiva,
si volse a retro a rimirar lo passo
che non lasciò già mai persona viva.”

Dante Alighieri, *La Divina Commedia*, Inferno, I, 22–27

Acknowledgments

I am deeply grateful to my professor, *Sai.* He has been a solid guide and a joyful hope, supporting the ventures of this work while holding it together. I will always remember his generosity and bring forward with me his kind nature.

To my friend *Vici* goes most of the merit for the completion of this thesis. It was because of his firm guidance and result-driven analytical expertise, that we as a team reached the results delineated here, and not only. I am very thankful to him for the complete professional dedication and friendship. Our discussion-driven walking-lunches and late evenings are an integral part of these pages.

To *Stefan* goes all of my gratitude for making this doctorate possible. It was because of his love and patience that I am able to present this work.

Thanks to my dear family who has walked with me, always. To their unconditional love and sacrifice, I dedicate this work. To my *mother* I am grateful for the strength and the joy she reserved to me. To my *father* I am thankful for the human values that have survived the tests of academic life. To my brother *Genti* I am grateful for his bright and dynamic mentoring, we have grown into each-other.

A number of colleagues and friends have contributed to the work presented here at some point in time, with passion: Derek Mason, Ulrike Menzel, Ros Roškar, Alexandra Trkola, Balazs Laurenczy, Juan Fuentes, Manuel Kohler, Cédric Weber, Skylar Cook, Pooja Bhat, Alex Yermanos, Simon Friedensohn, Laura Prochazka, Wenjing Kang, Tarik Khan, Dominik G. Grimm, William Kelton, Venelin Mitov, Cristina Parola.

Special thanks go to the ones that have supported significantly along the way towards the doctorate degree: Marcel Tigges, Rocco Falchetto, Itschak Lamensdorf, Brigitta Elhadj-Keller, Marcel Tanner, Renato Paro, Elisabetta Traggiai, Tanja Stadler, Andrew Bradbury, Annette Mollet, Peter E. Burckhardt, Jennifer Cochran, Thomas Kepler, David Epstein, Jake Glanville, Charalampos Tsourakakis, Phil Hodgkin.

At last, thanks *God!* whose grace, and infinite and unpredictable ways allowed me to learn, and live these beautiful years.

Ledi (Enkelejda Giovanna Miho)
Basel, May 10th 2017

Abstract

Systems immunology is the science of multi-scale quantitative analysis of the immune system, which is accomplished by combining experimental high-throughput measurements with high-dimensional computational analysis. Within the context of humoral immunity, B lymphocytes possess a vast and molecularly diverse repertoire of immunoglobulins, represented by surface-bound (B-cell receptor, BCR) and secreted receptors (antibodies), which protect against pathogens.

Recently, rapid advances in high-throughput sequencing have led to a deluge of immunoglobulin sequence data that require advanced analysis approaches in bioinformatics and computational biology. Therefore, the focus of this work has been to develop and use advanced systems analysis in order to determine the architecture, diversity, and evolution of antibody repertoires from large-scale sequence data at an unprecedented level of resolution.

Specifically, network analysis is an important and useful tool for the study of complex systems like antibody repertoires. The architecture of the antibody repertoire is defined by the network similarity landscape of its CDR3 sequences, which encode for antibody identity (clonality) and antigen specificity; this network architecture reflects the breadth of antigen-binding and correlates to humoral immune protection and function. However, most networks that have been constructed thus far embody thousands of antibodies, largely undersampling the millions present in an in vivo system. One of the major challenges that exists in antibody repertoire analysis is that computational complexity increases with the scale of large sequence data, often rendering unfeasible repertoire analysis on a single machine. In order to construct large-scale networks from high-throughput sequencing data (>100,000 unique antibodies) and uncover the unknown architecture of antibody repertoires, a novel high-performance computing platform was established.

Large-scale network analysis revealed three fundamental principles of antibody repertoire architecture across murine B-cell development: reproducibility, robustness and redundancy. Reproducibility of network structure explains clonal expansion and selection. Robustness ensures a functional immune response even under extensive loss of clones (50%). Redundancy in mutational pathways suggests that there is a pre-programmed evolvability in antibody repertoires. Thus, guidelines for a quantitative network analysis of antibody repertoires were delineated, setting the stage for the field of network systems immunology, and may guide the construction of synthetic repertoires for biomedical applications.

Network analysis is crucial to the systems-level understanding and identification of biological functions on other biomolecular systems as well. Specifically, the structure and function of proteins is directly related to their sequence. Constructing or analyzing large-scale networks remains a challenge because of the high-dimensionality space. In order to enable sequence-based cross-research areas to leverage the network analysis potential, imNet was implemented as an open-source standard software platform for the comprehensive generation and rapid analysis of large-scale sequence-similarity networks. Specifically, a parallelized, map-reduce distributed algorithm generates large-scale networks from input datasets in text format in the range of millions of sequences, not attainable previously due to memory constraints and computation time. Although imNet is primarily designed to handle the scale and depth of the enormous diversity present in immune repertoires, any sequence network can be constructed. The largest existing biological networks were thus constructed from different protein database sources (e.g., entire human proteome comprising n=293,700 peptide sequences; the entire UniProt database with n=553,231 sequences; n=14,937 sequences from the immune epitope database) and high-throughput sequences of antibody (n=6,348,502) and T cell repertoires from human (n=256,054) and mice (n=36,889) samples. In addition, imNet software supports the simulation of both *in silico* sequences and networks.

While network analysis provides insights regarding the architecture of the sequence landscape, systems analysis of immune repertoires can be leveraged to detect sequence signatures within repertoires. Sequence signatures from HIV-1 infected individuals that had developed broadly neutralizing antibodies, bNAbs, (n=9) compared to HIV-1 bNAb-negative (n=13) and uninfected individuals (n=7) were detected by systems analysis of antibody repertoire datasets. A reference database was constructed from known broadly neutralizing antibody sequences (bNAbs-DB) against HIV-1. The

bNAbs-DB enabled the identification of sequences with bNAb-related characteristics like somatic hypermutations, CDR3 length, germline frequencies, clonal frequency distributions, CDR3 similarity relations within repertoires and ultimately, sequence identity to database bNAb sequences in order to uncover unidimensional repertoire characteristics associated with bNAb status. Additionally, the bNAbs-DB was used to train a support vector machine for the detection of bNAb-like sequence features in repertoires of HIV-1 infected individuals who had developed bNAbs with a high prediction accuracy (80%). This technique allowed to capture bNAb signatures that were undetectable at the repertoire level when considering low dimensional data analysis like the average SHM or CDR3 length.

Although numerous immune repertoire sequencing datasets have been generated for immune systems analysis, data retrieval is arduous and publicly available data has not been aggregated into one database, thus hindering data cross-analysis and benchmarking. Therefore, systimsDB was constructed as a dedicated SQL database of in-house and publicly available BCR/TCR datasets, which resulted in around 7 billion sequences. systimsDB is searchable across datasets through a web-interface. Users can analyze entire annotated immune repertoires or partial sets of sequences. Selected sub-repertoires are downloadable in tabular output, and results can be graphically visualized and exported. The standardized preprocessing framework and downstream analysis in systemsDB ensures consistency in methods and reproducible results, thus enabling further usability of the database for benchmarking purposes.

In summary, this work shows the development of novel bioinformatics and statistical methods and captures the high-dimensional complexity of immune repertoires enabling deep immunological insight into the adaptive immune response.

Riassunto

L'immunologia dei sistemi si fonda sull'analisi multi-scala e quantitativa del sistema immunitario, che si esegue combinando misurazioni sperimentali ad elevato parallelismo con le relative analisi computazionali multi-dimensionali. Nel contesto dell'immunità umorale, i linfociti B possiedono un ampio repertorio di immunoglobuline, diverse a livello molecolare, che esistono sotto forma di recettori transmembrana (recettore delle cellule B, BCR) e secreti (anticorpi), e che proteggono dai patogeni.

RRecentemente, i rapidi sviluppi nel sequenziamento ad elevato parallelismo hanno portato ad una raccolta estesa di dati di sequenze di immunoglobuline, che necessitano approcci d'analisi avanzata in bioinformatica e biologia computazionale. Dunque, questo lavoro si focalizza sullo sviluppo e utilizzo di analisi avanzate di sistemi allo scopo di determinare l'architettura, la diversità e l'evoluzione dei repertori di anticorpi; simili strumenti permettono di processare le informazioni contenute nei grandi dati di sequenze, ottenendo un livello di risoluzione senza precedenti.

Nello specifico, l'analisi di reti è un mezzo importante e utile nello studio di sistemi complessi come il repertorio di anticorpi. L'architettura di un repertorio di anticorpi è definita dallo spazio di similarità di rete delle sue sequenze di CDR3, che codificano l'identità dell'anticorpo (clonalità) e la specificità verso l'antigene; quest'architettura di rete riflette l'estensione della specificità del repertorio verso l'antigene ed è correlata al ruolo e alla protezione svolta dall'immunità umorale. Tuttavia, la maggior parte delle reti costruite finora incorpora migliaia di anticorpi, ampiamente sottostimando i millioni presenti in un sistema *in vivo*. Una delle più grandi sfide esistenti nell'analisi di repertori di anticorpi è dovuta al fatto che la complessità computazionale aumenta con i gradi dati, rendendo non praticabile l'analisi su un singolo dispositivo. Per poter costruire reti di larga scala da dati di sequenziamento ad elevato parallelismo (>100,000 sequenze uniche di anticorpi) e scoprire l'architettura sconosciuta dei repertori di anticorpi, è stata stabilita una nuova piattaforma computazionale ad alte prestazioni.

L'analisi di reti di larga scala ha rivelato tre principi fondamentali dell'architettura dei repertori di anticorpi attraverso lo sviluppo delle cellule B nei topi: la riproducibilità, la robustezza e la ridondanza. La riproducibilità della struttura di rete spiega l'espansione e la selezione clonale. La robustezza assicura il funzionamento della risposta immunitaria anche in seguito ad una perdita significativa di cloni (50%). La ridondanza nei meccanismi mutazionali suggerisce che esiste un'evolvabilità programmata in anticipo nei repertori di anticorpi. Dunque, si sono quindi stabilite le line guida per l'analisi quantitativa di reti di anticorpi, ponendo le basi per il campo dell'immunologia dei sistemi basata sulle reti, che potrebbe guidare la costruzione di repertori sintetici per applicazioni biomediche.

L'analisi di reti è fondamentale per la comprensione a livello di sistema e l' identificazione di funzioni biologiche anche per altri sistemi biomolecolari. Nello specifico, la struttura e la funzione delle proteine sono direttamente collegate alla loro sequenza, ma la costruzione o l'analisi di grandi reti rimane una sfida a causa dello spazio alto-dimensionale. Allo scopo di consentire alle diverse aree di ricerca centrate sul sequenziamento di trarre vantaggio dall'analisi di reti, imNet è stato sviluppato come una piattaforma software standard per la completa costruzione e analisi rapida di reti di larga scala basate sulla similarità delle sequenze. In particolare, un algoritmo parallelizzato che usa il map-reduce per la computazione distribuita genera grandi reti da dati di testo nella gamma di milioni di sequenze, non ottenibili precedentemente per via dei limiti di memoria e tempo di calcolo. Nonostante imNet sia principalmente progettato per gestire la mole e la profondità dell'enorme diversità presente nei repertori immunitari, può essere utilizzato per la costruzione di qualunque rete di sequenze. Le più grandi reti biologiche esistenti sono state infatti costruite da diverse risorse di banche dati di proteine (ad es. l'intero proteoma umano con n=293,700 sequenze peptidiche; l'intera banca dati UniProt con n=553,231 sequenze; n=14,937 sequenze dalla banca dati degli epitopi) e sequenze ad elevato volume di anticorpi (n=6,348,502) e repertori di cellule T da campioni umani (n=256,054) e murini (n=36,889). Inoltre, il software di imNet supporta la simulazione di sequenze e reti *in silico*.

Mentre l'analisi di reti permette di comprendere l'architettura spaziale delle sequenze, l'analisi dei sistemi può essere utilizzata per rilevare caratteristiche distintive nei repertori. I segni distintivi delle sequenze da individui infetti da HIV-1 che hanno sviluppato anticorpi ampiamente neutralizzanti, bNAbs, (n=9) piuttosto che individui infetti da HIV-1 bNAb-negativi (n=13) e individui non infetti (n=7) sono stati identifi-

cati tramite analisi dei sistemi di repertori di anticorpi. Una banca dati di riferimento è stata generata da sequenze precedentemente note di anticorpi ampiamente neutralizzanti (bNAbs-DB) contro HIV-1. La bNAbs-DB ha permesso di identificare sequenze con caratteristiche collegate ai bNAb come ipermutazione somatica, lunghezza dei CDR3, la frequenza delle linee germinali, le distribuzioni delle frequenze clonali e, infine, l'identità alla banca dati di sequenze di bNAbs per poter individuare caratteristiche unidimensionali del repertorio associato con lo stato di sviluppo dei bNAbs. Inoltre, la bNAbs-DB è stata utilizzata, con un'alta accuratezza nella previsione (80%), per addestrare una macchina a vettori di supporto per identificare caratteristiche tipiche degli bNAbs nelle sequenze dei repertori di individui infetti da HIV-1 che hanno sviluppato bNAb. Questo metodo ha reso possibile individuare segni distintivi degli bNAb che non erano distinguibili a livello di analisi unidimensionali del repertorio come la valutazione dell'ipermutazione somatica e la lunghezza media dei CDR3.

Nonostante serie di dati da numerosi repertori sequenziati siano state generate per analizzare il sistema immunitario, il recupero di tali dati è faticoso e i dati pubblicamente disponibili non sono raccolti in un'unica banca dati, impedendo così l'analisi trasversale dei dati e il benchmark. Pertanto, systimsDB è stata costruita come una banca dati SQL dedicata ai dati prodotti localmente e pubblicamente disponibili di BCR e TCR, per un totale di circa 7 bilioni di sequenze. systimsDB è consultabile attraverso un'interfaccia web. Gli utenti possono utilizzare repertori completi annotati oppure serie parziali di sequenze. I sotto-repertori selezionati sono scaricabili tramite output tabulare e i risultati possono essere visualizzati in grafici ed esportati. La piattaforma standard di pre-elaborazione dati assicura consistenza nei metodi e risultati riproducibili, così permettendo il successivo utilizzo della banca dati come valore di riferimento.

Riassumendo, questo lavoro tratta lo sviluppo di nuovi metodi bioinformatici e statistici e rivela la complessità multi-dimensionale dei repertori immunitari, permettendo di arrivare a una comprensione profonda della risposta immunitaria umorale.

Table of contents

List of figures

List of tables

Chapter 1

Introduction

Immunology is the science of defense and integrity. The immune system protects an organism from pathogens (bacteria, virus and parasites) and maintains homeostasis by eliminating toxic or allergenic substances [3]. Specifically, the adaptive arm of the immune response is complex, characterized by large clonal diversity (each clone expresses a unique receptor) and antigen-specificity [4]. Systems immunology aims to investigate the immune response comprehensively by integrating high-throughput technologies with bioinformatics and computational biology [5–8]. Here we apply systems and network analysis to answer fundamental questions in immunology and develop bioinformatic tools to characterize the immune repertoire.

1.1 Adaptive immunity

The main components of the adaptive immune system in mammals are B- and T-lymphocytes (short: B and T cells) [9]. These immune cells arose approximately 500 million years ago and their main characteristic is their large diversity: each B and T cell expresses on the surface antigen receptors, each identifying uniquely a clone (cell lineage) and generated by somatic recombination. Germline gene rearrangements [10] produce combinatorial diversity to generate a vast repertoire of naïve B/T cells, which have not yet encountered their specific antigen, but represent a vast landscape of potential binding specificities. Thus, collections of B and T cells constitute diverse and complex immune repertoires, made of billions of unique clones identified through their complementarity determining region 3 (CDR3 of B cell receptors and antibodies, Fig. 1.1, 1.2). CDR3 is the most diverse region of the B/T cell receptors (or antibodies secreted from differentiated B cells) is the main contributor to antigen-recognition, thus determining binding [11, 12]. The generating mechanisms of immune repertoires contribute to their vast sequence space complexity.

1.1.1 Humoral immune response

B cells constitute the humoral arm of the adaptive immune response in vertebrates (Fig. 1.1) and confer protection against a plethora of pathogens. Pre-B cells are generated in the bone marrow, then migrate to the lymph nodes and the spleen as naïve B cells that are here activated by the antigen, thus proliferate and terminally differentiate into plasma cells homed in the bone marrow, which produce antibodies, the secreted form of the B cell receptor. The collection of B cell receptors and antibodies constitutes the antibody repertoire.

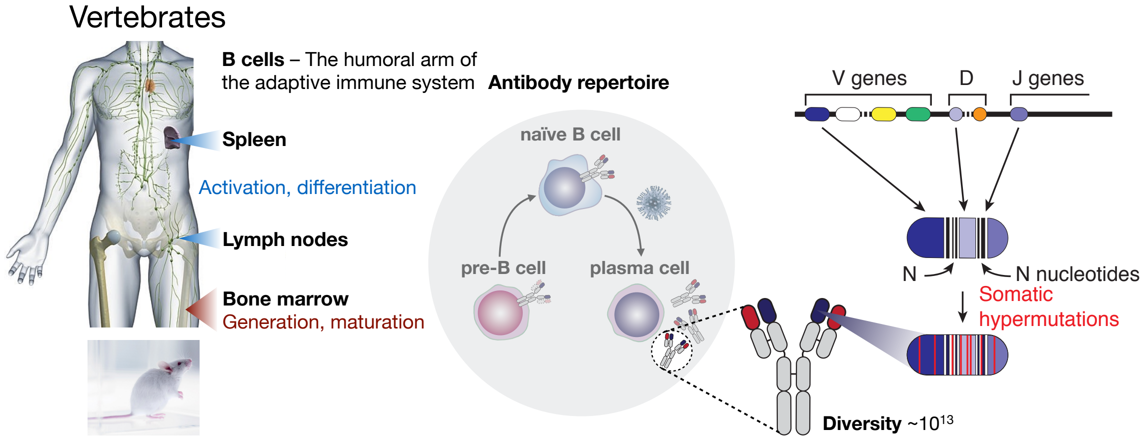


Fig. 1.1 Humoral immunity and the generating mechanisms of its diversity. B cell receptor repertoire evolution in vertebrates starts from the generation of pre-B cells in the bone marrow that develop into naïve B cells and differentiate into plasma cells that secrete antibodies. Each B cell receptor and antibody is consituted from heavy chain with a constant (grey) and a variable region (blue) and a light chain with a variable (red) and a constant region (grey). The variable region of the heavy chain is generated from the somatic recombination of V, D and J germline genes, the subsequent addition and subtraction of nucleotides and somatic hypermutations.

Antibodies have a heavy and a light chain, each with a variable region that differs extensively from one antibody to the next. The source of antibody diversity has been identified to be the VDJ somatic recombination [11]. The variable regions of an antibody are encoded by randomly selected V, D, J genes from their multiple copies. Further diversity is introduced by the addition and subtraction of nucleotides at the joints between the recombined V, D, J genes and somatic hypermutations which are point mutations into the rearranged V-regions of activated B cells [13]. The diversity the of antibody repertoires is daunting, and is estimated to be $\approx 10^{13}$ [4].

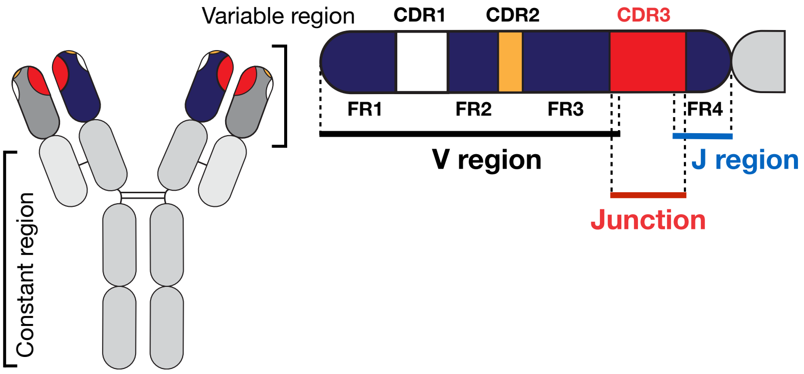


Fig. 1.2 B cell receptors and antibodies. B cell receptors and antibodies are made of heavy and light chains, each with a variable (dark colored) and constant region (light colored). The variable region of the heavy chain (dark blue) results from VDJ gene somatic recombination, addition and subtraction of nucleotides at the joints between the gene recombination sites, and somatic hypermutations. After alignment of to the reference genome V, D, and J genes, the V region, the junction and the J region are identified in the variable region of the antibody heavy chain. Framework (FR) and complementarity determining regions (CDR) of the antibody sequence are annotated according a numbering scheme (e.g., IMGT unique numbering [14]).

1.2 Sequences of immune repertoires

High-throughput sequencing of antibody repertoires has enabled a systems understanding and quantitative insight into the adaptive immune response [15]. Deep sequencing data captures the diversity of antibody repertoires, thus enabling a systems-level analysis of entire repertoires. Therefore, high-throughput sequencing data was analyzed in order to gain a systems understanding of the structure of antibody repertoires and detect global signatures at the repertoire level.

1.2.1 Sequenced murine antibody repertoires [1]

The experimental design of the analyzed mouse antibody repertoires assured high technological and biological coverage of the murine antibody repertoires along development [1]. The use of an inbred mouse model of 8–10 weeks old female C57BL/6 J mice (n = 19) housed under specific pathogen-free conditions, guaranteed the analysis of a fully sequenced genome and a completely annotated immunoglobulin genomic locus [16]. C57BL/6 mice were stratified into cohorts consisting of untreated (n = 5) or intraperitoneally prime-boost (after 3 weeks, with identical amounts of antigen in PBS) immunized mice, using three structurally different alum-precipitated antigens: 100 mg ovalbumin (OVA, n = 5), 100 mg 4-hydroxy-3-nitro-phenylacetyl conjugated to hen egg lysozyme (NP-HEL, n = 5), and 4 mg hepatitis B virus surface antigen (HBsAg, n = 4). Immunized mice were sacrificed 14 days post-secondary immunization; untreated control mice were sacrificed at corresponding age (Fig. 1.3).

Three key differentiation stages of the B cell development were isolated through fluorescence-activated cell sorting from two major lymphoid organs, spleen and bone marrow: pre-B cells (pBCs, $c-kit^{-}CD19^{+}IgM^{-}CD25^{+}PI^{-}$, bone marrow), naive follicular B cells (nBCs, $CD138^{-}CD19^{+}IgD2^{+}IgM^{+}CD232^{+}CD21^{+}PI^{-}$, spleen), and long-lived memory plasma cells (PCs, $CD138^{+}CD22^{-}MHCII^{-}CD19^{-}IgM^{-}PI^{-}$, bone marrow).

Amplification of antibody variable heavy chain (VH) libraries was performed using a forward primers specific to framework region 1, and either IgM- (pBC and nBC) or IgG- and IgM-specific (PC) reverse primers. Antibody library pools were sequenced on the Illumina MiSeq platform (2x300 cycles, paired-end). Total RNA sequencing led to 400 million full-length variable heavy chain region (VH) sequences from was performed

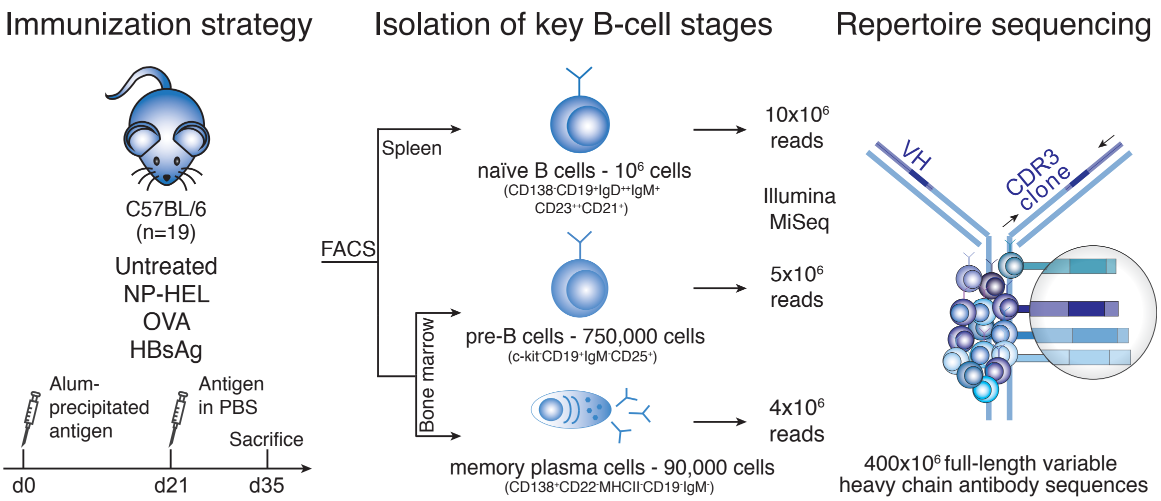

Immunization strategy
Isolation of key B-cell stages
Repertoire sequencing
C57BL/6
(n=19)
Untreated
NP-HEL
OVA
HBsAg
Alum-precipitated antigen
Antigen in PBS
Sacrifice
d0
d21
d35
FACS
Spleen
Bone marrow
naïve B cells - 10^6 cells
(CD138$^-$CD19$^+$IgD^{++}IgM$^+$ CD23^{++}CD21$^+$)
pre-B cells - 750,000 cells
(c-kit$^-$CD19$^+$IgM$^-$CD25$^+$)
memory plasma cells - 90,000 cells
(CD138$^+$CD22$^-$MHCII$^-$CD19$^-$IgM$^-$)
$10x10^6$ reads
Illumina MiSeq
$5x10^6$ reads
$4x10^6$ reads
VH
CDR3 clone
$400x10^6$ full-length variable heavy chain antibody sequences

Fig. 1.3 Sequencing the developing antibody repertoire: experimental design. Experimentally, antibody heavy chain high-throughput sequencing of pre-B cells (preBC), naive B cells (nBC), and memory plasma cells (PC) from four C57BL/6 mouse cohorts was performed.

(Fig. 1.3). Mean base call quality of all samples was in the range of Phred score 30 [1].

Sequencing data was processed (VDJ alignment, clonotyping) using the MiXCR software package (clonotype formation by CDR3 region) [17]. For downstream analyses, functional clonotypes were only retained if they were composed of at least four amino acids and had a minimal read count of 2 [18, 19]. The VH complementarity determining region 3 (CDR3) served as an accepted proxy for antibody clonality and specificity [12, 5]. Furthermore, repertoire sequencing reproducibility of the dataset was confirmed by technical replicates as described in Greiff et al [1].

1.2.2 Sequenced human antibody repertoires [2]

The data collection design leveraged samples of sequenced human PBMC from different individuals (total n=29) and constructed three cohorts (Fig. 1.4) and one reference database in order to detect bNAb-sequence signatures at the repertoire-level (Fig. 1.5). The organization of data allows to interrogate repertoires of individuals that have developed bNAbs (HIV-1 bNAb+) and compare them against two different negative controls (uninfected and HIV-1 bNAb-) and a positive control of bNAbs sequences collected in the bNAbs database (bNAbs-DB).

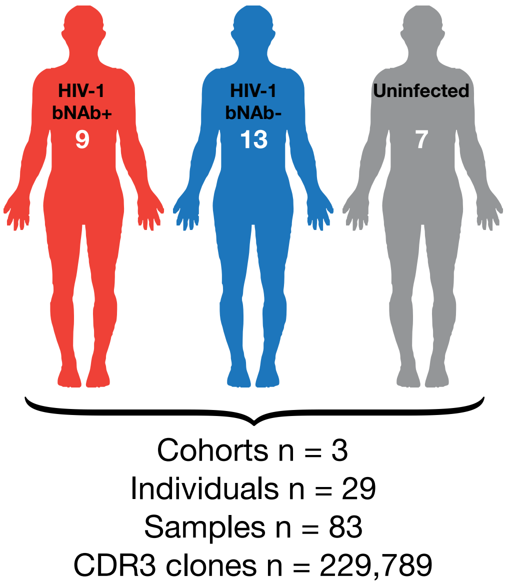


Fig. 1.4 Sequenced human antibody repertoires: data design. Data was collected from three cohorts (HIV-1 bNAb+ n=9, HIV-1 bNAb- n=13, and uninfected n=7) of sequenced human PBMC samples of 29 individuals.

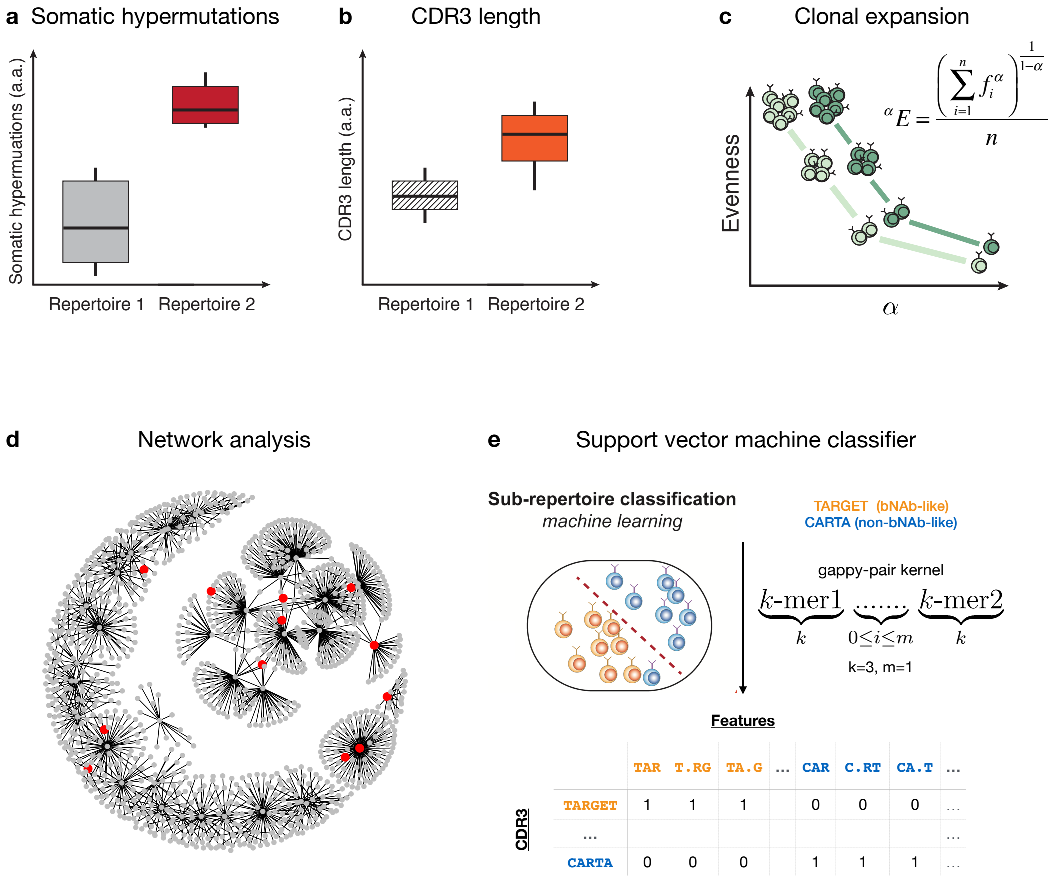
1.3 Systems quantification of immune repertoires

"...as long as the quantitation of the immune response remains elusive, immunology will remain a phenomenology" – J. K. Jerne [20]

Many discoveries since 1890 have elucidated the adaptive immune system [21]. However, a truly systems perspective and quantification of the large and complex sequence space B and T cell repertoires has only recently begun to delineate.

High-throughput sequencing of B and T cell repertoires coupled with computational methods [22, 23] enable the quantification of the adaptive immune response, thus bringing fundamental insights into health and disease status, vaccine development and guiding targeted therapeutics [24–28].

This work allows to (i) capture the architecture of the sequence space of B cell repertoires along development, revealing fundamental intrinsic principles through large-scale network analysis and (ii) characterize antibody repertoires from a systems prospective, revealing antigen-specific sequence signatures (Fig. 1.5, 2.3). At last, it proposes an immune repertoire database (systimsDB) that ensures reproducible results and allows to use the available data for benchmarking purposes and cross-validation analysis.


a Somatic hypermutations
Somatic hypermutations (a.a.)
Repertoire 1 Repertoire 2
b CDR3 length
CDR3 length (a.a.)
Repertoire 1 Repertoire 2
c Clonal expansion
Evenness
α
$^{\alpha}E = \frac{\left(\sum_{i=1}^{n} f_i^{\alpha}\right)^{\frac{1}{1-\alpha}}}{n}$
d Network analysis
e Support vector machine classifier
Sub-repertoire classification
machine learning
TARGET (bNAb-like)
CARTA (non-bNAb-like)
gappy-pair kernel
$\underbrace{k\text{-mer1}}_{k} \underbrace{\ldots\ldots}_{0\leq i\leq m} \underbrace{k\text{-mer2}}_{k}$
k=3, m=1
Features
CDR3

	TAR	T.RG	TA.G	...	CAR	C.RT	CA.T	...
TARGET	1	1	1		0	0	0	...
...								...
CARTA	0	0	0		1	1	1	...

Fig. 1.5 Systems analysis to detect signatures in antibody repertoires. Immune repertoires can be characterized and compared at a systems level with regard to their a) somatic hypermutations and b) CDR3 lengths, c) clonal expansion through evenness analysis, d) clonal architecture through network analysis, and e) classified for antigen-specific like-sequences using machine learning.

1.3.1 State-of-the-art of bioinformatic and statistical analyses for sequenced immune repertoires

Chapter 2 reviews the state-of-art of bioinformatic and statistical analyses of sequenced immune repertoires, illustrating the challenges and open questions in the field of systems immunology.

1.3.2 The architecture of the sequence space of immune repertoires

The fundamental principles of antibody repertoire architecture are revealed in **Chapter 3** by establishing novel large-scale network analysis as described in **Chapter 4**.

A historical paradigm shift enabled by high-throughput sequencing has led to a major transition from analyzing individual antibodies to capturing the entire diversity of antibody repertoires. The big data generated from sequenced repertoires, however, introduced significant unresolved computational challenges. The large numbers of sequences generated from very diverse antibody repertoires (e.g., $\geq 10^5$ unique antibodies) limited the practice of networks in immunology, although Nobel laureate Niels K. Jerne first proposed to use network theory for the study of the immune system more than 40 years ago [29]. Consequently, to obtain the architecture of antibody repertoires requires a second paradigm shift – the unification of immunology with informatics [30]. Thus, the architecture of antibody repertoires has remained unknown up-to-date.

The architecture of an antibody repertoire is defined by the network similarity landscape of its sequences and reflects the spectrum of antigen binding – thereby determining immunological protection and function. Here, a novel high-performance computing software to construct for the first time large-scale networks from high-throughput sequences of entire antibody repertoires was developed: imNet (**Chapter 4**).

The fundamental principles of reproducibility, robustness and redundancy of antibody repertoire architecture were uncovered. Reproducibility of network structure explains clonal expansion and selection signatures present in humoral immunity. Robustness of the architecture ensures a functional immune response even at extensive deletion of antibodies. Redundancy in mutational pathways provides pre-programmed evolvability of antibody responses.

These architectural principles serve as the blueprint for the construction of antibody repertoires, such as synthetic repertoires simulating natural immune systems, which can be used for immunotherapeutic and biomedical applications. By setting the stage for the field of *network systems immunology*, the attempt with this work is to help bring to fruition the vision laid out by Jerne.

1.3.3 Detecting sequence signatures with systems analysis

Systems analysis of antibody repertoires was used to detect broadly neutralizing antibody (bNAb) signatures in HIV-1 infected individuals that developed bNAbs (HIV-1 bNAb+) versus HIV-1 infected individuals that were negative for bNAbs (HIV-1 bNAb-) and uninfected individuals as described in **Chapter 5**. To this end, also a database of known bNAbs sequences was constructed (bNAbs-DB).

The systems characterization of immune repertoires allows for the detection of sequences with similar characteristics (CDR3 length, somatic hypermutations, V/J gene), however these characteristics often are not reflected at the systems level. When machine learning is applied to the diverse antibody repertoires at the sequence level, it detects bNAb-like sequences in HIV-1 bNAb+ versus uninfected individuals by using a trained model on the bNAbs-DB vs. sequences of HIV-1 bNAb-.

This systems approach can be applied to different datasets in search for antigen-like specificity or diagnostic-specific sequences in clinics.

1.3.4 A database of immune repertoire sequences

Although the systems analysis of immune repertoire sequences is progressing with the sequencing technology, immune repertoires have not been aggregated into a database, thus hindering the reproducible cross-analysis and benchmarking of immune sequences and repertoires from public data. systimsDB, a collection of B and T cell sequences ($\approx$ 7.5 billion) as described in **Chapter 6**, allows for the search, download and analysis of publicly available sequences of immune repertoires.

1.4 Articles incorporated in this work

This work incorporates authored and co-authored scientific articles which have been published or submitted for publication.

- **Chapter 1, Section 1.2.1**

  Greiff V*, Menzel U*, **Miho E**, Weber CR, Riedel R, Cook S, Valai A, Lopes T, Radbruch A, Winkler TH, Reddy ST. "Systems analysis reveals high genetic and antigen-driven predetermination of antibody repertoires throughout B-cell development." *Cell Reports*, 2017. [1]

- **Chapter 2**

  Greiff V, **Miho E**, Menzel U, Reddy ST. "Bioinformatic and statistical analysis of adaptive immune repertoires." *Trends in Immunology*, 2015. [5]

- **Chapter 3**

  Miho E*, Greiff V*, Roškar R, Reddy ST. "The fundamental principles of antibody repertoire architecture revealed by large-scale network analysis." (in review). Preprint available on bioRxiv, published online April 5, 2017 (http://dx.doi.org/10.1101/124578). [31]

- **Chapter 4**

  Miho E*, Roškar R*, Greiff V, Reddy ST. "imNet: software for the generation and analysis of large-scale sequence networks." (in submission). [32]

- **Chapter 5**

  Miho E et al. "Detection of sequence signatures of broadly neutralizing antibodies in HIV-1 individuals." (in submission). [2]

- **Chapter 6**

  Miho E, Friedensohn S, Laurenczy B, Fuentes Serna JM, et al. "SystimsDB: a systems database of immune repertoires." (in submission). [33]

1.5 Other co-authored articles

- Greiff V*, Weber CR*, Palme J, Bodenhofer U, **Miho E**, Menzel U, Reddy ST. "Learning The High-Dimensional Immunogenomic Features That Predict

Public And Private Antibody Repertoires." *The Journal of Immunology*, 2017 (in press). Preprint available on bioRxiv, published online April 18, 2017 (http://dx.doi.org/10.1101/127902). [34]

- Yermanos A, Greiff V, Krautler N, Menzel U, Dounas A, **Miho E**, Oxenius A, Stadler T, Reddy ST. "Comparison of methods for phylogenetic B-cell lineage inference using time-resolved antibody repertoire simulations (AbSim)." *Bioinformatics*, 2017. [35]

- Schanz M, Liechti T, Zagordi O, **Miho E**, Reddy ST, Günthard HF, Trkola A, Huber M. "High-throughput sequencing of human immunoglobulin variable regions with subtype identification." *PLOS ONE*, 2014. [36]

Chapter 2

Bioinformatic and statistical analysis of adaptive immune repertoires

High-throughput sequencing (HTS) of immune repertoires has enabled the quantitative analysis of adaptive immune responses and offers the potential to revolutionize research in lymphocyte biology, vaccine profiling, and monoclonal antibody engineering. Advances in sequencing technology coupled to an exponential decline in sequencing costs have fueled the recent overwhelming interest in immune repertoire sequencing. This, in turn, has sparked the development of numerous methods for bioinformatic and statistics-driven interpretation and visualization of immune repertoires. Here, we review the current literature on bioinformatic and statistical analysis of immune repertoire HTS data and discuss underlying assumptions, applicability, and scope. We further highlight important directions for future research, which could propel immune repertoire HTS to becoming a standard method for measuring adaptive immune responses.

2.1 Trends

- High-throughput immune repertoire sequencing is becoming a core technology for advancing basic, applied, and clinical immunology.
- Specialized bioinformatic and statistical methods for repertoire diversity and overlap analysis as well as for performing network and phylogenetic clustering enable the investigation of immune repertoire expansion, dynamics, architecture and evolution at an unprecedented level of detail.
- There is a divergence of the underlying assumptions, applicability, and scope of bioinformatic and statistical methods, thus compromising the consistency of data analyses within and across studies that need to be addressed in the near future.

2.2 Resolving the complexity of antigen receptor repertoires

B and T lymphocytes of the adaptive immune system have the ability to recognize foreign molecules via an immense array of antigen-binding receptors (B cell and T cell receptors, BCR/TCR) [37]. The diversity of lymphocyte repertoires (short: ‘immune repertoires’) is a result of genetic recombination and diversification mechanisms. Diversity is first created in the germline via recombination of variable V, (diversity D), and joining J gene segments [11], which form the antigen-binding variable region. Further diversification occurs through imprecise junction of these gene segments (addition of P- and N-nucleotides adjacent to the D segment), somatic hypermutation (SHM, in B cells), and combination of subunits (T cells) or heavy/light chains (B cells) [4].

Immune repertoire antigen-specificity and diversity is largely dominated by the junctional site of V (D)J recombination, which is known as the complementarity determining region 3 (CDR3) [12, 38]. The CDR3 has thus served for a long time as a natural identifier of lymphocyte clonality: B and T cells with an identical CDR3 are classified as belonging to the same clone. Upon antigen challenge, B cells are activated and undergo clonal selection and expansion forming a clonal lineage [39].

Immune repertoires are an important target of immunological and clinical research because they harbor information on both past and ongoing immune responses [4]. HTS now enables the quantitative analysis of these highly diverse immune repertoires at an unprecedented depth [28, 27, 26] and has already shown tremendous promise for investigating immune repertoire changes during chronic viral infections (e.g., HIV-1) [40–42], autoimmune diseases [43–46], and with aging [47, 48]. Continuous advances in sequencing technology have sparked the development of numerous bioinformatic and statistical methods which aim to maximize information extraction from immune repertoire HTS data. We review the current literature on bioinformatic and statistical analysis of immune repertoires. Specifically, we discuss steps of the HTS (bioinformatic) workflow which can influence the biological conclusiveness of a study, such as representative repertoire sampling, data error correction, and sequence germline annotation, as well as statistical approaches for estimating diversity and visualization of repertoire selection, architecture, and evolution.

2.3 Sampling and sequencing depth: How deep is deep enough?

While HTS provides a tremendous amount of depth for analyzing immune repertoires, biologically meaningful information on repertoire diversity is substantially dependent on the comprehensive sampling of the cell population studied (**biological sampling**, Fig. 2.1 A, see Glossary 2.8, [24]) and on the comprehensive read coverage of DNA or RNA molecules encoding BCRs and TCRs (**technological sampling**, Fig. 2.1 A). The choices of the organism and cell population are key to achieving optimal repertoire coverage; in humans, the most common source for lymphocytes is the peripheral blood compartment, which contains only 2.5% ($\sim 10^9$ B or T cells) of the estimated total number of cells ($\approx 10^{11}$) [49–52]. In mice the coverage of immune repertoires is less problematic because all lymphoid organs are readily accessible, and the number of lymphocytes ($\sim 10^8$ B or T cells [53]) lies significantly closer to the current state of the art in sequencing depth (10^4 and 10^7 sequencing reads per sample [46, 54]).

The consequences of insufficient biological sampling have been investigated previously by Warren and colleagues [55]: they showed that distinct 20 mL blood samples from the same individual captured only a portion of the TCR peripheral blood repertoire (biological undersampling). Furthermore, technological undersampling has been shown to compromise the detection of ‘public’ clones (clones shared across individuals), which are a common target in immune repertoire studies [56, 57]. In fact, several studies indicated that there was a positive correlation between sequencing depth and the number of public clones detected [43, 58, 59]. Thus, the biological conclusiveness of a study benefits from the implementation of **biological replicates** (test for biological undersampling [55, 60, 61]) (Fig. 2.1 A) and **technical replicates** (test for technological undersampling [18, 62, 15, 19]) (Fig. 2.1 A), which may be performed once for each cell population analyzed. It is important to note that biological undersampling can only be meaningfully addressed if sufficient technological sampling has been established [18]. Furthermore, species accumulation and rarefaction analyses may be performed to quantify the extent of (under)sampling [58, 18, 15, 63] (Fig. 2.1 A).

Because the size of the cellular compartments studied differs widely by research question, universal guidelines for ensuring comprehensive sampling are challenging to implement. Nevertheless, two general rules to consider are: (i) the number of sequencing reads should at least exceed the clonal diversity of the sample if complete read coverage is unattainable, and (ii) the lower the frequency of a clone, the higher the sequencing depth must be for its accurate capture [62]. While knowing the exact clonal diversity of a lymphocyte population before HTS is not possible, basic knowledge of cell numbers and clonal frequency distributions, as well as mathematical modeling [62], facilitate the estimation of the required sequencing depth. For example, antigen-specific or clonally expanded populations (e.g., memory B and T cells, plasma cells) [62] will have a clone-to-cell ratio that is well below 1 [44, 18, 64–67], and thus less sequencing reads would be required to obtain a good snapshot of the clonal diversity. By contrast, clonal frequency distributions of naive B and T cells have been shown to be more uniform [44, 64–68] (i.e., higher clone-to-cell ratios than clonally expanded populations), requiring a substantially higher number of reads for clonal diversity description. In the future, as sequencing depth continues to increase, repertoire coverage may become more comprehensive.

2.4 Bioinformatics tools for preprocessing of immune repertoire data

2.4.1 Combining experimental and computational approaches for error correction of immune repertoire HTS data

Regardless of the sequencing platform, HTS has not yet reached the level of accuracy of Sanger sequencing because it suffers from errors introduced during library ampli-

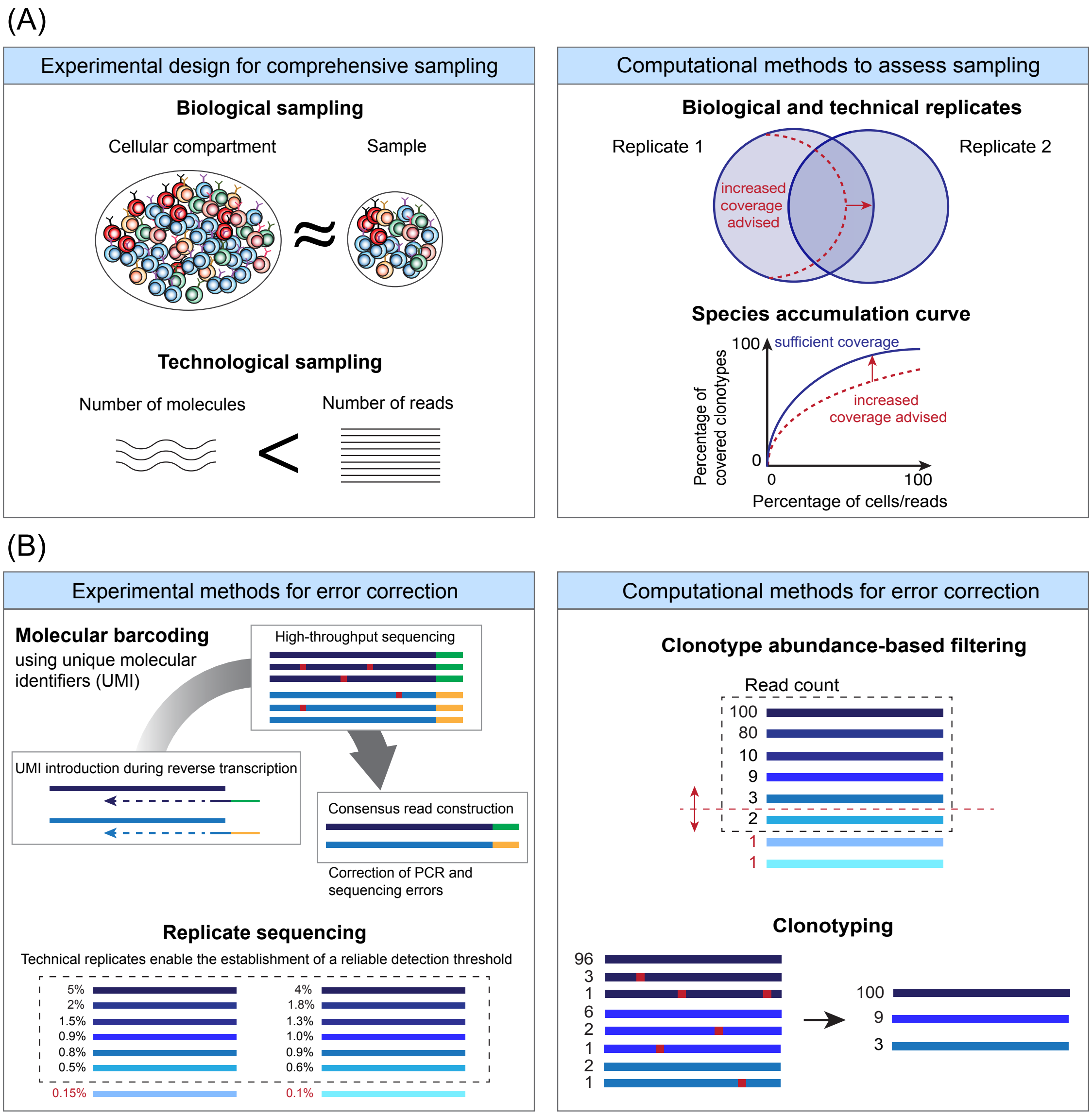

(A)
Experimental design for comprehensive sampling
Biological sampling
Cellular compartment
Sample
≈
Technological sampling
Number of molecules
Number of reads
<
Computational methods to assess sampling
Biological and technical replicates
Replicate 1
Replicate 2
increased coverage advised
Species accumulation curve
100
sufficient coverage
increased coverage advised
Percentage of covered clonotypes
0
0
100
Percentage of cells/reads
(B)
Experimental methods for error correction
Molecular barcoding
using unique molecular identifiers (UMI)
High-throughput sequencing
UMI introduction during reverse transcription
Consensus read construction
Correction of PCR and sequencing errors
Replicate sequencing
Technical replicates enable the establishment of a reliable detection threshold
5%
2%
1.5%
0.9%
0.8%
0.5%
0.15%
4%
1.8%
1.3%
1.0%
0.9%
0.6%
0.1%
Computational methods for error correction
Clonotype abundance-based filtering
Read count
100
80
10
9
3
2
1
1
Clonotyping
96
3
1
6
2
1
2
1
100
9
3

Fig. 2.1 Experimental and computational design. Considerations for the Experimental and Computational Design of an Immune Repertoire Study. (A) To comprehensively describe the population of interest, both biological and technological sampling deserve consideration. Reliable biological sampling ensures that the sampled population represents the diversity of the cellular compartment being investigated. For reliable HTS data it is equally important to calculate more reads than clones, or, if quantifiable, input molecules (DNA/RNA); this is referred to as technological sampling. The sufficiency of both biological and technological sampling can be assessed by clonal overlap analysis (e.g., Venn diagrams). Another typical means to assess sufficient sampling is by species accumulation/rarefaction curves; a curve that levels off indicates complete clonal coverage, and incomplete coverage is revealed by a curve with a positive slope (more cells/reads would reveal more unseen clones).

Fig. 2.1 (B) There exist different methods to account and correct for errors introduced by PCR and sequencing. Experimental methods comprise the addition of unique molecular identifiers (UMIs), which allow the construction of consensus reads. Replicate sequencing can be used to determine reliable detection cutoff. Errors can also be corrected computationally by heuristic abundance-based filtering of clones present with only a few reads or by clustering similar sequences based on a defined similarity threshold (clonotyping).

fication (experimental) or sequencing (HTS, bridge amplification, platform-specific) [69]. Therefore, both experimental and computational strategies have been devised to attenuate the impact of errors on biological conclusions. A shared principle of all error-correction approaches listed below is that they rely (either explicitly or implicitly) on high read numbers. Thus, high sequencing depth not only ensures comprehensive sampling but can also increase the accuracy of repertoire measurements.

It is a well-known statistical principle that a given entity converges to its true (‘expected’) value (law of large numbers) if sampled sufficiently often. This principle is leveraged by an error- correction approach that is based on unique molecular identifiers [UMIs, also referred to as unique identifiers (UIDs) or barcodes] (Figure 1B), which are degenerate nucleotide sequences of high diversity that uniquely tag each DNA or RNA molecule [60, 70–72]. Leveraging dedicated bioinformatic pipelines [71, 73], UMI methods in immune repertoire sequencing have been shown to achieve up to a 100-fold error reduction [60, 71], thus considerably reducing artificial repertoire diversity [74]. However, a study by Shugay and colleagues indicated that increased RNA input (increasing from ng to mg) required a considerable increase in sequencing depth (10^6 to 10^7 sequencing reads) and a switch in sequencing platform (Illumina MiSeq to HiSeq) to ensure consensus read construction (presence of multiple sequencing reads with identical UMIs) [71]. Therefore, to effectively use UMI approaches for error correction, technological over-sampling is needed, which may not always be feasible for highly-diverse and large cell populations (i.e., naive B and T cells). In addition to error correction, UMI methods have also been applied to the problem of pairing TCR / and TCR b chains [75], and BCR heavy and light chains [76].

Another approach to experimental error correction is the use of technical replicates which can be used to establish **reliable clonal detection cutoffs** [61, 18, 19] (Fig. 2.1 B). While these cutoffs exploit the multiplicity of reads per clone as detection confidence [55], it has been indicated that hotspot PCR or sequencing errors are reproducible across technical replicates [71]. In these situations, the only approach for error correction would be UMI-based [71]. However, it should be noted that UMI-based approaches can still benefit by using technical replicates to establish sensitivity thresholds of error correction [77, 78].

There is a vast array of approaches that could be considered as computational error correction. The simplest would be filtering HTS datasets (before any V(D)J annotation, Fig. 2.3) for low-quality reads (e.g., **Phred score**) using dedicated software packages such as pRESTO or FastQC [73, 79]. Subsequently, several studies have employed heuristic clonal abundance cutoffs (e.g., removal of clones with only 1–5 reads, Fig. 2.1 B) [60, 61, 18, 66] to decrease artificial diversity. Warren and colleagues showed that abundance filtering is superior to strict quality filtering in decreasing artificial diversity [55]. Furthermore, Bolotin and colleagues demonstrated that aggressive quality filtering can even lead to loss of a significant portion of the data [80]. In fact, lower-quality reads may be recovered from paired-end sequencing [81] (the inherently lower-quality 30 ends of sequencing reads gain in confidence via an overlapping region in both forward and reverse reads) or by merging lower-quality reads with reads of higher quality and identical or very similar clonal identifiers (clonotyping, see below and Fig. 2.1 B) [71, 80, 17, 82]. This ensures error correction and maximal data preservation while removing artificial diversity. Overall, however, it should be noted that clear guidelines for quality and abundance filtering do not yet exist [83, 84].

2.4.2 Clonotyping: Defining clonality from error-prone high-throughput sequencing data

The investigation of the complexity of lymphocyte clonality in health and disease represents the core purpose of the majority of analytical approaches (Fig. 2.2). While the definition of clonality in a biological sense is widely accepted (all lymphocytes having the same BCR or TCR belong to the same clone, see above), its translation to HTS data is challenging owing to the influence of PCR and sequencing errors, and of SHM.

A common approach is to cluster sequencing reads with high CDR3 homology (measured by edit string distance) as well as identical CDR3 length and V/J gene usage, and to refer to these as ‘clonotypes’ [39, 85]. Using publicly-available resources

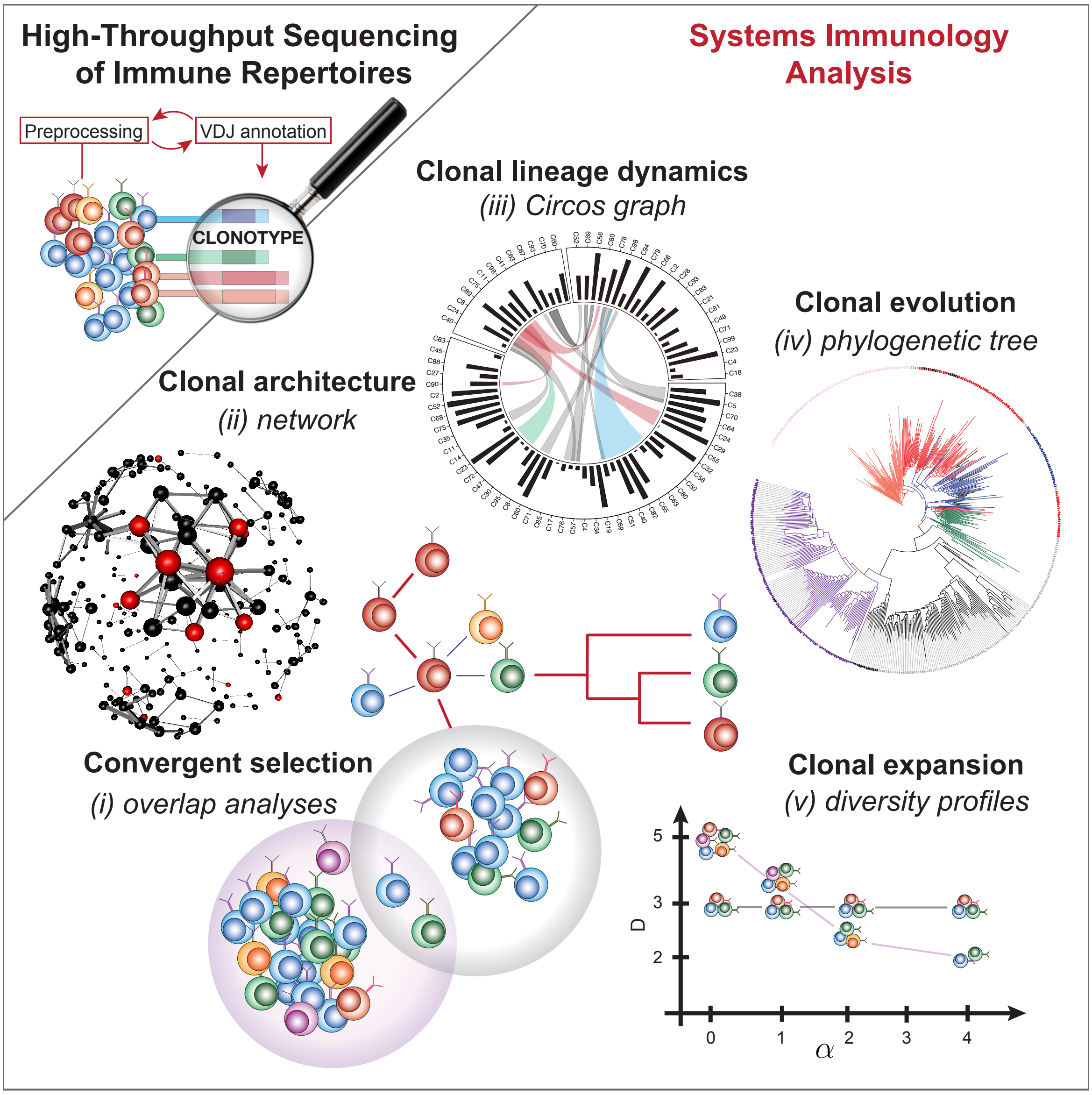


Fig. 2.2 Statistical analysis and visualization of high-throughput immune repertoires. The complexity of immune repertoire data necessitates sequence-dependent and sequence-derived analysis. Statistical analyses rely predominantly on clonotyped data and are therefore preceded by a workflow composed of raw data preprocessing (read filtering, error correction), germline annotation (Tab. 2.2), and clonotyping. Sequence-dependent approaches (i) visualize convergence of repertoires by quantifying clonal overlap (Venn diagrams; overlap indices such as Morisita–Horn, not shown); (ii) display the clonal architecture of repertoires (networks) highlighting denser (clonal expansion) or sparser regions of the repertoire (each vertex is a clone, the size of each vertex is proportional to its abundance, red color highlights selected clones); (iii) reveal dynamics of clones (Circos graphs) shared across samples (sections) by visualizing their change in frequency (bars); or (iv) retrace clonal evolution (phylogenetic trees) helping for instance the visualization of the phylogenetic relation of different clonal lineages (color-coded).

Fig. 2.2 Sequence-derived approaches consist mainly of (v) diversity (D) profiles (Box 1), which enable the comparison of repertoire diversity and clonal expansion (each line represents the diversity profile of one repertoire, that in purple being more clonally expanded). Legend: each color represents one clone.

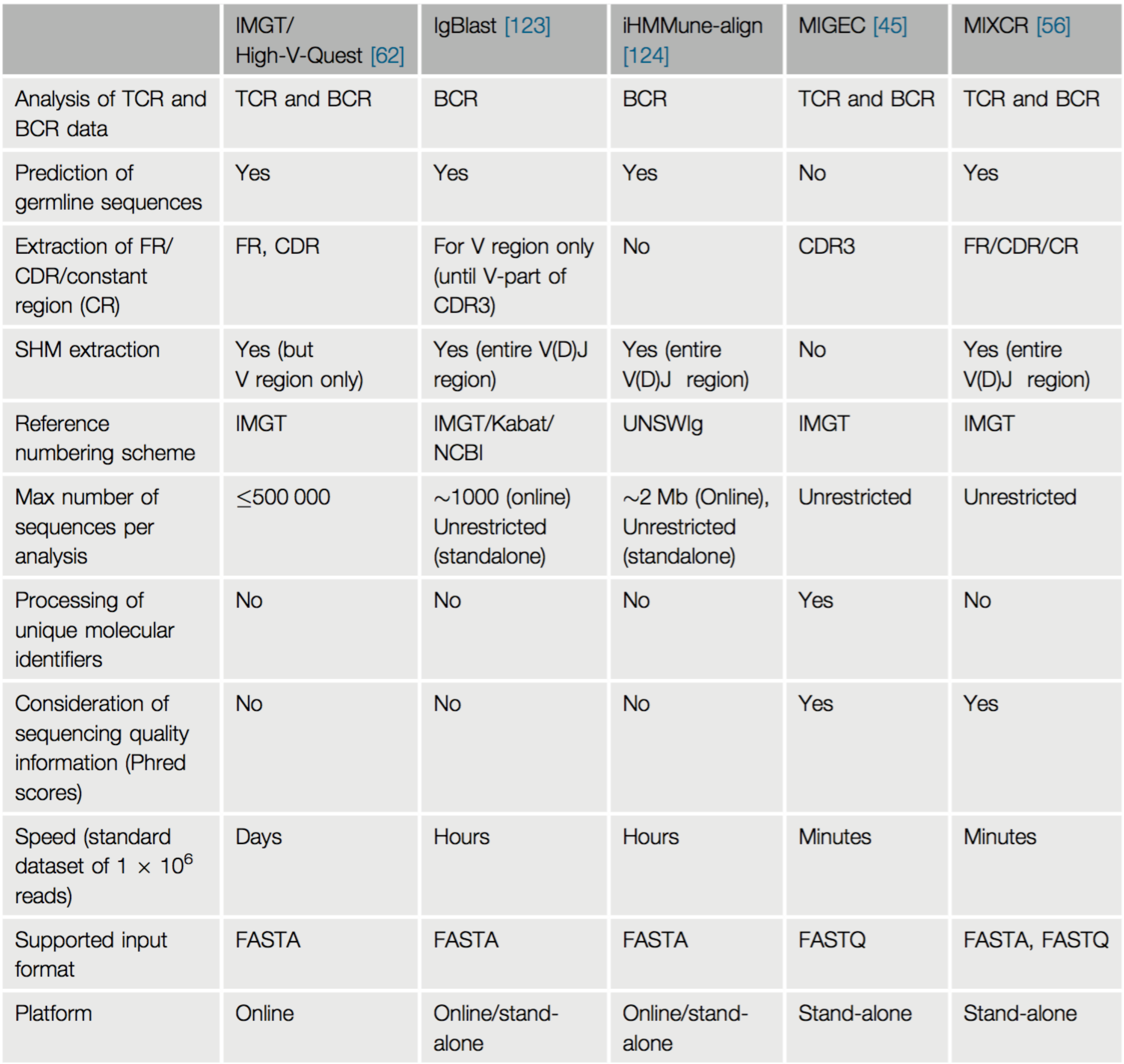
[71, 17, 86–89]and in-house developed software [90, 91], clustering by CDR3 homology at the nucleotide level has been performed in the following ways: (i) inferring unmutated common ancestors [92, 93]; (ii) absolute edit distance cutoffs in hierarchical clustering linkage trees [93], allowing a range of mismatches (one [80, 94, 95], three [96] [71], or five [40]) in sequences within one clonotype; or (iii) clustering by using relative thresholds (90% [90, 97], 95% [98], 97.25% [42], 100% [67] [41]). At the amino acid level, clonotypes have been built based on 80–100% CDR3 homology [43, 18, 19, 99, 100].

Clonotyping reduces the influence of PCR and sequencing errors on clonal diversity estimations but also, in the case of B cells, serves to group clones that belong to the same clonal lineage. A robust clonotype definition is, therefore, a defining step in every immune repertoire HTS study because it has a large impact on biological conclusions drawn (especially in diversity analyses, Fig. 2.2). Tipton et al. recently defined clonotypes by experimental validation as sequences with CDR3 (hamming) nucleotide identity of >85% using replicate sequencing [44].

2.4.3 Annotation of immune repertoire data

Raw sequencing reads require annotation for downstream statistical analysis (Fig. 2.2). Annotation tools vary widely with regard to the extent of output information, which can range from the sole identification of the CDR3 to comprehensive information (e.g., germline gene usage, framework regions, CDRs, and the extent of SHM, Fig. 2.3). While annotation speeds differ across several orders of magnitude (minutes to potentially days, Fig. 2.3), IMGT (International Immunogenetics Information System) has become the germline and **numbering scheme** database of choice for the main annotation platforms (Fig. 2.3). Because annotation accuracy can vary widely across different platforms [17], the use of simulated V(D)J repertoires [17, 101] may now offer the potential to help standardize annotation algorithms.

One limitation of germline reference databases is their inherent incompleteness: recent studies have highlighted the uncertainty regarding the extent of germline poly-



	IMGT/ High-V-Quest [62]	IgBlast [123]	iHMMune-align [124]	MIGEC [45]	MIXCR [56]
Analysis of TCR and BCR data	TCR and BCR	BCR	BCR	TCR and BCR	TCR and BCR
Prediction of germline sequences	Yes	Yes	Yes	No	Yes
Extraction of FR/ CDR/constant region (CR)	FR, CDR	For V region only (until V-part of CDR3)	No	CDR3	FR/CDR/CR
SHM extraction	Yes (but V region only)	Yes (entire V(D)J region)	Yes (entire V(D)J region)	No	Yes (entire V(D)J region)
Reference numbering scheme	IMGT	IMGT/Kabat/ NCBI	UNSWIg	IMGT	IMGT
Max number of sequences per analysis	≤500 000	~1000 (online) Unrestricted (standalone)	~2 Mb (Online), Unrestricted (standalone)	Unrestricted	Unrestricted
Processing of unique molecular identifiers	No	No	No	Yes	No
Consideration of sequencing quality information (Phred scores)	No	No	No	Yes	Yes
Speed (standard dataset of 1 × 10^6 reads)	Days	Hours	Hours	Minutes	Minutes
Supported input format	FASTA	FASTA	FASTA	FASTQ	FASTA, FASTQ
Platform	Online	Online/stand-alone	Online/stand-alone	Stand-alone	Stand-alone

Fig. 2.3 Characterization of the main annotation platforms.

morphism in humans [102] and even among widely utilized mouse strains with defined genetic backgrounds (e.g., BALB/c, C57BL/6) [16].

2.5 Immune repertoire analysis and visualization

2.5.1 Methods for quantifying clonal convergent selection, dynamics, architecture, and evolution

The use of Venn diagrams is a classic approach to studying convergence (or overlap) of repertoires (e.g., quantification of shared antigen-specific [54, 66, 103] or evolutionarily conserved public clones [59]) (Fig. 2.2). While Venn diagrams merely quantify the clonal sequence overlap, the **Morisita–Horn index** [104] (and other overlap indices [105]) quantifies the convergence of both clonal sequences and respective abundance across samples. The use of **Circos graphs** represents a recent advance in the visualization of overlap from complex large datasets [106]; for example, these plots have been used for studying the dynamics of B cell clonal expansion after influenza vaccination [54], and visualized the contribution of specific subsets of naive B cells to the compartment of antibody-secreting cells in autoimmune disease patients [44]. Unfortunately, Venn diagrams and Circos plots do not scale well with increasing numbers of samples, thus rendering the visualization of the clonal overlap of more than 10 datasets virtually impossible.

Although clonotyping and clonal lineage reconstruction are widespread in the literature, the quantification of overlap of entire clonotypes/clonal lineages across samples remains an unresolved issue of great importance. This is due to mathematical challenges in determining the overlap of sets (repertoire of clonotypes, clonal lineages), which are themselves composed of sets (sequences within clonotypes, clonal lineages). This problem has been partly circumvented by considering either core clonotypes (most-reliable and abundant sequences within clonotypes) [80] or by considering partial overlap of clonal lineage members [54].

Large-scale connectivity analysis between and within repertoires on both non-temporal and temporal scales has been attempted using **network** [94, 95, 107] and phylogenetic analyses [42, 44, 108] (Fig. 2.2). Network analysis was used for the visualization of differences in repertoire architecture of individuals of differential immunological status (e.g., healthy and cancer or HIV-attained individuals) by highlighting dense (highly connected clonally expanded clones) and sparse repertoire regions [64, 94, 95] (Fig. 2.2). These networks are usually constructed by drawing edges between clones (termed vertices or nodes) which differ by a given number of amino acid/nucleotide changes. The size of the vertices may be drawn relative to the abundance of a clone within the repertoire. This strategy enables one to relate clonal sequence architecture to clonal frequency distributions, thus further highlighting regions of the repertoire that have undergone potential disease-specific clonal expansions (Fig. 2.2). Immune repertoire networks have been visualized through the use of software packages such as igraph [94, 95, 107] and Gephi [64, 109].

In contrast to networks, phylogenetic analysis allows the reconstruction of clonal lineage evolution [42, 44, 108, 110] (Fig. 2.2). They have recently been applied for tracing a lineage of HIV-1 broadly-neutralizing antibodies over the timecourse of 15 years [42]. In addition, phylogenetic clustering was used to determine the pairing of antibody heavy and light chains from HIV-1 repertoires to discover novel neutralizing antibody variants [91]. The foundation of all phylogenetic tree construction algorithms represents a sequence alignment, which is fed into phylogeny (clonal tree) inferral algorithms such as IgTree [44, 108] or Phylip, the latter of which had been originally developed for applications in ecology and macroevolution [111–113]. Visualization of trees is often performed using Dendroscope [42, 114]. To date, neither network nor phylogenetic methods are well adapted to the complexity of immune repertoire HTS data, and additional work will be necessary to fully exploit these analytical techniques.

2.5.2 Methods for quantification of clonal Diversity, clonal expansion, and SHM

Recent studies suggest that, in general, lymphocyte repertoires are quasi-distinct in clonal composition (see discussion of this phenomenon and associated references in Greiff et al. [68]). This restricts sequence-dependent comparisons of immune repertoires across individuals to the comparably small number of public clones, thus disregarding the wealth of information present in entire immune repertoires. However, the comparisons of sequence-derived characteristics, such as diversity, clonal expansion, and SHM count, can be performed at the whole-repertoire level, thereby complementing sequence-dependent analyses [68].

The quantification of clonal expansion and repertoire diversity (Box 1) represents a major goal in HTS repertoire studies because it yields information on the current immunological status of a host [54, 68, 94, 115], which is particularly important for disease and vaccine profiling. The mathematical foundations of biological diversity assessment were developed decades ago for ecological research [116]. Several dedicated **R packages** already exist for diversity index calculations [117–120].

Box 1. Repertoire Diversity Analysis

The diversity (${}^{\alpha}$D) of a repertoire of S clones is usually calculated as follows (Hill diversity, [121])

$$ {}^{\alpha}D = \left(\sum_{i=1}^{S} f_i^{\alpha} \right)^{\frac{1}{1-\alpha}} \tag{2.1} $$

f_i is the frequency of the ith clone weighted by the parameter α. Special cases of this Diversity function correspond to popular diversity indices in the immune repertoire field: species richness ($\alpha = 0$), the exponential Shannon–Weiner ($\alpha \to 1$), the inverse of the Simpson index ($\alpha \to 2$), and the Berger–Parker index ($\alpha \to$). The higher the value of α, the higher becomes the influence of the higher-abundance clones on the diversity. Owing to the mathematical properties of the diversity function (Schur concavity [122]), two repertoires may yield qualitatively different ${}^{\alpha}D$ values depending on the diversity index used (see Figure 1 in Greiff et al. [68]). Diversity profiles, which are vectors of several diversity indices, have, therefore, been suggested to be superior to single diversity indices [68] and are increasingly used in repertoire analyses [68, 118, 123]. Fig. 2.2 shows two diversity profiles of two immune repertoires of differential clonal expansion. Of note, Chao et al. published recently a rarefaction framework for the Hill diversity formula [124], and this will enable the estimation of diversity profiles in the case of undersampled data.

To quantify clonal expansion, diversity can be divided into evenness $\frac{{}^{\alpha}D}{{}^{0}D}$ and species richness ${}^{0}D$ [68, 125]. Evenness ranges between 1 (uniform clonal population, every clone occurring in the frequency of $\frac{1}{{}^{0}D}$ and $\approx \frac{1}{{}^{0}D}$), in which case one clone completely dominates the immune repertoire.

Diversity indices are highly dependent on comprehensive and accurate sequencing [105]. While error correction approaches function to limit overestimation of repertoire diversity (see above), comprehensive sampling of repertoires is challenging to achieve owing to their heavy-tailed clonal frequency distributions – that is, few highly-abundant clones and many low-abundance clones [68, 126–128]. The precision of diversity calculation in case of insufficient sampling can be increased using diversity index estimators [105, 124, 129]. Furthermore, Laydon and colleagues recently published a novel rarefaction-based method for estimating total repertoire size [129, 130], which offers advantages to commonly used estimators of species richness such as [124, 130] and Good–Turing [105, 131].

Adding to the problems in repertoire diversity analysis, it has been found that single diversity indices might lead to contradicting qualitative outcomes depending on the diversity measure used ([122] and Box 1). This could yield qualitatively different conclusions regarding the clonal expansion status of a given repertoire [68], and would be especially problematic in the example of clinical lymphoma and immune disorder monitoring [68, 94, 132, 133]. Several groups [102, 123] including ourselves [68] have recently published a potential solution to this problem in the form of diversity profiles (Box 1).

While diversity profiles are suitable for comparative analyses of clonal diversity and expansion, estimates of total repertoire size remain an unresolved issue. However, estimates of total repertoire size remain an unresolved issue [55, 130, 134]. Reasons for this are: (i) so far it has remained a challenge to cover immune repertoires in their entirety (except for smaller ones such as that of zebrafish [15]); (ii) the discrimination between rare clones and sequencing errors remains a challenge [55]; and (iii) the absence of a validated framework for the diversity profile estimation of undersampled immune repertoire data (Box 1).

SHM is a defining step of clonal selection and expansion during the generation of the B cell response [11]. Annotation methods in Fig. 2.3 determine SHM counts independently of clonal-relatedness by alignment with germline reference databases, while other approaches assess SHM phylogenetically [108]. For fundamental immunology, the elucidation of SHM patterns is of high importance for the understanding of how activation-induced deaminase (AID) targets V(D)J regions [135] . In HIV, high SHM counts are a hallmark of HIV-specific broadly neutralizing antibodies [136], an outstanding feature that could be exploited for discovery of novel therapeutic candidates. The inability, however, to unequivocally separate SHM from PCR and sequencing error [39], as well as the incompleteness of reference germline databases [102], remain fundamental problems in obtaining true absolute SHM counts. Therefore, performing relative SHM analyses, which quantify the differences of SHM counts between cell populations of interest, together with appropriate controls (e.g., naive B cells, or synthetic spike-ins), may be preferable to interpretations based on absolute SHM counts.

2.6 Concluding remarks

The increased application of immune repertoire HTS to research in immunodiagnostics [68, 137], immune response profiling [54, 110, 138, 139], antibody engineering [42, 140, 141], and lymphocyte development [142, 143] continues to expand the rapidly developing field of systems immunology. However, a lack of standardization in bioinformatic and statistical analysis renders the comparison of results across studies challenging. The sensitivity of immune repertoire data analyses would dramatically benefit from the establishment of standards for HTS data pre-processing (e.g., quality filtering, clonotype definition, etc.). Moreover, further development of visualization methods for these high-dimensional, highly diverse, and interconnected data will improve the knowledge gained from immune repertoires (see Outstanding Questions).

Systems immunology-driven studies hold the promise of resolving some longstanding questions in adaptive immunity: (i) what principles drive immune repertoire construction; (ii) what is the size and extent of variation of the expressed immune repertoire; and (iii) are immune repertoires complete in the sense that they could recognize any antigen [134, 144]? Answering these questions will require detailed knowledge of interindividual germline variance [102], statistical models of repertoire generation [145–148], and continuous technological advances in DNA sequencing technology and computational biology.

2.7 Outstanding questions

- How to standardize HTS and the analysis of immune repertoires? An experimental framework mimicking the large diversity of immune repertoires for the unbiased validation of HTS library preparation methods (PCR, primer bias, and error correction) is missing. Similarly, a standardized repertoire simulation framework for validating bioinformatics processing and analysis pipelines remains to be developed.

- How to visualize the connectivity of immune repertoires? Network and phylogenetic analyses are currently visually and computationally unsuitable for large datasets (>10 000 sequences; one small HTS dataset usually has >100 000 sequences). In addition, both phylogenetic and network results can vary substantially in the parameter values used (clonotyping parameters, alignment method, molecular clock model, substitution model).

- How to analyze antibody repertoire evolution across time-course (longitudinal) samples? The establishment of a mathematical framework for clonotype/lineage overlap and phylodynamic network analyses on complex datasets will be necessary to investigate antibody (immune) repertoire evolution on a large scale.

2.8 Glossary

Biological replicates: HTS of different samples of the same underlying cell population (e.g., partitioning of PBMC, peripheral blood mononuclear cells). Biological replicates are used to assess biological sampling.

Biological sampling: the cell population sampled must be an approximate representation of the cellular compartment being investigated to allow meaningful conclusions to be drawn from the data.

Circos graph: a circular layout plot for the visualization of quantitative and qualitative relationships in complex and large datasets. In immune repertoire HTS data, it is used mainly to visualize frequencies of overlapping clones across timeresolved longitudinal data.

Clonal frequency distribution: for a given immune repertoire, the clonal frequency distribution describes the distribution of the number of sequencing reads (read abundance) that are allocated to each clonotype (commonly referred to as ‘clone size’). The underlying power law of clonal frequency distributions is commonly visualized by plotting the logarithm of clonotype frequency as a function of the logarithm of clonal rank.

Clustering of sequences: clustering is the process of grouping a set of similar sequences (nt/aa sequences defined as strings of characters) in the same group based on a given sequence identity threshold. Hierarchical clustering is a connectivity algorithm that forms clusters of sequences based on their string distance.

Morisita–Horn overlap index: this is used to compare species (e.g., clone, germline genes) sequence and abundance overlap between any two immune repertoires. It is defined as

$$MH = \frac{2\sum_{i=1}^{S} x_i y_i}{\sum_{i=1}^{S} x_i^2 + \sum_{i=1}^{S} y_i^2} \tag{2.2}$$

where S is the number of unique species, and x and y denote the frequency of the ith species in either repertoire. The MH index ranges between 0 (no overlap) and 1 (complete species overlap and identical species frequencies).

Network: a network is a measurable pattern of relations among subunits. It represents a graph composed of a set of objects (vertices, nodes) and links (edges).

Numbering schemes: complementarity determining regions and framework regions are identified as amino acid strings by different numbering schemes (i.e., IMGT, Kabat, Clothia). Numbering schemes define the start and ending positions of BCR and TCR regions.

Phred score (Q): a measure for quality base calling. It is defined as $P = -log_{10}P$, where P is the base-calling error probability. For example, if $Q = 30$ for a given base, the probability that the base was called incorrectly is $P = 10^3$.

R package: R is a statistical programming environment, and its package system enables the flexible and constant addition of newly developed statistical approaches.

Reliable clonal detection cutoff: clones in datasets of technical replicates are ranked in decreasing order of frequency and tested for simultaneous presence in all replicates to construct a list of reliably detected clones, which together are at least for example 90% (cutoff) present in all replicates. The reliable detection cutoff is valid for all HTS datasets prepared with experimental conditions identical to those of the technical replicates. Importantly, the meaningful application of reliable detection cutoffs depends on (near)-complete sample coverage.

Species accumulation and rarefaction analysis: species accumulation curves display the rate at which new clones are discovered with increasing number of sequencing reads. By contrast, rarefaction curves are used to estimate the number of clones at a particular level of sampling.

String distance: measures dissimilarity between any two sequences (e.g., germline reference sequence and sequencing reads for V(D)J annotation or two CDR3s for clonotyping) by counting the minimum number of operations required to transform one string into the other. Levenshtein or edit distance accounts for insertions, deletions and substitutions.

Technical replicates: replicate sequencing of the same immune repertoire library. A strict definition would be the resequencing of the same library, whereas a more lenient definition would consider also molecular replicates (separate library preparation of the same genetic material) adequate provided that biological replicates have been performed to exclude biological undersampling. Technical replicates are used to assess technological sampling.

Technological sampling: ensuring that the number of sequencing reads exceeds the molecular diversity, or at least, the clonal diversity of the underlying sample.

Unique molecular identifiers (UMIs): pseudo-random sequences of several degenerate nucleotides (usually 8–12), which are added during library preparation by reverse transcription or ligation. Sequencing reads with identical UMIs are merged (consensus read construction) thus increasing the confidence in each base call, and consequently reducing the extent of PCR and sequencing error.

Chapter 3

The fundamental principles of antibody repertoire architecture

The antibody repertoire is a vast and diverse collection of B-cell receptors and antibodies that confer protection against a plethora of pathogens. The architecture of the antibody repertoire, defined by the network similarity landscape of its sequences, is unknown. Here, we established a novel high-performance computing platform to construct large-scale networks from high-throughput sequencing data (>100,000 unique antibodies), in order to uncover the architecture of antibody repertoires. We identified three fundamental principles of antibody repertoire architecture across B-cell development: reproducibility, robustness and redundancy. Reproducibility of network structure explains clonal expansion and selection. Robustness ensures a functional immune response even under extensive loss of clones (50%). Redundancy in mutational pathways suggests that there is a pre-programmed evolvability in antibody repertoires. Our analysis provides guidelines for a quantitative network analysis of antibody repertoires, setting the stage for the field of network systems immunology, and may direct the construction of synthetic repertoires for biomedical applications.

3.1 Introduction

The high diversity of antibody repertoires enables broad and protective humoral immunity, thus understanding their system sequence-related properties is essential to the development of therapeutics and vaccines [149, 150]. The source of antibody diversity has long been identified to be the V-, (D- in the heavy chains) and J-gene somatic recombination [10]. Further additions and deletions of nucleotides at the junctions of the gene segments generate a large collection of antibodies and B-cell receptors, which is called the antibody repertoire [4, 15]. Antibody identity (clonality) and antigen specificity are primarily encoded in a portion of the variable heavy chain, at its junctional site of recombination, by the highly diverse complementarity determining region 3 (CDR3) [11]. Therefore, the similarity landscape of CDR3 sequences constitutes the clonal architecture of an antibody repertoire, which reflects the breadth of antigen-binding and correlates to humoral immune protection and function. However, the fundamental principles that govern antibody repertoire architecture have remained unknown, thereby hindering a profound understanding of the immune system.

Recently, different aspects of network analysis have been employed to investigate antibody repertoires in health and disease. Antibody repertoire networks represent CDR3 sequence-nodes connected by similarity-edges [94, 151, 152, 95, 64]. Sequence-based networks have first been used to show immune responses defined by similarity between clones, a proxy for clonal expansion [151]. Network connectivity was later also used to discriminate between diverse repertoires of healthy individuals and clonally expanded repertoires from individuals with diseases like chronic lymphocytic leukemia [94] and HIV-1 infection [95]. A predominant part of network analysis has involved visualization of clusters and the display of clonal composition [94, 151, 152, 95, 64]. Yet, visualization does not provide quantitative insights into the architecture of antibody repertoires and is limited to the informative graphical display of a few hundred nodal clones. It has been shown that the natural antibody repertoire exceeds the informative visualization threshold by at least three orders of magnitude [37, 153], a limit that previous research did not explore given the lower biological coverage (10^2-–10^3 unique clones analyzed). Consequently, computational methods for constructing large-scale networks with more than 10^3 nodes have remained underdeveloped in systems immunology [30]. Furthermore, only networks expressing clonal similarity relations of 1 nucleotide (nt) or a.a. between sequences have been analyzed so far. Thus, the lack of quantitative investigation of an (exceedingly) small subset of the antibody repertoire, with respect to clone numbers and network size, has limited the biological insight into the repertoire architecture.

To reveal the fundamental principles of antibody repertoire architecture, we implemented a large-scale network analysis platform coupled to high-coverage antibody repertoire high-throughput sequencing data to answer the following questions: (i) Does sequence similarity among clones show reproducible signatures across individuals? (ii) How robust are antibody repertoires to removal of a fraction of clones, given their kinetics and rapid turnover? (iii) To what extent is the repertoire architecture intrinsically redundant? (See Fig. 3.1).

3.2 A high-performance computing platform for the generation and analysis of large-scale antibody repertoire networks from high-throughput sequencing data

The global landscape of clonal similarities is vast and complex; for example, the size of the distance matrix of all-against-all sequences is 10^{10} for a representative repertoire of 10^5 clones (murine naïve B cells). In order to extract the construction principles of antibody repertoires from the high-dimensional similarity space, we developed a large-scale network analysis approach, which was based on representing CDR3 clones as sequence-nodes connected by similarity-edges. Specifically, we developed a computational platform that leverages the power of distributed cluster computing able to compute the extremely large distance matrices required for entire repertoires ($\geq 10^5$ CDR3, Fig. 3.2). Our implementation utilized the Apache Spark [154] distributed computing framework to partition the work among a cluster of machines (Fig. 3.2 B). The construction of large-scale networks is computationally demanding: the construction of a large network from 1.6 million nodes (simulated sequences) required 15 minutes if the calculation was performed simultaneously on 625 computational cores (Fig. 3.2 C). Importantly, computational costs could have been lowered by performing network analysis on a subsample of the repertoire (e.g., 10^3) similar to previous approaches [94, 151, 152, 95, 64]. However, it has been previously demonstrated that sub-networks are statistically not representative of entire networks and typical network quantities as evidence by sampling-sensitive network measures such as degree distribution, betweenness, assortativity and clustering [155, 156]. Thus, it was imperative to construct and analyze large-scale networks based on a similarity distance matrix that covers the full clonal diversity of biological repertoires.

Biological comprehensive sampling of antibody repertoires was ensured by the usage of high-throughput RNA sequencing data (400 million full-length antibody sequence reads) from murine B-cell populations, isolated at key stages in humoral development [153]. Data was analyzed from pre-B cells (pBC), naïve B cells (nBC) and memory

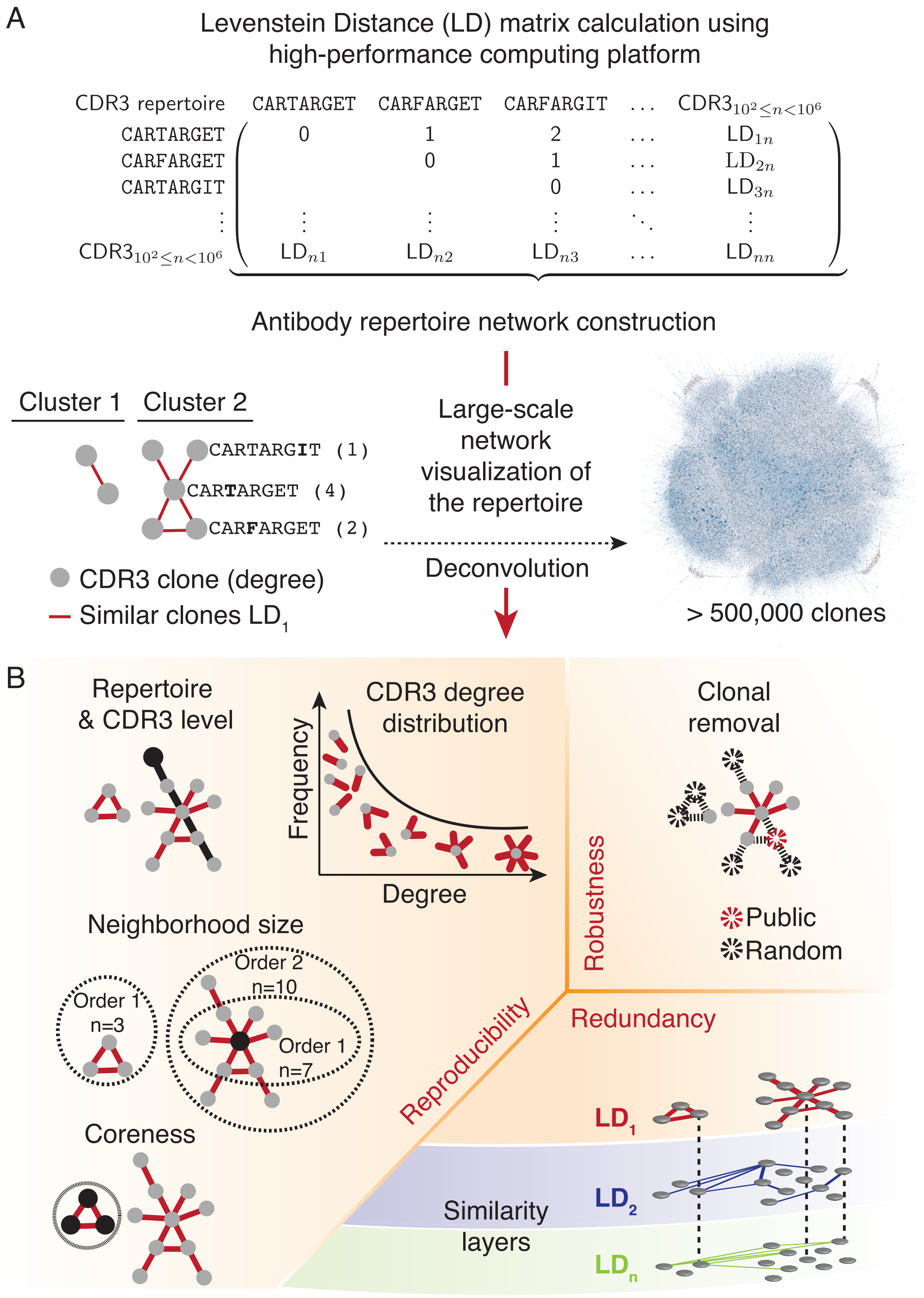


Fig. 3.1 Large-scale network analysis reveals the architecture of antibody repertoires and its three main principles.

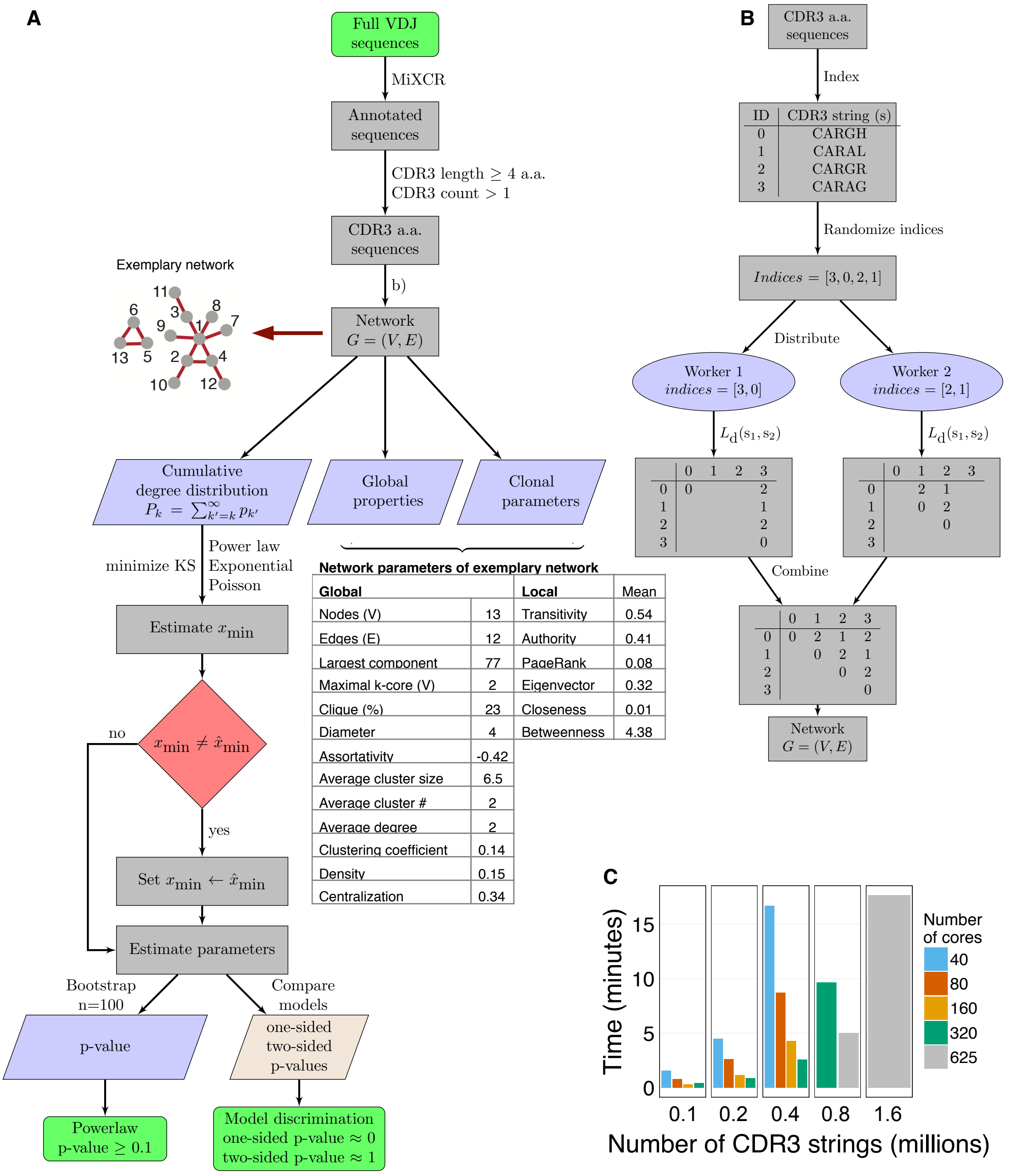
Fig. 3.1 (A) Large-scale networks (>500’000 nodes) of antibody repertoires were constructed from the Levenshtein distance (LD, edit string distance) matrix of CDR3 clonal sequences (a.a) using a custom high-performance computing platform (see Methods 3.8). Networks represent antibody repertoires of similar CDR3 nodes connected by edges when amino acid CDR3 sequences differ by a predetermined LD. All clones of a repertoire connected at a given LD form a similarity layer (LD_n). (B) Deconvolution of the complexity of antibody repertoire architecture was performed by quantifying (i) its reproducibility through global and clonal (local) properties, (ii) robustness to clonal deletion and (iii) redundancy across its similarity layers in the sequence space (Fig. 3.2).

plasma cells (PC) isolated from 19 mice, which were stratified into 1 unimmunized and 3 immunized cohorts. The experimental design of the data allowed for the assessment of antibody architecture across several important parameters: i) across key stages of B-cell development, ii) before (pBC, nBC) and after antigen-driven clonal selection and expansion (PC), (iii) differences in the complexity of the antigen (HBsAg, OVA and NP-HEL), and (iv) across a scale of different repertoire sizes (10^2—10^5 of unique CDR3 clones). The experimental data provided maximal technological and high biological coverage [153] allowing to cover the clonal diversity of the antibody repertoires, and thus to capture their global similarity landscape and architecture.

For each sample (n=57, from 19 mice and 3 B-cell stages), antibody repertoire architecture was based on the pairwise sequence similarity of all clones (Levenshtein distance (LD) matrix, hereafter referred to as similarity layer, Fig. 3.1 A). When two sequences were similar within a defined threshold, they were connected in the repertoire network (i.e., similarities of 1 a.a. differences were captured in similarity layer 1 (LD_1), 2 a.a. in LD_2 and so on).

3.3 Global patterns of antibody repertoire networks are reproducible

In order to quantify the extent to which architectural patterns are reproducible across antibody repertoires, we analyzed the base similarity layer in antibody repertoires (similarity layer LD_1). The base layer of the network organization provides information regarding the minimal differences (e.g., 1 a.a.) of all antibody sequences that compose



Network parameters of exemplary network

Global		Local	Mean
Nodes (V)	13	Transitivity	0.54
Edges (E)	12	Authority	0.41
Largest component	77	PageRank	0.08
Maximal k-core (V)	2	Eigenvector	0.32
Clique (%)	23	Closeness	0.01
Diameter	4	Betweenness	4.38
Assortativity	-0.42		
Average cluster size	6.5		
Average cluster #	2		
Average degree	2		
Clustering coefficient	0.14		
Density	0.15		
Centralization	0.34		

ID	CDR3 string (s)
0	CARGH
1	CARAL
2	CARGR
3	CARAG

	0	1	2	3
0	0			2
1				1
2				2
3				0

	0	1	2	3
0		2	1	
1		0	2	
2			0	
3				

	0	1	2	3
0	0	2	1	2
1		0	2	1
2			0	2
3				0

Fig. 3.2 High-performance computing platform to construct and analyze large-scale networks from entire antibody repertoires. (A) Data preprocessing, network construction and model fits to degree distribution (see Methods 3.8, Degree distribution fits for further details). Network parameters (global and mean local/clonal) are shown for the exemplary network. (B) Software schematics showing the distributed parallel computing platform used to partition the work among a cluster of many workers. (C) Computation time to construct large-scale networks depends on the number of CDR3 sequences and the number of cores used.

the repertoire. Solely the base similarity layer, LD_1, has previously been analyzed to describe antibody repertoires as networks [94, 151, 152, 95, 64]. Although repertoires varied highly among mice (74–85% clonal variability, (Fig. 3.4 A), we found a remarkable cross-mouse consistency in patterns of clonal interconnectedness (similarity of antibody sequences) within each B-cell stage: the number of edges among clones ($E_{\mathrm{pBC}} = 230,395$, $E_{\mathrm{nBC}} = 1,016,928$, $E_{\mathrm{PC}} = 45$), the size of the largest component (pBC=46%, nBC=58%, PC=10%, Fig. 3.3 A) and cluster composition (Fig. 3.4 B) varied negligibly across mice (see Methods 3.8, Network analysis). Thus, although sequences varied substantially between individuals, the overall structure of the network remained similar.

However, along B-cell development, PC repertoires were five-fold more disconnected than pBC and nBC networks (largest component pBC=46%, nBC=58%, PC=10%, Fig. 3.3 A), and their centrality was concentrated on specific clones compared to the homogeneously connected clones in pBC and nBC networks (centralization, $z_{\mathrm{PC}} = 0.05$, density $D_{\mathrm{PC}} = 0.01$, $z_{\mathrm{pBC,\ nBC}\approx 0}$, $D_{\mathrm{pBC,\ nBC}} \approx 0$, Fig. 3.4) C). Compared to pBC networks, nBC were on average 4–5 times larger and showed a higher average degree ($k_{\mathrm{pBC}=3}$, $k_{\mathrm{nBC}=5}$, $k_{\mathrm{PC}=1}$, Fig. 3.4 B) although both pre- and naïve B-cell repertoires had identical diameter ($d_{\mathrm{pBC,nBC}} = 26$, $d_{\mathrm{PC}} = 6$, Fig. 3.3 B), indicating a similar coverage of the sequence space. We observed that clones in pBC and nBC repertoires connected to comparable clones in terms of degree (assortativity [157**?** , 158], $r_{\mathrm{pBC}} = 0.48$, $d_{\mathrm{nBC}} = 0.41$), while PC networks were consistently disassortative: their highly connected clones were linked to clones with few connections ($r_{\mathrm{PC}} = -0.09$, Fig. 3.3 B). The characterization of the global patterns of antibody repertoire networks indicated that pBC, nBC and PC repertoires were reproducible. pBC and nBC clones cover a larger diversity space than clones in PC repertoires, where sequence similarity showed to be centralized and targeted towards certain clones.

3.4 Clonal features of antibody repertoire networks are reproducible

Antibody repertoire architecture was also reproducible at the level of clonal (local) features in pBC and nBC networks, which were characterized by a low variability (coefficient of variation, CV) across various clonal parameters. The low variability of clonal parameters in pBC and nBC networks ($CV_{\mathrm{pBC}} = 2-28\%$, $CV_{\mathrm{nBC}} = 1-24\%$) was

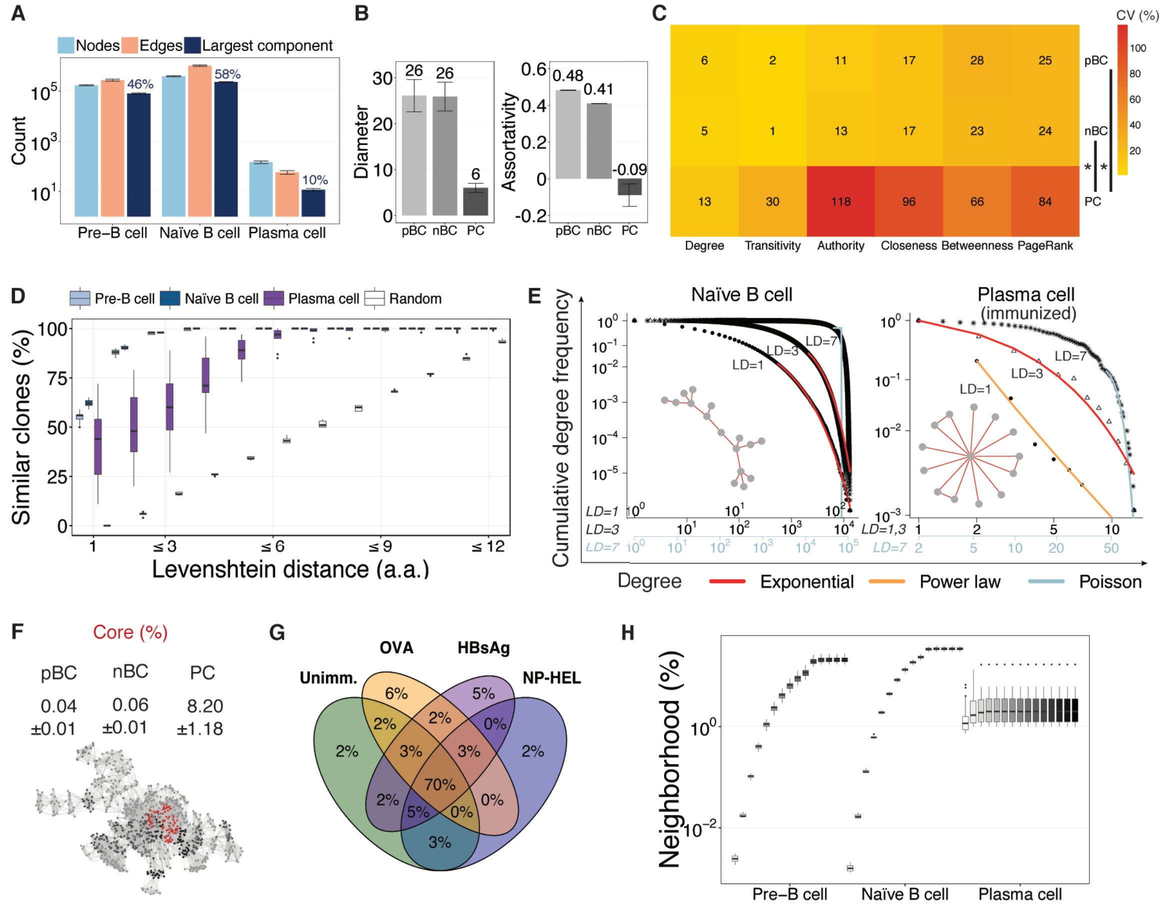


Fig. 3.3 Global and clonal properties of antibody repertoire networks are reproducible. (A) Network size of antibody repertoires. The y-axis indicates the absolute number count of CDR3 nodes, CDR3 edges (similarities) and CDR3 clones in the largest component. The mean percentage of the CDR3s belonging to the largest component is shown on top of the largest component bar. (B) Global properties, diameter and assortativity coefficient are shown for pre-B cells (pBC), naïve B cells (nBC) and plasma cells (PC). (C) The mean value of the coefficient of variation for clonal properties in pBC, nBC and PC repertoires. Wilcoxon test, $p_{\text{pBC,nBC/PC}} < 0.05$ (see Methods 3.8). (D) Percentage of clones connected to at least one other clone in the repertoire at LD_1, $LD_{\leq 2}$, . . . , $LD_{\leq 12}$ in pre-B cells, naïve B cells, plasma cells and randomly constructed CDR3 strings. (E) The power-law (orange), exponential (red) and Poisson (grey) distributions were fit to the cumulative degree distributions of naïve B cell and plasma cell (unimmunized) repertoires of a mouse for similarity layers $LD_{1,3,7}$ (log-log scale). Representative clusters are shown for LD_1. (F) Percentage of CDR3 clones (mean±s.e.m) that compose the maximal core. Subgraph of the maximal k-core (red), and k-1 (black), k-2 (dark grey) and k-3 (light grey) cores in a pBC sample. (G) Percentage overlap of CDR3 germline V-genes in the maximal core of nBC repertoires. (H) Normalized neighborhood size for orders n=1—10, 15, 20, 30, 40, 50 across CDR3 clones (similarity layer LD_1). Barplots show mean±s.e.m. (Fig. 3.2).

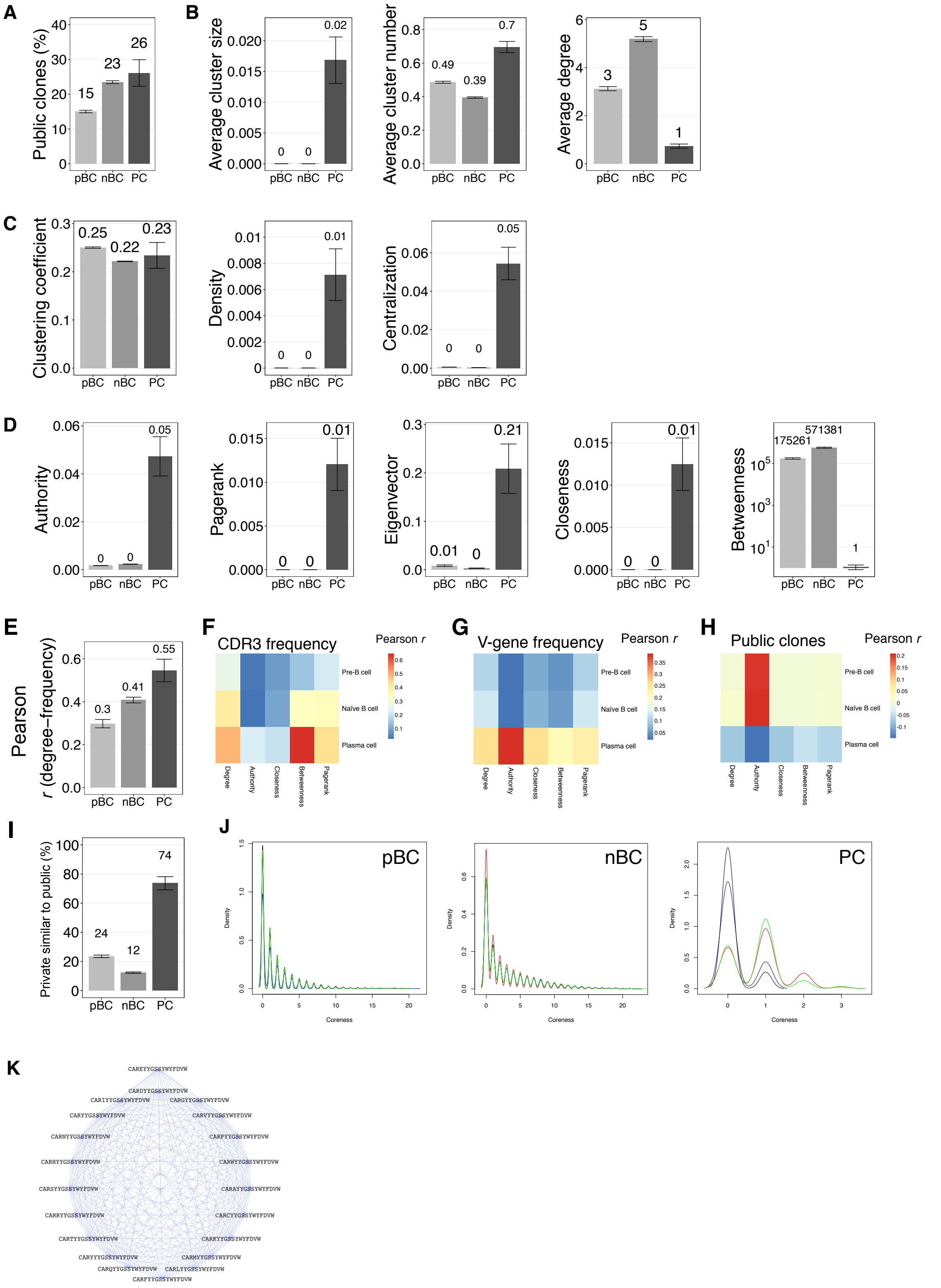


Fig. 3.4 Global and clonal (local) network parameters of antibody repertoires of pre-B cells (pBC), naïve B cells (nBC) and plasma cells (PC).

Fig. 3.4 (A) Percentage of public clones, shared CDR3 clones between mice in pre-B cell (pBC), naïve B cell (nBC) and plasma cell (PC) repertoires (mean±s.e.m, mice n=19). (B—C) Global properties: Cluster analysis shows the average normalized cluster size and cluster number in the antibody repertoire networks. Average degree, clustering coefficient, density and (degree) centralization characterize the networks at the global level (mean±s.e.m, mice n=19). (D) Local properties: authority, PageRank, eigenvector, closeness and betweenness describe each clone in the network. Average values are shown for each B cell population, pre-B cells (pBC), naïve B cells (nBC) and plasma cells (PC). Barplots show mean±s.e.m, mice n=19. (E) Pairwise Pearson correlation (r, mean±s.e.m) of CDR3 degree with CDR3 frequency in pre-B cells (pBC), naïve B cells (nBC) and plasma cells (PC) antibody repertoire networks. (F) Pairwise Pearson correlation of local properties with CDR3 frequency (median, mice n=19). (G) Pairwise Pearson correlation of local properties with germline V-gene frequency (mean, mice n=19). (H) Pairwise Pearson correlation of CDR3 clonal (local) properties with public (1) vs. non-public (0) CDR3 clones (mean, mice n=19). (I) Percentage of public clones similar (connected) to at least one other public CDR3 clone sequence by cohort (mean, mice n=19). (J) Coreness density distribution for the unimmunized cohort of pre-B cells (pBC), naïve B cells (nBC) and plasma cells (PC). The x-axis shows the k-core (after removing sequentially shells of nodes of degree k-1). Line colors depict different mice.

in contrast to the higher variability observed in PC repertoires ($CV_{\mathrm{nBC}} = 13 - 118\%$, Fig. 3.3 C). Specifically, low variability across different individuals was observed in several average clonal parameters such as degree, transitivity, authority and PageRank, closeness and betweenness. Variation analysis of the similarity degree indicated that the average number of similar clones to each of the clones in a repertoire varied marginally in pBC and nBC ($CV_{\mathrm{pBC,nBC}} = 5, 6\%$). Transitivity showed that the similarity between clones both similar to a third CDR3 clone varied only negligibly between individuals ($CV_{\mathrm{pBC,nBC}} = 1, 2\%$). Authority and PageRank showed that the centrality of a CDR3 in the repertoire topology varied respectively $CV_{\mathrm{pBC,nBC}} = 11\%$ and 25% across individuals, suggesting that individual repertoires were centered variably around certain CDR3 clones which were centers of highly connected (similar) clonal regions compared to less connected regions in the same repertoire network.

Closeness analysis revealed that an analogous number of similarity edges were required to access every other CDR3 from a given CDR3 clone in antibody repertoire networks of different individuals, as the similarity of a clone to every other CDR3 clone in the repertoire varied by $CV_{\mathrm{pBC,nBC}} = 17\%$. Betweenness, the “bridge” function of a clone in sequence similarity, varied slightly across individuals with $CV_{\mathrm{pBC,nBC}} = 28\%$, suggesting a comparable structure of the similarity route function of CDR3 sequences in these repertoires. These characteristics reflect the transversal diversity of pBC and nBC antibody repertoires where the clones cover a larger space and their similarity is more homogenously distributed at the global repertoire level.

Although a higher variability was detected across PC repertoire networks (Fig. 3.3 C), clonal parameters were specific to B-cell stages ($p_{\text{pBC,nBC/PC}} < 0.05$): PC clones possessed higher centrality compared to pBC and nBC (closeness [159], eigenvector [160], and PageRank [161]), while antigen-inexperienced clones bridged sequence similarity (betweenness [162], Fig. 3.4 D). Furthermore, in contrast to pBC and nBC, PC network clonal parameters correlated with CDR3 frequency (clonal degree median $r_{\text{Pearson}} = 0.55$, betweenness $r_{\text{Pearson}} = 0.82$) suggesting that clonally expanded CDR3 sequences were structural centers of similar clones (Fig. 3.3 E, F). CDR3 authority correlated positively with germline V-gene frequency in PC clones ($r_{\text{Pearson}} = 0.39$), denoting the potential role of the V-gene usage in the centralization of these networks (Fig. 3.3 G). Thus, certain high frequency V-genes predispose clones to be highly connected and similar to other clones.

3.5 The structure of antibody repertoires is reproducible and depends on the immune status

Network analysis revealed that antibody repertoires were constricted along B-cell development throughout all similarity layers. At LD_1, 44–62% of clones were similar (connected) to at least one other clone in all B-cell stages, revealing a high degeneracy in clonal generation and selection (Fig. 3.3 D), compared to nearly 0% of connections in randomly generated CDR3 strings representative of an unbiased repertoire, simulating an unbiased V(D)J recombination, with random deletions and insertions (see Methods 3.8). This indicated that around half of the antibodies in the repertoires had similar clones, thus showing a constriction of the biological repertoire compared to its unbiased random repertoire diversity.

In order to understand if such degeneracy in CDR3 sequence similarity translated into reproducible repertoire network structures [160], we determined the clonal empirical degree distribution. The degree distribution is a distinctive feature of different types of networks and it provided an immediate indication of how similarities (degrees) between antibody sequences were distributed in repertoires. Analysis of the cumulative degree distribution revealed that antigen-inexperienced pBC, nBC and unimmunized PC repertoires were exponentially distributed (LD_1), whereas PC repertoires of immunized cohorts were power-law distributed (base similarity layer LD_1, Fig. 3.3 E, Fig. 3.5 D, E, F, G). Clusters of connected CDR3 clones showed a typical tree-like structure for pBC and nBC, and a star-like structure for PC. The structure of the network suggested an extended and chain-wise sequence similarity of the antibody clones in pBC and nBC repertoires and targeted expansion of certain clones in PC after immunization.

In order to prove the tree-/star-like hypothesis and further investigate the sequence similarity space, we performed k-core (Barabási and Oltvai, 2004) decomposition (Fig. 3.3 F, G, H) and neighborhood analysis. The k-core decomposition revealed that the largest k-cores (after all external shells with $k<k_{\max}$ were removed, where k is the degree, i.e. number of similar clones, see Methods) of pBC and nBC (0.04% and 0.06% of CDR3 clones in k-core, respectively) were 200-fold smaller than those of PC (8.2%, Fig. 3.3 F). Antigen-inexperienced repertoires were characterized by larger coreness values (>20), signifying a more layered structure of CDR3 similarity (Fig. 3.4 J, K) and confirming their tree-like structure. Furthermore, the high convergence of V-genes at the core-level of antibody repertoire networks (pBC=50%, nBC=70%, PC=1–10%, Fig. 3.3 G), in contrast with the low exact CDR3 sequence core-overlap (Fig. 3.5 A, B, C), suggested a genetically determined origin of the structure.

The average CDR3 neighborhood size, which designated the set of similar CDR3 clones along each sequential step of similarity from a certain clone (orders n=1–50), was order-independent in PC and plateaued at 2% of the network, confirming that PC clones were connected to one central clone in a star-like similarity structure, reflecting clonal selection and expansion signatures. Neighborhood size [163], so the number of similar clones to each clone, increased order-wise in antigen-inexperienced cells up to 34% (Fig. 3.3 H), signifying tree-like similarity structures that enable maximal exploration of sequence space within the genetically predetermined repertoire constriction space, suggesting that antibody repertoires are evolutionarily wired to respond to diverse antigenic stimuli.

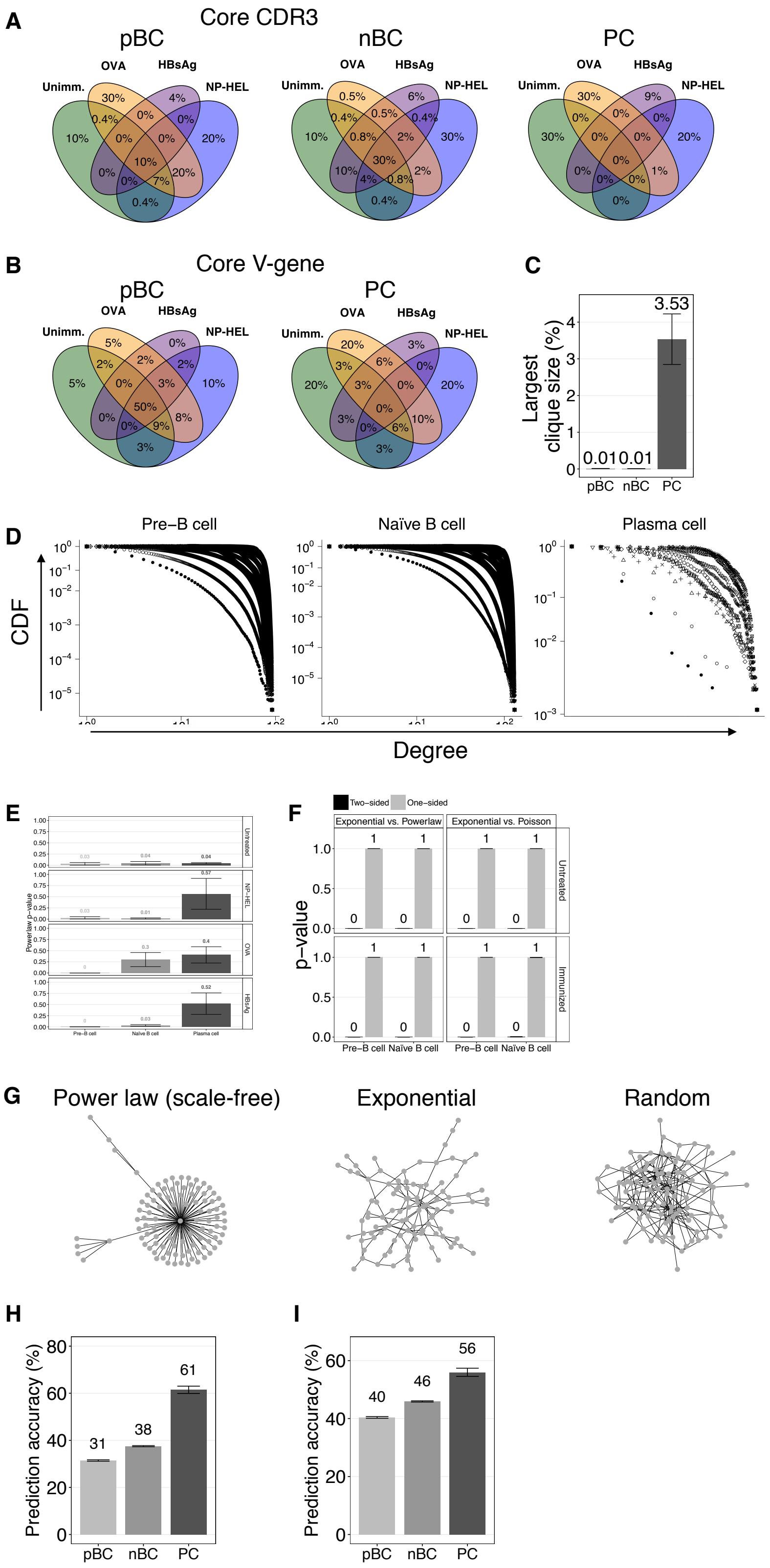


Fig. 3.5 Core and structure (degree distributions of CDR3 similarity) analysis, and similarity layer prediction of antibody repertoire networks.

Fig. 3.5 (A) Maximal core CDR3 clones overlap in pre-B cells (pBC), naïve B cells (nBC) and plasma cell (PC) repertoire networks. (B) Maximal core germline V-genes overlap in pre-B cells and plasma cell. (C) Percentage of the largest cliques (completely connected subgraph, mean±s.e.m, mice n=19) along B-cell development. (D) Cumulative degree distributions (CDF). Each distribution line (different symbols) depicts one similarity layer LD1–12. (E) p-values (mean±s.e.m) of the power-law fit for each cohort. (F) One-sided and two-sided p-values (mean±s.e.m) for the discrimination between the exponential (one-sided p-value=1, two-sided p-value=0) and the power-law fits. (G) Graphics of power-law ($\gamma = 2.2$), exponential and random network models of 100 nodes. (H) Prediction accuracy (Q^2, leave-one-out cross-validated R^2, mean±s.e.m) of selected distant similarity layers LD_{4-12} from LD_1. (I) Prediction accuracy (Q^2, mean±s.e.m) of all similarity layers (LD_{2-12}) from LD_1.

3.6 Antibody repertoires are highly robust systems

We hypothesized that the reproducible architecture of antibody repertoires may have evolved to be robust to fluctuations in clonal composition. It is known that antibody repertoires are very dynamic systems characterized by a high turnover rate [164, 165, 110, 96]. Therefore, we investigated the robustness of antibody repertoire architecture to clonal removal (deletion).

It is has been recently established that individual repertoires have public clones, which are defined as identical clones present in multiple individuals [56]. While mostly distinct, antibody repertoires possessed a fraction of public clones (15–26% along B-cell development, Fig. 3.4 A). Given their regular presence, we determined if public clones were essential to the maintenance of antibody repertoire architecture. We found that the highest authority clones were public (Fig. 3.6 A) and up to 74% of private clones (specific to an individual) were connected to at least one public clone (Fig. 3.3 I). To quantify the extent to which public clones maintain the architecture of antibody repertoires, we tested the effect of removing public clones on CDR3 degree distributions. In pBC and nBC, removal of all public clones transformed their network structure from exponential to power-law; in contrast, removal of public clones led to no change in PC network structure, Figure 3B). To assess if such a structural shift was specifically due to the deletion of public clones, we removed (repeatedly) random subsets of clones representing a similar fraction of public clones. The structure of antibody repertoires was robust along B-cell stages at up to 50% removal of random clones. The same structural shift in repertoire structure caused by the deletion of public clones could only be replicated by removing 90% of random clones (Fig. 3.6 C). Therefore, public clones represent pillars that are critical for maintaining the architecture of an antibody repertoire, and the robustness of this architecture suggests a functional immunity is preserved even after extensive (random) loss of antibody clones (or B cells).

3.7 Antibody repertoires are evolutionary redundant

Redundancy is a hallmark of robust systems; for example, redundancy in genes with the same function is the main mechanism of robustness against mutations in genetic networks [166]. To investigate the extent of redundancy within antibody networks, we examined whether their architecture at the base similarity layer (LD_1) was manifested in higher order similarity layers (LD>1). Differences greater than 1 a.a. between antibody sequences could represent the potential personal scenarios of antibody repertoire evolution, a result of successful survival through selective processes. Specifically, if a clone connected to many other clones in the LD_1 similarity layer mutates into a similar clone at a specific a.a. position, this potential clone will be connected to many clones in the LD_2 similarity layer. Thus, higher order similarity layers can serve as surrogates for the evolution of potential antibody repertoires from antigen-inexperienced B-cell populations.

To quantify the extent of redundancy across similarity layers, we calculated the prediction accuracy of LD_1 versus similarity layers LD_{2-12} using a leave-one-out cross-validation approach (Fig. 3.6 D, Fig. 3.5 H and I). Specifically, quantitative redundancy was low in PC ($LD_1 \rightarrow LD_{2-3}$ prediction accuracy was 28% on average); however, LD_1 of pBC and nBC predicted CDR3 degree profiles of proximal similarity layers LD_{2-3} with 80% accuracy (Fig. 3.6 D and E), thereby indicating a high redundancy in antibody repertoire architecture. This high redundancy is explained by the structure of the antibody networks (Fig. 3.3 E-H). Although the distance between proximal similarity layers (LD_1 to LD_3) seems small ($1-3$ a.a. CDR3 sequence differences), it represents $\approx 20\%$ of potential change in clonal a.a. sequence (99% of CDR3 clones are 4-20 a.a. long), which is in the range of highly mutated antibodies (e.g., broadly neutralizing HIV-specific [167]). Therefore, redundancy in the antigen-inexperienced repertoire is maintained throughout a large sequence space and provides details on the pre-programmed evolvability [168, 169] of antibody responses.

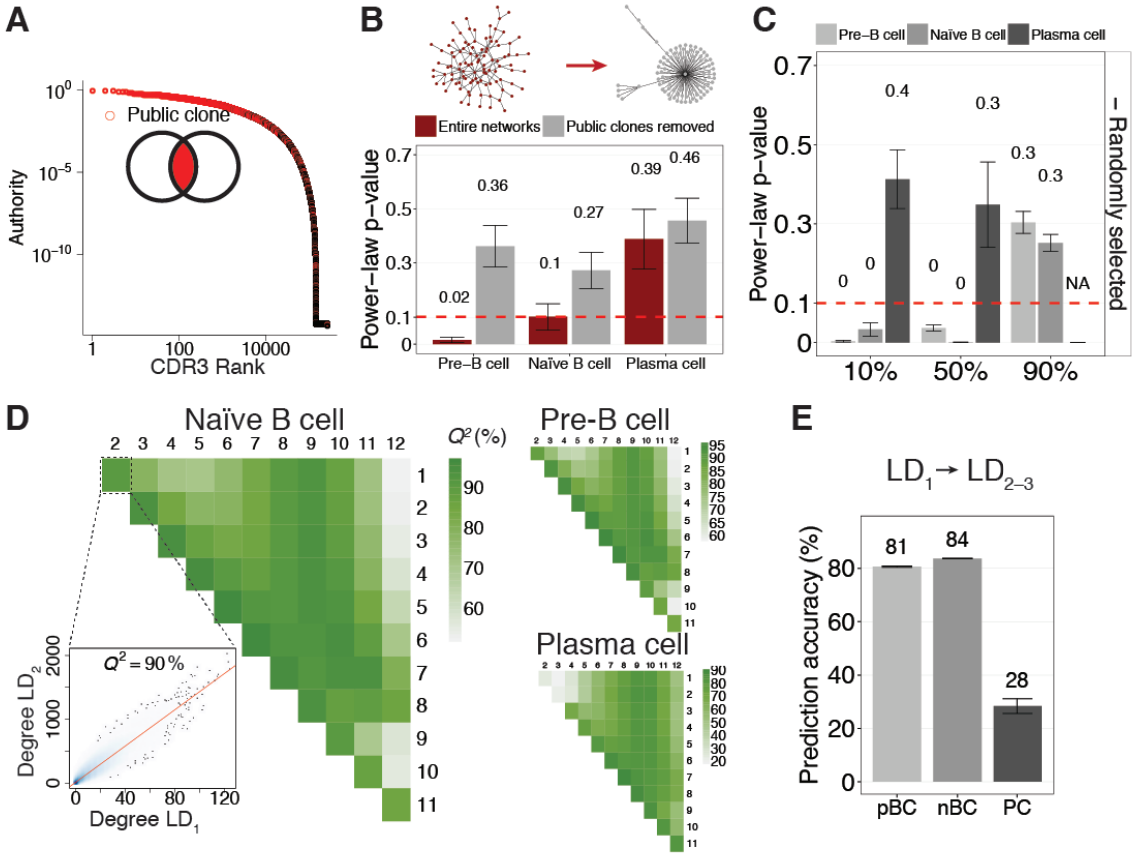


Fig. 3.6 The architecture of antibody repertoires is robust and redundant. (A) CDR3 clones of an exemplary naïve B-cell repertoire have been ordered from increasing to decreasing frequency (CDR3 rank). Public clones are color-coded in red. (B) Bootstrapped p-values of the power-law fit are shown for complete antibody repertoires and after removing public clones. Power law is a good fit to degree distributions for p-values above the dashed red line (p-value = 0.1). Examples of exponential (red) and power-law (grey) networks are shown on the top panel. (C) CDR3 clones were removed randomly at 10%, 50% and 90% from each original repertoire (20 times) and the power-law distribution was fit to the cumulative degree distributions of the remaining CDR3 clones. A p-value=0.1 is indicated as a red dashed line. In PC samples a fit was not feasible after removal of 90% of CDR3 clones (NA). (D) Heatmaps indicate the mean prediction accuracy (Q^2, leave-one-out cross-validated R^2) of similarity layer LD1 versus similarity layers LD_{2-12}. The scatterplot shows Q^2 for LD_1 vs. LD_2 for each CDR3 clone. (E) Prediction accuracy (Q^2) for LD_1 vs. LD_2 and LD_2. Barplots show mean±s.e.m.

3.8 Methods

3.9 Dataset

The dataset analyzed was produced as described in Greiff et al. [1]. Briefly, murine B-cell populations of pre-B cells (pBC, IgM, bone marrow), naïve follicular B cells (nBC, IgM, spleen), and memory plasma cells (PC, IgG, bone marrow) were sorted using fluorescence-activated cell sorting (FACS) from C57BL/6J mice unimmunized (n=5) or prime-boost immunized with alum-precipitated antigens: nitrophenylacetyl-conjugated hen egg lysozyme (NP-HEL, n=5), ovalbumin (OVA, n=5) or Hepatitis B virus surface antigen (HBsAg, n=4). Following total RNA extraction, full-length antibody variable heavy chain (VDJ) libraries were generated by a two-step PCR process, as described previously [19]. Libraries were sequenced using the Illumina MiSeq (2x300bp) platform. Mean Phred-scores of raw data were ≥ 30. Approximate paired-end reads (full-length VDJ) were: pBC $5 * 10^6$ reads, nBC $10 * 10^6$ reads and PC $4 * 10^6$ reads.

3.10 Data preprocessing and CDR3 clonal analysis

Antibody sequences have been preprocessed and VDJ annotated with MiXCR [17] and further filtered to retain only those sequences that had CDR3 length ≥ 4 a.a. and occurred more than once in each CDR3 repertoire data set (Fig. 3.2A). Clones were defined by 100% a.a. sequence identity of CDR3 regions. CDR3 regions were defined by MiXCR according to the nomenclature of the Immunogenetics Database (IMGT) [170].

3.11 Network construction

To construct networks (graphs), a sparse triangle matrix of pairwise Levenshtein distances (LD) between CDR3s must first be computed. For small samples (up to 100,000 unique CDR3 sequences) such a calculation is relatively quick on a single computer. However, due to the N^2 complexity of required calculations, computing the pairwise matrix for samples of >100,000 unique CDR3 sequences becomes prohibitively expensive. To perform these computations, we developed software that utilizes the Apache Spark (http://spark.apache.org/) distributed computing framework to partition the work among a cluster of many machines (Fig. 3.2B). We chose specifically Apache Spark because i) its deployment is very flexible with regard to underlying computing infrastructure and ii) for similarity layers LD>1, the networks become extremely large and difficult to process. In these cases, our package can take advantage of the Spark GraphFrames distributed graph library [171], which allows scaling to even larger samples with millions of sequences (Fig. 3.2C). With this approach we were able to compute the distance matrices for large samples (>100,000 unique CDR3 sequences) within minutes (Fig. 3.2B,C).

In addition to the computational complexity inherent in creating the distance matrix, the construction of networks for large LD is very computationally and time-wise costly. We therefore avoided constructing networks altogether for calculating the node degrees and instead used a map-reduce distributed algorithm. For practical purposes, the construction of small networks was performed using the Networkx library [172]. For generating and outputting the largest graphs to disk in common network formats, we used the efficient graph-tool library (https://graph-tool.skewed.de/ [173]). For manipulating and analyzing the largest networks, our software package took advantage of the Spark GraphFrames distributed graph library [171].

The software was developed in python (https://www.python.org/) using the Numpy / Scipy [174] scientific libraries for matrix and array manipulation and Apache Spark [154] as the distributed backend. Our software package for antibody repertoires imNet is described in Chapter 4 and includes tutorials and demos, including scripts to set up the distributed computation environment on commonly-used compute cluster infrastructure. The results shown in this work were obtained using 1–625 cores of the Euler parallel-computing cluster operated by ETH Zürich.

3.12 Degree distribution fits

Degrees (number of similar CDR3 sequences to a specific CDR3 sequence) were calculated for each of the similarity layers LD_{1-12} for each CDR3 sequence in each sample. CDR3 with zero degrees that were not similar to any other CDR3 in the network were excluded in order to fit degree distributions. The power-law, exponential and Poisson distributions were fitted to the empirical degree distributions of the networks, constructed as described in Network construction, by estimating $x_{\min}$ (estimated lower degree threshold by minimizing the Kolmogorov-Smirnoff statistic [175]) and optimizing model parameters using the poweRlaw [176] package. We first discriminated if the power-law distribution could describe the best fit to the degree distribution by bootstrapping 100 times the power-law p-value obtained from each sample after estimating $x_{\min}$. Following the approach described by Virkar and Clauset [177], a p-value ≥ 0.1 indicated that the power-law distribution described the degree distribution (Fig. 3.2A). To determine the degree distribution in cases where the power law was not the best distribution fit (p-value < 0.1), we compared the exponential and the Poisson fits. Two-sided p-value ≈ 0 indicated that the fitted models could be discriminated, and one-sided p-value ≈ 1 indicated that the first (for example exponential) model was the best fit for the data [176].

3.13 Robustness of the architecture of antibody repertoire networks

Public clones were defined as clones shared among subjects in a cohort (Fig. 3.4). In order to assess the robustness of the architecture of antibody repertoire networks we removed public clones from each sample-repertoire. As controls, we performed repeated removal (20 times) of randomly selected clones in the size of public clones. The p-values for the power-law fit were calculated after 100x bootstrapping for each repertoire; one-sided and two-sided p-values were used for the comparison between the exponential and the Poisson fits (see Section 3.12).

3.14 Network analysis

Drawing from network theory [178], we translated the concepts of network analysis [?] to antibody repertoires. An antibody repertoire network is an undirected *graph* $G = (V, E)$ described as a set of *nodes* (CDR3 vertices, V) together with a set of *connections* (similarity edges, E), representing the adjacency matrix of pairwise Levenshtein distances (LD) between CDR3 a.a. sequences A=$\begin{bmatrix} 0 & \dots & LD_{1n} \\ \vdots & \ddots & \vdots \\ LD_{n1} & \dots & LD_{nn} \end{bmatrix}$.

In the context of antibody repertoires, we let N = |V| and L = |E|. The order of a graph N represents the number of its unique CDR3 clones (nodes). The size of a graph L is the number of its CDR3 similarity connections (edges). The degree k, that represents the edges connected to a node, describes the count of all similar CDR3 clones to a CDR3 based on LD. Because the degree indicates how active a node is, it could be interpreted as a measure of how central a CDR3 clone is in the antibody repertoire network. In simpler terms, it quantifies the number of CDR3 clones that are similar to a certain CDR3, and thus the potential development or the evolutionary routes to this CDR3.

The *average degree* $\langle k\rangle \equiv \frac{\sum_{i=0}^{n} k_i}{N} = \frac{2L}{N}$ is the average number of similar CDR3 clones. The *degree distribution* $P(k) = N_k/N$, defined as the fraction of nodes with degree $k(N_k)$ in total nodes, represents the fraction of CDR3 clones that have the same number of similar CDR3s. The *cumulative degree distribution* describes the fraction of nodes with degree greater than or equal to $P_k = \sum_{k'=k}^{\infty} p_{k'}$. In Erdös–Rényi (ER) random graph models, degrees follow a Poisson distribution in the limit of large numbers of nodes, while degree distributions have an exponential tail in exponential networks [179].

Global characterization [**?**] described the network as a whole, such as degree distribution, centralization, largest component, diameter, clustering coefficient, assortativity and coreness. The *centralization* analysis indicates if the network is homogeneous (clones are connected in the same way) or is centered around certain nodes (highly connected clonal regions compared to less connected regions in the same network). The *largest component* is the largest cluster of connected CDR3 clones. The *diameter* (d) is the maximum distance (shortest path between two nodes) between any pair of CDR3 sequences. The *clustering coefficient* (C) represents the probability that neighbors of a node are also connected, which translates in antibody repertoires as the probability that CDR3 clones similar to a specific CDR3 are also similar among one another. Network *density* (D) is the ratio of the number of edges (CDR3 similarities) and the number of all possible edges in the network. The *assortativity* coefficient (r) indicates if nodes in a network connect to nodes with similar characteristics. It is positive if nodes tend to connect to nodes that are similar to them (i.e. highly connected CDR3 sequences are similar and connect to highly connected CDR3 sequences), and negative otherwise. *Coreness* is a measure of the network's cohesion and allows one to understand the global network structure and is useful in comparing complex networks by analyzing the subsets of CDR3-cores that form layers in the antibody repertoire. K-core decomposition is a process that is performed by iteratively removing shells of all vertices of degree less than k ($k<k_{\max}$) leaving the k-cores of a network (its connected component). The k-core of a graph is the maximal subgraph in which each node has at least degree k. We have computed the maximal k-core of antibody repertoire networks (the innermost core, $k_{\max}$) and the core distribution along k degrees.

Clonal (local) characterization of antibody repertoires was performed by analyzing local properties of the networks [**?**]. The importance of CDR3 clones was measured by calculating the authority [180], eigenvector [161] and PageRank [162] scores of each node in repertoire networks. In particular, the *authority* (a) of nodes is defined as the principal eigenvector of the transpose matrix t($\mathbf{A}$)*$\mathbf{A}$, where $\mathbf{A}$ is the adjacency matrix of the network. Eigenvector centrality indicates the centrality of a CDR3 clone, not only dependent on the number of similar CDR3 (number of degree, connections) but also on the quality of those connections: CDR3-nodes with high eigenvector values are connected to many other nodes which are, in turn, connected to many others (and so on). *PageRank* measures the importance of the similarity between two CDR3 clones within the network extending beyond the approximation of a CDR3 importance or quality. *Closeness* (centrality [160]) (c) was calculated to measure how many steps were required to access every other CDR3 from a given CDR3 clone in antibody repertoire networks. We calculated the normalized closeness by multiplying the raw closeness by n-1, where n was the number of nodes in the network. Clique analysis identified maximally-connected subgraphs (a subset of nodes) in which every CDR3 was similar to every other CDR3 sequence and the largest clique was the maximal completed subgraph which had more nodes than any other clique in the network. The node *betweenness* (b) is the number of geodesics (shortest paths) going through a node and indicates the "bridge" function of a CDR3 sequence. Network properties were calculated using the `igraph` [107] R package.

3.15 Quantifying the predictive performance (Q^2) of linear regression models

The predictive performance (Q^2) of each linear regression model ($Y = X\beta + \varepsilon$) was calculated using leave-one-out cross-validation (LOOCV): $Q^2 = (1 - \frac{PRESS}{TSS}) * 100$ where PRESS is the predictive error sum of squares $\sum_{j=1}^{n}\left(Y_j\hat{Y}_{[j]}\right)^2$ with $\hat{Y}_{[j]}$ denoting the prediction of the model when the j-th case is deleted from the training set and TSS is the total sum of squares $\sum_{i=1}^{n}\left(Y_j\bar{Y}\right)^2$ [181]. $\boldsymbol{X}$ and $\boldsymbol{Y}$ are CDR3 degree vectors of repertoires at each LD_{1-12}. LOOCV was performed using the forecast R package [182]. Cross-validation was used because, in contrast to regular regression analysis, it enables the quantification of the predictive performance of each regression model.

3.16 Simulated networks

Networks (nodes V=$10^2 - 10^5$) were simulated with the ER, exponential and power-law models using base R [183] and igraph [107]. Random networks were simulated according to the ER model, exponential networks were simulated setting a probability of a connection between two nodes p=0.5 and scale-free networks were simulated using the Barabási-Albert model [184].

3.17 Graphics

Graphic representations were produced using base R ([183]) and the ggplot2 R package [185]. Heatmaps were produced using the NMF package [186]. Networks and network clusters visualization were performed using igraph [107] employing the Fruchterman–Reingold force-directed and Kamada–Kawai layout algorithms. Large-scale networks (Fig. 3.1A) were visualized using Gephi (version 0.9.1) [109]; node size was scaled 10–100 proportional to the degree of a node and a blue to grey color gradient was applied to nodes from high to low degrees.

3.18 Statistical significance

Statistical significance was tested using the Wilcoxon rank-sum test. Results were considered significant for p<0.05.

3.19 Data and software availability

Antibody repertoire sequencing data analyzed is available with ArrayExpress accession number: E-MTAB-5349. Software is available at https://github.com/rokroskar/imnet.

Chapter 4

imNet: software for the generation and analysis of large-scale sequence networks

Due to advances in high-throughput omics technologies, the generation of large-scale datasets of biological molecules (DNA, RNA, proteins, metabolites) is now routine. Network analysis on biomolecular systems is crucial for a systems-level understanding and identification of biological functions [187].

Specifically, the structure and function of proteins is directly related to their sequence [188]. For example, adaptive immunity is driven by immune repertoires, where different antibodies and T cell receptors (TCRs) confer recognition and protection against the enormous variety of pathogens. Studying the sequence similarities of antibodies and TCRs through large-scale network analysis would enable the detection of molecules with similar function, which could be used for diagnostic or medical applications.

However, most biological networks that have been constructed thus far embody thousands of biomolecules, largely undersampling the millions present in an in vivo system. Constructing or analyzing large-scale networks remains a challenge because of the high-dimensionality space: e.g., there are 10^5 peptides that lead to a space of 10^{10} potential similarity relations in the human proteome [189].

Here we report imNet, an open-source software for the comprehensive generation and rapid analysis of large-scale sequence-similarity networks (Fig. 4.1). We implemented a parallelized algorithm for generating large-scale networks from datasets in the range of millions of sequences, not attainable previously due to memory constraints and computation time. Our implementation utilizes the Apache Spark [154] distributed computing framework to partition the work among a cluster of machines (Fig. 4.1).

imNet is primarily designed to handle the scale and depth of the enormous diversity present in immune repertoires. However, any sequence-based network can be constructed (e.g., entire human proteome [189]). If a cluster is not available, imNet can also perform the calculations on a single computer for smaller samples ($< 10^4$). imNet takes as input sequences in text format and outputs either the network in the GraphML format or a degree vector obtained using a map-reduce distributed algorithm. The software supports the simulation of *in silico* sequences and networks.

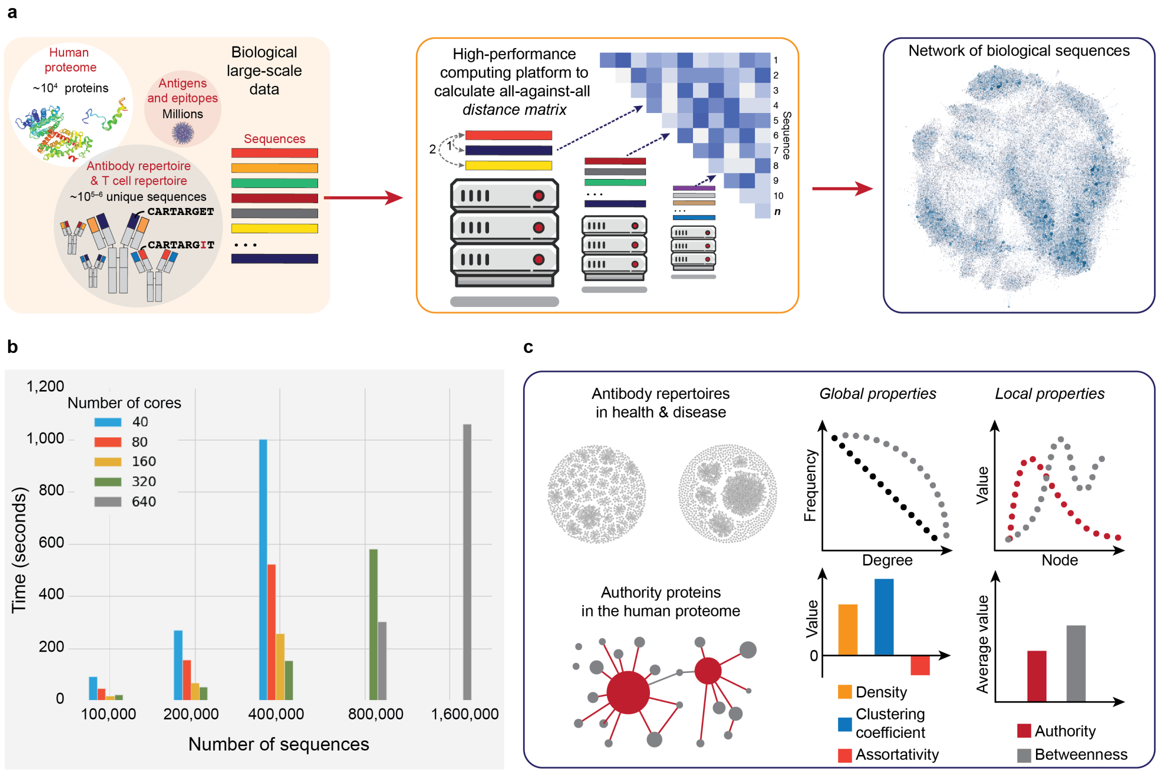


Fig. 4.1 imNet data analysis. (A) Construction of large-scale distance matrices from biological sequences using a high-performance parallel computing approach to partition the work among a cluster of machines to generate networks. (B) Computational time-scaling of the performance of imNet according to the number of sequences (strings) and number of parallel cores (machines). (C) Exemplary applications and downstream analysis of networks.

The approach that imNet uses to analyze sequences is highly efficient; in a recent study, we constructed immunology networks from $> 10^5$ antibody sequences [31]. The construction of large-scale networks from 10^6 sequences requires $\approx 10^{12}$ edit distance calculations resulting in ≈ 300 hours of computation on a modern single-core machine, whereas imNet performed 1,000 times faster on 600 cores (Fig. 4.1). In addition to rapidly constructing large-scale networks, the software computes network degree distributions and provides a framework for implementing subsequent network analyses such as density, clustering coefficient, assortativity, authority and betweenness (Fig. 4.1).

We demonstrated the application of imNet in answering fundamental questions in antibody repertoire development [31]. In the future imNet may be used for a variety of applications, such as the design of synthetic immune repertoires or the analysis of large sequence databases (e.g., cancer genomes).

imNet has been developed as a python library and the source code is available for download: https://github.com/rokroskar/imnet

4.1 imNet User Manual

4.1.1 System requirements

imNet is written for Python 2.7. It uses the Apache Spark (http://spark.apache.org) framework for distributed computations.

4.1.2 Source

The source code for the latest version of imNet can be found on GitHub: https://github.com/rokroskar/imnet

4.1.3 Installation

1. **Using `pip`**

   The simplest way to install imNet is with `pip`:

   ```
   $ pip install imnet
   ```

   In addition to installing the imNet python library and its dependencies, this will also install the `imnet-analyze` script into your python bin directory.

2. **From source**

   Make sure to have installed all the dependencies – see below.

   Clone the repository and install:

```
$ git clone https://github.com/rokroskar/imnet.git
$ cd imnet
$ python setup.py install
```

If you make changes to the cython code, you will need cython and a usable C compiler.

4.1.4 Dependencies

You only need to install dependencies separately if you are installing imNet from source or want to develop. If you just want to run imNet, install it with `pip` (see above) that will install the dependencies automatically.

The following basic python libraries needed by imNet are installable via `pip` or `conda`:

- `click`
- `findspark`
- `python-Levenshtein`
- `scipy`
- `networkx`
- `pandas`
- `cython (optional)`

If your goal is to analyze large samples (> 10,000 strings), distributing the computation is strongly advised. imNet currently uses the Apache Spark (http://spark.apache.org) distributed computation framework. We won't go into the details of installing and running spark here; you can download it and unpack the archive at any location. The minimum requirement is to set the `SPARK_HOME` environment variable to point to the directory where you unpacked spark, e.g.

```
$ export SPARK_HOME=/path/to/spark
```

If you are running spark on a cloud resource, please refer to the official spark documentation for instructions on how to start up a spark cluster. To allow imNet to run via spark you will need to provide the spark URL of the ‘spark-master‘.

If your resource is an academic HPC (high-performance computing) cluster, we recommend that you use `sparkhpc` (https://github.com/rokroskar/sparkhpc) for managing spark clusters. `sparkhpc` greatly simplifies spawning and managing spark clusters.

4.1.5 Usage

Basic usage

imNet takes as input a list of strings either supplied directly by the user or read from a file.

1. **Command-line**

   Refer to the command-line help for usage, e.g.

   ```
   $ imnet-analyze --help
   ```

   Usage:

   ```
   $ imnet-analyze [OPTIONS] COMMAND [ARGS]
   ```

   Options:

   `--spark-config TEXT` Spark configuration directory

   `--spark-master TEXT` Spark master

   `--kind [graph|degrees|all]` Which kind of output to produce

   `--outdir TEXT` Output directory

   `--min-ld INTEGER` Minimum Levenshtein distance

   `--max-ld INTEGER` Maximum Levenshtein distance

   `--sc-cutoff INTEGER` For a number of strings below this cutoff, Spark will not be used

   `--spark / --no-spark` Whether to use Spark or not

   `--help` Show this message and exit

   Commands:

   `benchmark` Run a series of benchmarks for graph

   `directory` Process a directory of string files (e.g., CDR3 sequences) `file` Process an individual file with strings

`random` Run analysis on a randomly generated set of strings

```
$ imnet-analyze random --help
```

Usage:

`imnet-analyze random [OPTIONS]` Run analysis on a randomly generated set of strings

2. **Run analysis on a randomly generated set of strings for testing**

   Options:

   `--nstrings INTEGER` Number of strings to generate

   `--min-length INTEGER` Minimum number of characters per string

   `--max-length INTEGER` Maximum number of characters per string

   `--help` Show this message and exit

3. **Tutorial**

   For a tutorial on using the imNet python library, look at the example notebook: https://github.com/rokroskar/imnet/blob/master/notebooks/example_workflow.ipynb

4. **Sample input and output**

   Several inputs and the corresponding network outputs from running imNet on sample datasets include the human proteome, B and T cell receptors (human and mice), epitopes and peptide sequences.

4.2 imNet software description, algorithms and pipeline overview

The primary function of imNet is the construction of the upper triangle of the symmetric distance matrix given a set of unique strings. This calculation can either be used to construct a network graph, or alternatively, output the degree vector for each node specifying the degree of connectedness at each separation distance.

4.2.1 Generation of the distance matrix

To calculate the distance between two strings, we use the highly-efficient `python-Levenshtein` package (https://pypi.python.org/pypi/python-Levenshtein/). The calculation of a single pair distance takes a few times 10^{-6} seconds, leading to a few hundred hours of single-core computation time for a sample of 10^6 sequences.

To reduce the time-to-solution, we implemented an algorithm to compute the distance matrix in parallel, using the Apache Spark distributed computing framework to handle the details of the parallelization. The reduction in time-to-solution is approximately linear with the number of cores used for computation (see Fig. 4.1b).

In addition to implementing the distance matrix calculation, we also implement a map-reduce algorithm for efficiently calculating the degree distribution directly from the distributed matrix data.

The all-to-all distance calculation is inherently an $O(N^2)$ calculation. If the requested maximum Levenshtein distance is small (< 3) we make use of a Vantage-point tree algorithm to reduce the maximum number of string comparisons that need to be made. In the best case, this algorithm has complexity of $O(nlogn)$, which gives a way to scale the analysis to even larger samples in the future.

4.2.2 Network analysis

Global and local network parameters can be analyzed from the starting network (Fig. 4.1c). Analysis of the network clustering coefficient, density, assortativity, authority and betweenness can be performed using `igraph` functions `density()`, `transitivity_undirected()`, `assortativity()`, `authority_score()`, `betweenness()` (http://igraph.org/python/doc/igraph.GraphBase-class.html [107]).

4.2.3 Data and benchmarks

For benchmarking purposes, we generated random strings of 4–20 amino acids from a normal distribution. We generated networks from 1,600,000 randomly sampled strings of length 4–20 amino acids (a.a.) and 12–60 nucleotides.

We used input sequences from different protein database sources and high-throughput sequences of antibody and T cell repertoires from human and mice samples.

Specifically, the exemplary network of mouse antibody repertoire comprised 535,061 sequences [31], while the network from the mouse T cell repertoire was constructed from 36,889 sequences [43]. The network of human antibody repertoire comprised 6,348,502 sequences [190] and T cell repertoire network was generated from 256,054 sequences

[46]. We constructed an epitope network of 14,937 sequences from the immune epitope database [191].

Furthermore, we constructed networks of the entire human proteome from 293,700 peptide sequences [189] and from the entire UniProt [192] database using the total of 553,231 protein sequences manually annotated.

Chapter 5

Detection of broadly neutralizing antibody sequence signatures in HIV-1

Motivation: A small fraction of HIV-1 infected individuals develops broadly neutralizing antibodies (bNAbs) that respond to a multitude of HIV-1 strains. However, bNAbs are very different in their sequences, and rare within and across individuals. High-throughput sequencing of antibody repertoires (Ig-seq) enables a high-resolution and quantitative description of humoral immune diversity, reporting full-length antibody sequences. Therefore, Ig-seq coupled to advanced computational tools can be used to determine if sequence-associated signatures of bNAbs exist in HIV-1 patients and identify potential bNAb-like sequences.

Results: Here we report the detection of sequence signatures in antibody repertoires from HIV-1 infected individuals that have developed bNAbs compared to HIV-1 bNAb-negative and uninfected individuals. We have compiled a database of bNAbs sequences that we have used as reference to inquire sequence characteristics and to train a support vector machine for features of bNAbs sequences. We investigated somatic hypermutations, complementarity determining region 3 (CDR3) length, germline frequencies, clonal frequency distributions, CDR3 similarity relations within repertoires and ultimately, sequence identity to database bNAb sequences in order to uncover unidimensional repertoire characteristics associated with bNAb status. We report that we can discriminate with 95% accuracy between bNAbs and sequences of HIV-1 bNAb-HTS data using machine learning. This technique allowed to capture bNAbs signatures that were undetectable at the repertoire level when considering low dimensional data

analysis like the average SHM or CDR3 length. Our results offer a global systems characterization of HIV-1 repertoires compared to uninfected, predict bNAbs-like sequence features using machine learning, and open the way to the advancement of computational methods for the de novo prediction of bNAbs sequences from HIV-1 bNAb+ repertoires.

Availability and Implementation: https://github.com/enkelejdamiho/bNAbs-DB

5.1 Introduction

The enormous repertoire of B-cell receptors (10^9 BCR, antibodies [37] protects against various pathogens. A small fraction (1%) of individuals chronically infected with human immunodeficiency virus-1 (HIV-1) develop broadly neutralizing antibodies (bNAbs) against the virus [193, 194], in spite of its rapid mutation rate and high genomic variability [195]. Clonally expanded B cells undergo rounds of affinity maturation to respond to the mutating virus [92, 149, 168] and secrete bNAbs that vary in breadth and potency against the multitude of HIV-1 strains. bNAbs suppress viraemia in HIV-1 infection, showing remarkable potential as therapeutics [196, 197] and guides for vaccine design [198, 199, 149].

The enormous clonal composition of antibody repertoires can now be captured at high resolution through high-throughput sequencing [27, 5]. Advances in single B-cell cloning and high-throughput sequencing methods have sustained the characterization of monoclonal antibodies and BCR repertoires [140, 93, 92], supporting the identification of more potent and broader neutralizing antibodies directly from HIV-1 infected individuals [200–202].

However, bNAbs are rare within and across patients; only around 30 clones of bNAbs ($\approx$ 90 somatic variants), have been identified [40, 98]. Thus, although bNAb have been reported to display some specific characteristics (e.g., high SHM and long CDR3 [203–205, 149]), it is unknown if these features are reflected at the repertoire level. It is thus unclear if bNAb-like repertoire characteristics could allow for the detection and prediction of bNAb status from high-throughput sequencing data. To what extent unidimensional data like SHM and CDR3 length capture the bNAb status-specific information of antibody repertoires?

We characterized sequenced antibody repertoires from HIV-1 infected individuals that have developed bNAbs and two controls, HIV-1 infected that have not developed bNAbs (HIV-1 bNAb-) and uninfected individuals, in order to assess if reference-bNAb sequence characteristics at the repertoire level could reveal bNAb status. We set to identify repertoire signatures associated with the presence of bNAbs. This aim sustains the progress to a potentially successive de novo identification of bNAbs sequences from HTS data.

In order to determine if global analysis of HIV-1 antibody repertoires revealed the presence of sequence characteristics indicative of the emergence of bNAbs, we constructed a reference database of bNAbs sequences (bNAbs-DB). bNAbs are characterized by a high number of SHM, elongated CDR3 sequences, and no specific V-gene germline (although, a few bNAbs displayed restricted IGHV genes [206]). We investigated the SHM and CDR3 length according CDR3 frequency. We then set to analyze the clonal frequency distribution (evenness profiles) and the structure of CDR3 similarity relations (networks) in order to investigate clonal-based signatures of bNAbs in repertoires. In order to investigate potential germline signatures, we analyzed the V- and J- frequencies. Ultimately, we used sequence identity to bNAbs-DB and machine learning to detect bNAb status in HIV-1 bNAb+ cohorts. We trained SVM with bNAbs-DB sequences and tested it in HIV-1 bNAbs- individuals; we then used this model to detect bNAb status by predicting bNAb-like CDR3 sequences in HIV-1 bNAb+ compared to uninfected individuals.

5.2 Methods

5.2.1 Datasets

We compiled three high-throughput sequencing datasets of antibody repertoires from HIV-1 bNAb+, HIV-1 bNAb- and HIV-1-uninfected individuals (total number of individuals n=29, total number of samples n=83). Primers covered all V genes. Read statistics after preprocessing, as described below, are indicated in Fig. 5.4.

Dataset 1

B-cell repertoire from HIV-1 infected individuals with broadly neutralizing antibodies (n=9). High-throughput sequencing data of heavy and light chains from African donor

17 of the IAVI Protocol G cohort (n=1, IAVI donor 17, PGT121 class of antibodies [207], resulting from sequencing the 5'-RACE PCR of B-cell transcripts with Ion Torrent Personal Genome Machine (PGM), was provided by Prof. Jiang Zhu [208]. Data from eight additional donors characterized as bNAb producers (n=8) that had broad neutralizing plasma antibodies by previously described criteria from the CHAVI chronic HIV-1 infection cohort [209] and/or had bNAbs isolated from blood B cells as described [210, 211] as described by Kepler et al., were provided by Prof. Thomas Kepler [212].

Dataset 2

HIV-1 infected human donors (n=13). High-throughput sequencing data of heavy chains from peripheral blood mononuclear cells (PBMCs) of HIV-1 infected donors (n=8, part of the SPARTAC trial) collected over 2 years, five of whom received anti-retroviral therapy during the first half of the study period, was downloaded from Hoehn and colleagues [95], European Nucleotide Archive accession number ERP000572. Data was downloaded for patients P1 weeks 0, 4, 16, 24, 52, 72, 120 (7 time points), P2 weeks 0, 4, 12, 16, 24, 48, 60, 108 (8 time points), P3 weeks 0, 4, 12, 16, 24, 52, 60, 108 (8 time points), P4 weeks 4, 12, 16, 24, 52, 60, 108 (7 timepoints), P5 weeks 4, 12, 16, 24, 52, 60, 108 (7 time points), P6 weeks 4, 12, 16, 24, 52, 60, 108 (7 time points), P7 weeks 0, 4, 12, 16, 24, 52, 60, 108 (8 time points), P8 weeks 0, 4, 16, 24, 52, 60, 108 (7 time points). Libraries were sequenced by 150 bp paired-ended MiSeq (Illumina) and read depth of full data had a median of 567,936 reads per patient per time point (Hoehn et al., 2015). Additional high-throughput sequencing data from PBMC samples (n=5) of HIV-1 bNAb negative donors was provided by Prof. Thomas Kepler [212].

Dataset 3

B cell repertoire of uninfected human donors (n=7). B-cell heavy and light chains sequenced the 5'-RACE PCR of B-cell transcripts with Ion Torrent Personal Genome Machine from HIV-1-uninfected individuals (n=2) were provided by Prof. Jiang Zhu [208]. Sequences of antibody repertoires from five HIV-1-uninfected individuals were published by Bashford-Rogers and colleagues [94]. Heavy chains (IGH) were amplified from PBMCs isolated from healthy volunteers (n=5). High-throughput sequencing was performed on the VH gene using Roche 454. The raw data was downloaded from European Nucleotide Archive accession number ERP002120.

5.2.2 Broadly neutralizing antibody database against HIV-1

We constructed a bNAbs database (bNAbs-DB, https://github.com/enkelejdamiho/bNAbs-DB) by collecting sequences of bNAbs from on-line databases like bNAber [213] (http://www.bnaber.org/), GenBank, Los Alamos HIV Molecular Immunology (http://www.hiv.lanl.gov/), Abysis (http://bioinf.org.uk/abysis/), scientific literature and personal communications. The sequences were IMGT annotated [214]. The database encompasses 90 bNAbs (Tab. 5.2), which belong to 10 families according to their binding site (Tab. 5.1). To avoid bias towards bNAbs with numerous variants which most likely have a similar number of SHM and CDR3 lengths, we selected the most broadly neutralizing and potent representatives within each group of similar monoclonal antibodies in each clonal family (indicated in bold in Tab. 5.1).

Binding site	**bNAbs family**
gp120 adjacent to CD4BS	**HJ16**
gp120 CD4BS	**12A12**, 12A21
	3BNC117, 3BNC62, 3BNC60, 3BNC55
	8ANC131, 8ANC134, **8ANC195**
	b12
	CH98
	CH103, CH104, CH106
	VRC-CH30, VRC-CH31, VRC-CH32, **VRC-CH33**, VRC-CH34
	VRC-PG04, VRC-PG04b
	VRC-PG20, VRC01, VRC02, VRC03, VRC06, VRC06b, **NIH45-46**, VRC23, VRC23b
	1NC9
gp120 V1-V2	**CH01**, CH02, CH03, CH04
	PG9, PG16
	PGC14
	VRC26.01, VRC26.02, **VRC26.03**, VRC26.04, VRC26.05, VRC26.06, VRC26.07, VRC26.08, VRC26.09, VRC26.10, VRC26.11, VRC26.12
gp120 V1-V2, quaternary structure	PGT141, **PGT142**, PGT143, PGT144, PGT145
gp120 V3	**10-1074**
	2G12
	447-52D
	PGT121, PGT122, PGT123
	PGT125, PGT126, PGT127, **PGT128**, PGT130, PGT131, VRC24
	PGT135, PGT136, PGT137
	HGN194
	2219, **2557**, 2558
gp160	**3BC176**, 3BC315
gp41	**5H_I1-BMV-D5**
gp41 MPER	**10E8**, 7H6
	2F5
	4E10
	Z13
	m66, **m66.6**
gp41 NHR	**HK20**
gp41-gp120 quartenary interface	**PGT151**, PGT152
	35O22

Table 5.1 bNAbs clonal lineages contained in the bNAbs-DB. The most broadly neutralizing and potent representatives within each bNAbs clonal family are indicated in bold.

bNAb	SHM (a.a.)	CDR3 length (a.a.)	V gene	D gene	J gene	Reference
10-1074	20	26	IGHV4-59	IGHD3-3	IGHJ6	Mouquet et al., PNAS, 2012
10E8	27	22	IGHV3-15	IGHD3-3	IGHJ1	Huang et al., Nature, 2012
7H6	27	22	IGHV3-15	IGHD3-3	IGHJ1	Huang et al., Nature, 2012
12A12	34	15	IGHV1-2	IGHD4-17	IGHJ2	Scheid et al., Science, 2011
12A21	31	15	IGHV1-2	IGHD5-12	IGHJ2	Scheid et al., Science, 2011
2F5	14	24	IGHV2-5	ND	IGHJ6	Buchacher et al., AIDS, 1994
2G12	31	16	IGHV3-21	IGHD1-26	IGHJ3	Buchacher et al., AIDS, 1994
3BC176	34	21	IGHV1-2	IGHD5-12	IGHJ3	Klein et al., JEM, 2012
3BC315	24	21	IGHV1-2	IGHD5-12	IGHJ3	Klein et al., JEM, 2012
3BNC117	34	12	IGHV1-2	IGHD6-25	IGHJ2	Scheid et al., Science, 2011
3BNC62	36	12	IGHV1-2	IGHD4-17	IGHJ2	Scheid et al., Science, 2011
3BNC60	38	12	IGHV1-2	IGHD3-3	IGHJ2	Scheid et al., Science, 2011
3BNC55	33	12	IGHV1-2	IGHD6-25	IGHJ2	Scheid et al., Science, 2011
447-52D	10	21	IGHV3-15	ND	IGHJ6	Buchbinder et al., AIDS, 1992
4E10	18	20	IGHV1-69	IGHD6-19	IGHJ4	Buchacher et al., AIDS, 1994
5H_I1-BMV-D5	6	12	IGHV1-69	IGHD1-14	IGHJ4	Miller et al., PNAS, 2005
8ANC131	38	18	IGHV1-46	IGHD3-16	IGHJ6	Scheid et al., Science, 2011
8ANC134	37	18	IGHV1-46	IGHD3-16	IGHJ6	Scheid et al., Science, 2011
8ANC195	41	22	IGHV1-3	IGHD3-3	IGHJ4	Scheid et al., Science, 2011
b12	23	20	IGHV1-18	IGHD1-1	IGHJ6	Burton et al., PNAS, 1991
CH01	28	26	IGHV3-20	IGHD3-10	IGHJ2	Bonsignori et al., J.Virol., 2011
CH02	22	26	IGHV3-20	IGHD3-10	IGHJ2	Bonsignori et al., J.Virol., 2011
CH03	22	26	IGHV3-20	IGHD3-10	IGHJ2	Bonsignori et al., J.Virol., 2011
CH04	23	26	IGHV3-20	IGHD3-10	IGHJ2	Bonsignori et al., J.Virol., 2011
CH98	39	50	IGHV3-30	IGHD1-IR1	IGHJ4	Bonsignori, J. Clin. Invest., 2014
CH103	21	15	IGHV4-31	IGHD6-13	IGHJ1	Liao et al., Nature, 2013
NIH45-46	40	18	IGHV1-2	IGHD1-26	IGHJ2	Scheid et al., Science, 2011
PG9	19	30	IGHV3-33	IGHD1-1	IGHJ6	Walker et al., Science, 2009
PG16	21	30	IGHV3-33	IGHD3-3	IGHJ6	Walker et al., Science, 2009
PGT121	23	26	IGHV4-59	IGHD3-3	IGHJ6	Walker et al., Nature, 2011
PGT122	25	26	IGHV4-61	IGHD3-3	IGHJ6	Walker et al., Nature, 2011
PGT123	28	26	IGHV4-59	IGHD3-3	IGHJ6	Walker et al., Nature, 2011
PGT125	27	21	IGHV4-38	IGHD3-16	IGHJ5	Walker et al., Nature, 2011
PGT126	23	21	IGHV4-38	IGHD3-16	IGHJ5	Walker et al., Nature, 2011
PGT127	25	21	IGHV4-39	IGHD3-16	IGHJ5	Walker et al., Nature, 2011
PGT128	29	21	IGHV4-39	IGHD3-10	IGHJ5	Walker et al., Nature, 2011
PGT130	29	21	IGHV4-39	IGHD3-10	IGHJ5	Walker et al., Nature, 2011
PGT131	29	21	IGHV4-39	IGHD3-10	IGHJ5	Walker et al., Nature, 2011
PGT135	29	20	IGHV4-39	IGHD3-9	IGHJ5	Walker et al., Nature, 2011
PGT136	28	20	IGHV4-39	IGHD2-8	IGHJ5	Walker et al., Nature, 2011
PGT137	33	20	IGHV4-39	IGHD2-15	IGHJ5	Walker et al., Nature, 2011
PGT141	27	34	IGHV1-8	IGHD4-17	IGHJ6	Walker et al., Nature, 2011
PGT142	29	34	IGHV1-8	IGHD4-17	IGHJ6	Walker et al., Nature, 2011
PGT143	27	34	IGHV1-8	IGHD4-17	IGHJ6	Walker et al., Nature, 2011
PGT144	30	34	IGHV1-8	IGHD4-17	IGHJ6	Walker et al., Nature, 2011
PGT145	27	33	IGHV1-8	IGHD4-17	IGHJ3	Walker et al., Nature, 2011
PGT151	27	28	IGHV3-30	IGHD3-10	IGHJ6	Falkowska et al., Immunity, 2014
PGT152	28	28	IGHV3-30	IGHD3-3	IGHJ6	Falkowska et al., Immunity, 2014
VRC-CH30	37	15	IGHV1-2	IGHD3-16	IGHJ4	Wu et al., Science, 2011
VRC-CH31	37	15	IGHV1-2	IGHD5-12	IGHJ1	Wu et al., Science, 2011
VRC-CH32	36	15	IGHV1-2	IGHD6-13	IGHJ4	Wu et al., Science, 2011
VRC-CH33	36	15	IGHV1-2	IGHD3-10	IGHJ4	Wu et al., Science, 2011
VRC-CH34	36	15	IGHV1-2	IGHD6-19	IGHJ4	Wu et al., Science, 2011
VRC-PG04	42	16	IGHV1-2	IGHD2-8	IGHJ2	Wu et al., Science, 2011
VRC-PG04b	42	16	IGHV1-2	IGHD2-15	IGHJ2	Wu et al., Science, 2011
VRC-PG20	36	15	IGHV1-2	IGHD3-10	IGHJ1	Zhou et al., Immunity, 2013
VRC01	41	14	IGHV1-2	IGHD2-21	IGHJ2	Wu et al., Science, 2010
VRC02	39	14	IGHV1-2	IGHD5-12	IGHJ2	Wu et al., Science, 2010
VRC03	39	16	IGHV1-2	IGHD2-21	IGHJ1	Wu et al., Science, 2010
VRC23	30	14	IGHV1-2	IGHD2-15	IGHJ2	Georgiev et al., Science, 2013
VRC23b	32	14	IGHV1-2	IGHD2-21	IGHJ1	Georgiev et al., Science, 2013
VRC24	29	26	IGHV4-4	IGHD3-22	IGHJ5	Georgiev et al., Science, 2013
HK20	14	15	IGHV1-69	IGHD6-6	IGHJ3	Corti et al., PlosONE, 2010
HJ16	45	21	IGHV3-30	IGHD3-3	IGHJ2	Corti et al., PlosONE, 2010
HGN194	14	11	IGHV5-51	IGHD4-17	IGHJ4	Corti et al., PlosONE, 2010
Z13	23	19	IGHV4-59	IGHD2-15	IGHJ6	Zwick et al., J. Virol., 2001
m66	9	23	IGHV5-51	IGHD3-10	IGHJ6	Zhu et al., J. Virol., 2011
m66.6	9	13	IGHV5-51	IGHD3-10	IGHJ4	Zhu et al., J. Virol., 2011
1NC9	36	21	IGHV1-46	IGHD5-24	IGHJ4	Scheid et al., Science, 2011
2219	16	17	IGHV5-51	IGHD4-17	IGHJ3	Gorny et al., J. Virol., 2002
2557	23	17	IGHV5-51	IGHD3-22	IGHJ3	Gorny et al., J. Virol., 2004
2558	16	16	IGHV5-51	IGHD1-26	IGHJ4	Gorny et al., J. Virol., 2004
35O22	34	16	IGHV1-18	IGHD5-24	IGHJ4	Gorny et al., Mol. Immunol., 2009
PGC14	22	15	IGHV1-69	IGHD3-10	IGHJ5	Gorny et al., Mol. Immunol., 2009
VRC06	47	17	IGHV1-2	IGHD2-21	IGHJ5	Gorny et al., Mol. Immunol., 2009
VRC06b	41	17	IGHV1-2	IGHD4-17	IGHJ1	Walker et al., Science, 2009
VRC26.01	16	37	IGHV3-30	IGHD3-9	IGHJ3	Li et al., J. Virol., 2012
VRC26.02	16	37	IGHV3-30	IGHD3-3	IGHJ3	Li et al., J. Virol., 2012
VRC26.03	14	37	IGHV3-30	IGHD3-3	IGHJ3	Doria-Rose et al., Nature, 2014
VRC26.04	14	37	IGHV3-30	IGHD3-3	IGHJ3	Doria-Rose et al., Nature, 2014
VRC26.05	19	37	IGHV3-30	IGHD3-3	IGHJ3	Doria-Rose et al., Nature, 2014
VRC26.06	17	38	IGHV3-30	IGHD3-3	IGHJ3	Doria-Rose et al., Nature, 2014
VRC26.07	18	37	IGHV3-30	IGHD3-3	IGHJ3	Doria-Rose et al., Nature, 2014
VRC26.08	16	39	IGHV3-30	IGHD3-3	IGHJ3	Doria-Rose et al., Nature, 2014
VRC26.09	22	39	IGHV3-30	IGHD3-3	IGHJ3	Doria-Rose et al., Nature, 2014
VRC26.10	17	37	IGHV3-30	IGHD3-3	IGHJ3	Doria-Rose et al., Nature, 2014
VRC26.11	22	37	IGHV3-30	IGHD3-3	IGHJ3	Doria-Rose et al., Nature, 2014
VRC26.12	21	37	IGHV3-30	IGHD3-3	IGHJ3	Doria-Rose et al., Nature, 2014
CH104	21	15	IGHV4-31	IGHD6-13	IGHJ1	Doria-Rose et al., Nature, 2014
CH106	22	15	IGHV4-31	IGHD6-13	IGHJ1	Doria-Rose et al., Nature, 2014

Table 5.2 Database of bNAbs: bNAbs-DB. Characteristics of the heavy chains from somatic variants (n=90) are reported. SHM, Somatic Hypermutations.

5.2.3 Human immunoglobulin germlines

Human immunoglobulin heavy and light chain V, D, and J germline genes were downloaded from IMGT/GENE-DB [215].

5.2.4 Data annotation and preprocessing

Forward and reverse reads were paired using PANDAseq where applicable [81]. Data was submitted to IMGT/HighV-QUEST for annotation [214]. Annotated full-length VDJ sequences were pre-processed in downstream analyses by retaining (i) productive sequences (IMGT-defined), (ii) CDR3 of a minimal length of 3 a.a. and (iii) minimal abundance of 2 CDR3, where the same CDR3 sequence was present at least twice in the dataset and singletons were excluded. For all analyses, CDR3 abundances were calculated based on exact amino acid sequences (100% identity).

5.2.5 Networks

Networks were constructed from the Levenshtein distance (edit distance) matrix calculated of all-against-all CDR3 a.a. sequences in each repertoire [31]. The Levenshtein distance that measured insertions, deletions and substitutions was calculated with the stringDist function of Biostrings package in R. Each node of the network represents a CDR3 clone and each edge (link) connects CDR3 sequences that are 1 edit distance away (1 a.a. different). Networks construction, analysis and visualization was performed using the `igraph` R package [107].

5.2.6 Germline V/J gene analysis

V and J gene frequencies in antibody repertoires were calculated for each donor class (HIV-1 bNAb+, HIV-1 bNAb- and uninfected) and the bNAb database. The number of V-gene germlines was calculated from each donor class. In addition, the number of V-gene germlines represented in an antibody repertoire was calculated from each donor sample and time-point, where repertoires from different time-points of HIV-1 infection were included. Identity to the bNAbs and to the germline is calculated as the percentage of identity between V regions (a.a.) of sequences after calculation of the Levenshtein (edit) distance.

5.2.7 Evenness profiles

It was previously shown that evenness profiles can be used to capture a repertoire's state of clonal expansion [68]. Evenness profiles were calculated in a range of $\alpha = 0$ to $\alpha = 10$ with a step size of 0.2 as previously described by Greiff and colleagues. Briefly, clonal diversity was defined as ${}^{\alpha}\mathrm{D}(f) = (\sum_{i=1}^{n} f_i^{\alpha})^{\frac{1}{1-\alpha}}$, where f is the clonal frequency distribution with f_i being the frequency of each clone and n the total number of clones [121, 216, 217]. The α-values represent weights, which means as α increases, higher frequency clones are weighted more. The α-parameterized Diversity creates for a given array of alpha values a diversity index *profile* (short: diversity profile or $\underset{\alpha_D}{\rightarrow}$). Evenness describes the extent to which a given species frequency vector is distanced from the uniform distribution species frequency vector and is defined as (${}^{\alpha}\mathrm{E}$): ${}^{\alpha}\mathrm{D} = SR * {}^{\alpha}\mathrm{E}$ where SR is the species richness ($SR =^{\alpha=0}$ D), the number of unique clones in a repertoire dataset.

5.2.8 Hierarchical clustering

We clustered Evenness profiles [68] on their Pearson correlation matrix. Each tile in the heatmap represents a pairwise Pearson correlation coefficient between Evenness profiles of two given samples. Hierarchical clustering of the Pearson correlation matrix was performed using the standard UPGMA (unweighted pair group method with algorithm mean) clustering algorithm and was visualized as heatmap using the `aheatmap()` function from the `NMF` R package [186].

5.2.9 Sequence-based support vector machine analysis

We discriminated bNAb-like versus non-bNAb-like CDR3 clones based on CDR3 sequence using the KeBABS R package [218, 153]. Briefly, KeBABS enables kernel-based analysis of biological sequences using a position-independent gappy pair kernel that divides sequences into features of length k with gaps up to length m. For example, the sequence CARTA is decomposed by the gappy pair kernel with parameters k = 1 and m = 2 into monomers with gaps of zero to two amino acids in between: CA, C.R, C..T, AR, A.T, A..A, RT, R.A and TA [218].

We first calculated the balanced accuracy of the three classes of sequences versus each other: HIV-1 bNAb+, HIV-1 bNAb- and uninfected. We then built a support vector machine (SVM) model from equilibrating the input sequences for the classes. We trained the classifier by setting 80% of sequences as a training dataset and 20% of the sequences as a test dataset. After searching the parameter space for the optimal model by nested cross-validation, parameters were set to $k = 1$, $m = 2$, $C = 1$ (C is the cost for the misclassification of a sequence). Thus, the feature space used by the gappy pair kernel for a.a. sequences is $20^{2*k} * (m + 1) = 1.2 * 10^3$.

Prediction accuracy of class discrimination was quantified by calculating the balanced accuracy (0.5 * (Specificity + Sensitivity)), where specificity was defined as TN/(TN+FP) and sensitivity as TP/ (TP + FN) with TP, TN, FP, FN being true positive, true negative, false positive and false negative, respectively.

5.2.10 VDJ sequence alignment with `blastp`

Full bNAbs VDJ sequences were set as a reference database and sequences from samples were aligned to the reference database using protein-protein BLAST (BLASTP 2.2.31+). The fasta file from full VDJ a.a. sequences of HIV-1-infected donor sample was set as `query` and bNAbs VDJ full a.a. fasta file was set as `db`. The output top-aligned hits were analyzed.

5.2.11 Determination of statistical significance

Significance was tested using the Wilcoxon rank-sum test if not indicated otherwise. Tests were regarded as significant if $p < 0.05$.

5.3 Results

5.3.1 Construction of a bNAbs database and characterization of broadly neutralizing antibody sequences

In order to uncover features of bNAbs from high-throughput sequences of antibody repertoires, we compiled a database of 90 bNAb sequences (Tab. 5.1, 5.2). Somatic variants have been grouped according their HIV-1 binding sites in 10 families and 34 bNAb most potent clonal representatives (Tab. 5.1). The database was characterized for bNAb features of somatic hypermutations (SHM), CDR3 length, V/J germline gene usage and CDR3 similarity.

Many bNAbs develop an extremely large number of SHM, likely the result of co-evolution in response to persistent and rapidly mutating HIV-1 [204]. Extensive SHM,

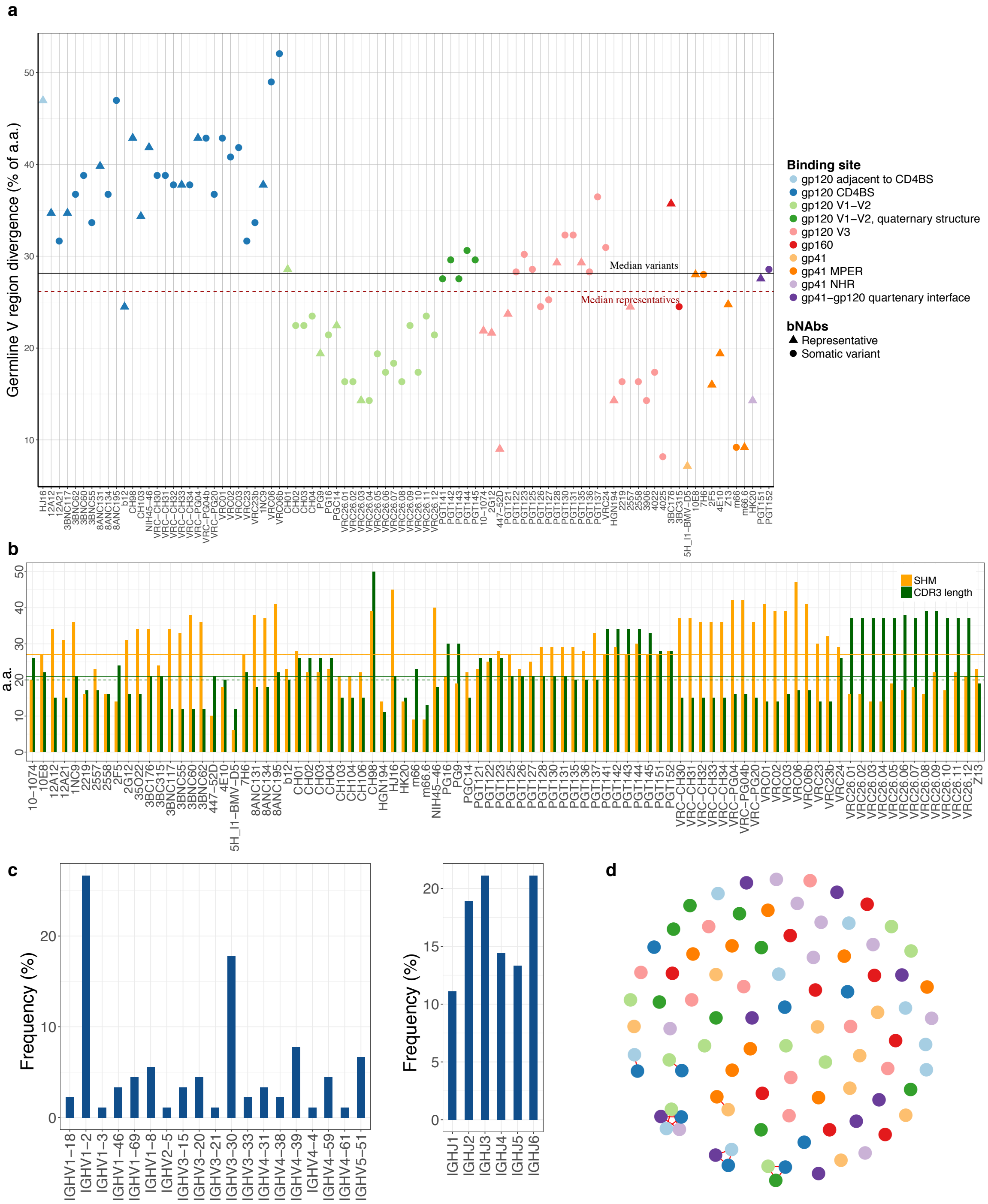


Fig. 5.1 Characterization of the bNAbs sequence database. (A) V-gene germline divergence (% a.a.). Epitope specificity is indicated by the color-code; somatic variants are indicated by dots, potent representatives of the group by triangles. (B) CDR3 length (a.a.) and somatic hypermutations (SHM) of each bNAb variant in the database. Median values are indicated by the dashed horizontal line for representative bNAbs and solid line for all bNAbs variants in database. (C) Frequency of V- and J-gene germlines in the database. (D) CDR3 a.a. similarity network of bNAbs color-coded by the binding site shows that a few bNAbs variants connect in similarity clusters, where similarity-linked bNAbs differ by 1 a.a. in the CDR3 sequence (e.g. bNAbs binding to gp120 site, blue cluster).

insertions and deletions are critical for broad neutralization and breadth [136, 93, 42]. Thus, we characterized the divergence from V-gene germline (% a.a.) of all variants of bNAbs sequences in our database (Fig. 5.1a, b). In order to avoid bias towards bNAbs with numerous variants which most likely have a similar number of SHM and CDR3 lengths (e.g., VCR26 has 12 variants), we have selected the most broadly neutralizing and potent representatives within each group of similar monoclonal antibodies in each clonal family, which are indicated in bold in Tab. 5.1. The selected bNAb representatives showed a median divergence of 26% a.a. from germline compared to the 28% of all variants. The most divergent were bNAbs specific for the CD4 binding site depicted in blue. Although some bNAbs, e.g., bNAbs specific for the gp41 binding site, presented a large range of divergence from V-gene germline (10–35%), potent representatives did not feature as the most V-gene germline-divergent within the group. Selected bNAb representatives, as well as bNAb variants in the database have a median of 27 a.a. SHM (min = 6, max = 45/47 a.a., Fig. 5.1b, Fig. 5.2a). CD4-binding-site (CD4BS) bNAbs directed against gp120 epitopes were the most somatically hypermutated with a median of 36 a.a. changes (Fig. 5.2b).

Furthermore, some bNAbs have very long complementarity determining regions 3 (CDR3), likely due to the fact that very long CDR3s enable them to traverse the structurally complex HIV-1 envelope proteins that are heavily shielded by glycans [219]. The median of CDR3 a.a. sequences was 21 a.a. in all bNAb variants and 20 a.a. in representatives (min = 11 a.a., max = 50 a.a., Fig. 5.1b, 5.2c). Indeed, bNAbs targeting the variable regions 1 and 2 (V1/V2) of HIV-1 gp120 envelope glycoprotein protected by extraordinary sequence diversity and N-linked glycosylation, were the the longest (>30 a.a.) among bNAbs (Fig. 5.2b).

Although it has been reported that some bNAb specificities are restricted to certain germlines, HIV-specific bNAbs do not appear to privilege a determinate IGHV [220]. bNAbs variants showed a more prevalent frequency of IGHV1-2, IGHV3-30 and IGHV4-39; IGHJ2, IGHJ3, and IGHJ6 were the most frequent J-gene germlines in the database (Fig. 5.1c).

Nevertheless, bNAbs have very different sequences (Fig. 5.1a) as shown by the wide range of the germline V region divergence even within bNAbs binding the same site, and the disconnected network of bNAbs (similar bNAbs that differ only by 1 a.a. in

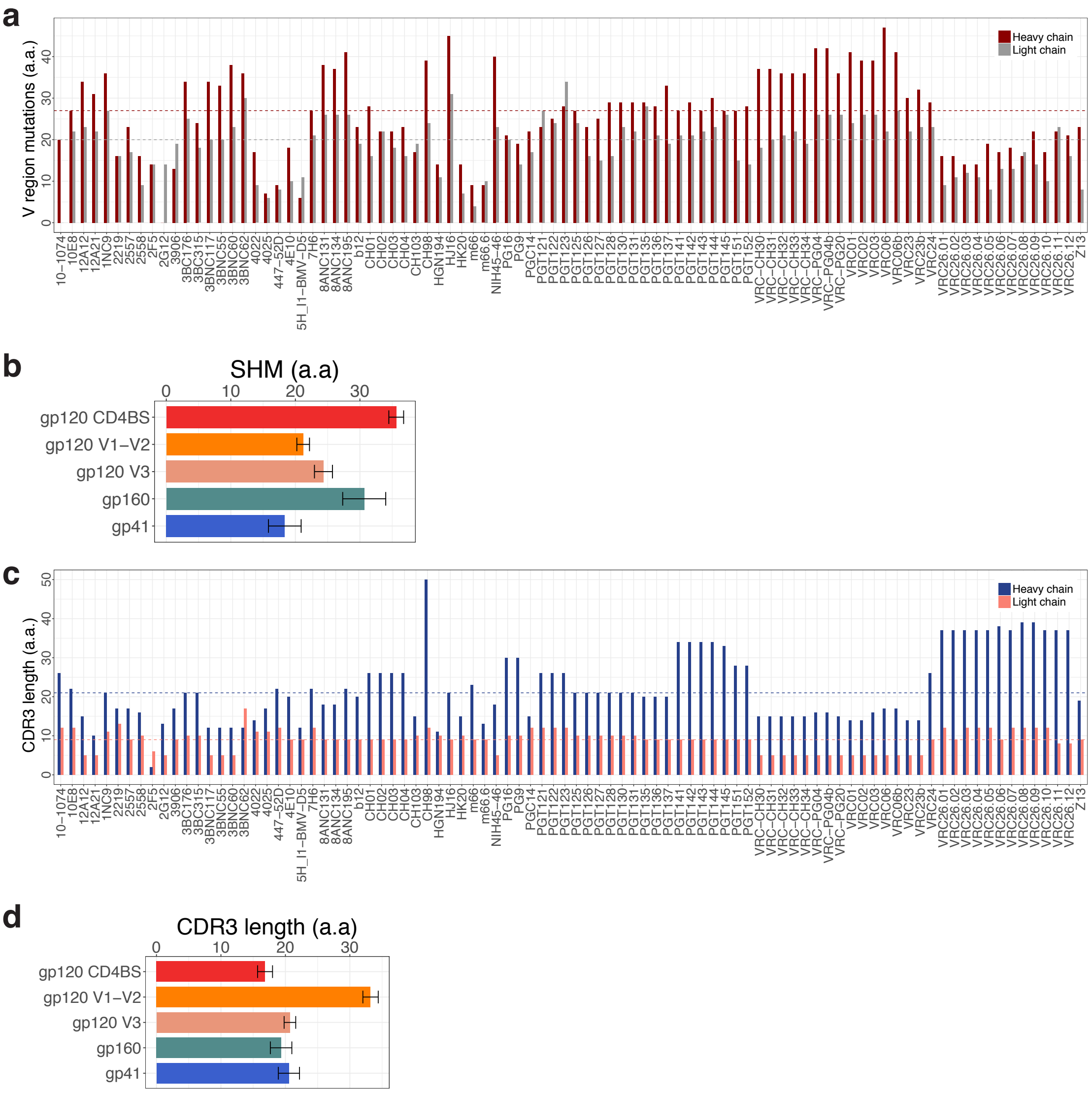


Fig. 5.2 SHM and CDR3 lengths in bNAbs-DB. (A) Somatic hypermutations (a.a) in database are shown for heavy and light chains. Median values are indicated with horizontal dashed lines. (B) Somatic hypermutations (a.a) in regard to the binding site of the bNAbs-DB. (C) CDR3 length (a.a) in database are shown for heavy and light chains. Median values are indicated with horizontal dashed lines. (D) CDR3 legnth (a.a) in regard to the binding site of the DB bNAbs.

their CDR3 are connected, Fig. 5.1d).

5.3.2 Global antibody repertoire analysis of SHM and CDR3 length in HIV-1 bNAb-, bNAb+ and uninfected individuals

bNAbs are characterized by a high number of SHM, elongated CDR3 sequences, and no specific V-gene germline usage (although, of note that the VRC01 and 8ANC131-class of bNAbs displayed IGHV genes restricted to IGHV1–2 and IGHV1–46 [206, 205]. We investigated the SHM and CDR3 length according to CDR3 frequency for each class of individuals (Fig. 5.3a).

More sequences with higher SHM than found in the reference bNAbs (>27 a.a. SHM) were present in HIV-1 bNAb+ (n=436, 0.3%) compared to HIV-1 bNAb- (n=386) and uninfected repertoires (n=170). However, the higher absolute number of hypermutated clones did not translate into a distinctive SHM signature at the repertoire level that could differentiate for bNAb status (0.3% hypermutated sequences in HIV-1 bNAb+, 0.4% in HIV-1 bNAb-, versus 1.8% in uninfected repertoires), and differences in average SHM in the repertoires were not significant (Fig. 5.3b). Longer CDR3 (above median length of bNAbs, 21 a.a.) sequences were eight fold more present in both HIV-1 bNAb+ (n=6,148) and HIV-1 bNAb- (n=6,266) repertoires compared to the uninfected and one clone (n=793) when absolute numbers were considered. However, long CDR3 sequences were constituted only 4.7% of HIV-1 bNAb+ repertoires compared to 6.9% in HIV-1 bNAb-, and 8.3% in uninfected repertoires, thus not allowing to discriminate bNAb status. Interestingly, sequences that respected the combination of SHM and CDR3 length parameters above bNAbs reference (>21 a.a. long and >27 a.a. SHM) were found only in HIV-1 bNAb+ (n=35, 0.03%) and uninfected individuals (n=19, 0.2%).

While CDR3 length distribution was comparable in the three categories of repertoires (Fig. 5.3a, top panel), SHM distribution showed that repertoires from uninfected individuals showed higher densities of clones in the range of 15–20 a.a. mutations compared to HIV-1 bNAb+/- repertoires (Fig. 5.3a, right panel). The analysis of SHM and CDR3 length showed that while bNAb+/- repertoires contain more clones above the bNAb database reference, average SHM and CDR3 lengths of repertoires do not

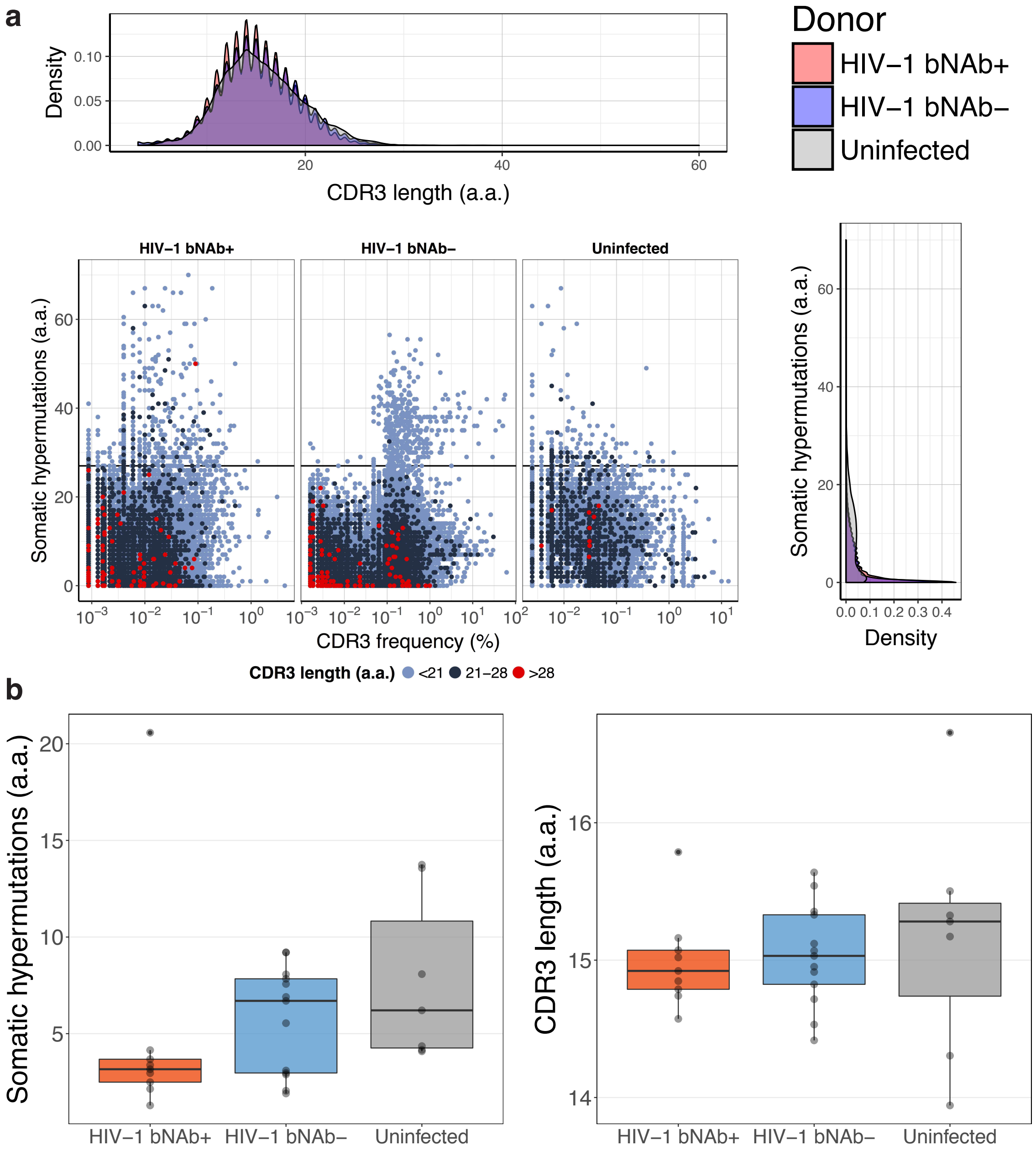


Fig. 5.3 CDR3 frequency, SHM and CDR3 length characterization. (A) HIV-1-infected that have developed bNAbs (HIV-1 bNAb+), HIV-1-infected that have not developed bNAbs (HIV-1 bNAb-) and HIV-1-uninfected donors (uninfected) have been characterized for CDR3 frequency, SHM and CDR3 length. CDR3 frequency of clones, indicating somatic hypermutations (a.a.) and color-coded according CDR3 length category (< 21 a.a., 21–28 a.a. and > 28 a.a.) of the donors separated by their class: HIV-1 bNAb+, HIV-1 bNAb- and uninfected individuals. Horizontal line represents the bNAb SHM median (27 a.a.) Top panel: Density of CDR3 lengths (a.a.). Right panel: Density of V region mutations (a.a.). (B) Somatic hypermutations (a.a.) for each individual donor are shown as dots. Differences not significant. (C) CDR3 length (a.a.) for each individual. Differences not significant. Bar plots show median±s.e.m.

serve as indication of bNAb status. Of note that CDR3 sequences showed comparable a.a. frequencies (Fig. 5.4).

5.3.3 While V/J germline analysis does not discriminate bNAbs status, V region similarity to known bNAbs is higher in HIV-1 bNAb+ repertoires

Although bNAbs recognize HIV-1 through their CDR3, the VRC01-class and the 8ANC131-class of bNAbs have shown to bind CD4BS mainly through CDR2 (located entirely in the V region) and displayed IGHV genes restricted to IGHV1–2 and IGHV1–46 [206, 205]. Therefore, we compared V- and J-gene germline species richness and frequencies among classes in order to determine whether these could discriminate samples by bNAb status.

In total 49 different V-genes were present in the HIV-1 bNAb+ repertoires' class, 51 in HIV-1 bNAb- donors and 49 in uninfected (Fig. 5.5a). Interestingly, the differences in the number of V-gene germlines were significant between HIV-1 bNAb+ (mean±s.e.m = 43.6±0.4) and HIV-1 bNAb- (mean±s.e.m = 36.3±0.6) individuals, but these were not significant compared to uninfected donors (mean±s.e.m = 35.1±4.8, Fig. 5.5b).

HIV-1 bNAb-/+ and uninfected individuals showed similar V-gene frequencies, and HIV-1 bNAbs+ and HIV-1 bNAbs- individuals used V-genes independently of bNAb status ($r_{\text{Pearson}} = 0.95$, Fig. 3a). Their V-gene frequencies correlated less with uninfected individuals (respectively, $r_{\text{Pearson}} = 0.88$, 0.84, Fig. 3a). Correlation of J-gene usage was high across all three donor categories ($r_{\text{Pearson}} = 0.98$–0.99, Fig. 5.3a) and all 6 J-genes were present.

In order to investigate whether the difference in the number of V-gene germlines represented in repertoires translated to a distinctive fingerprint in V-region similarity to bNAbs for HIV-1 bNAb+ donors compared to controls (HIV-1 bNAb- and uninfected), we calculated the sequence similarity to bNAbs and divergence from germline (Fig. 3c). Indeed, twice as much sequences with a V region above 90% a.a. similarity to bNAb V regions were detected in the HIV-1 bNAb+ class (n=2,602; 0.02%) compared to HIV-1 bNAb- (n=919; 0.01%) and uninfected donors (n=101; 0.01%). V region identity to bNAbs could discern the class of repertoires. When the whole VDJ region

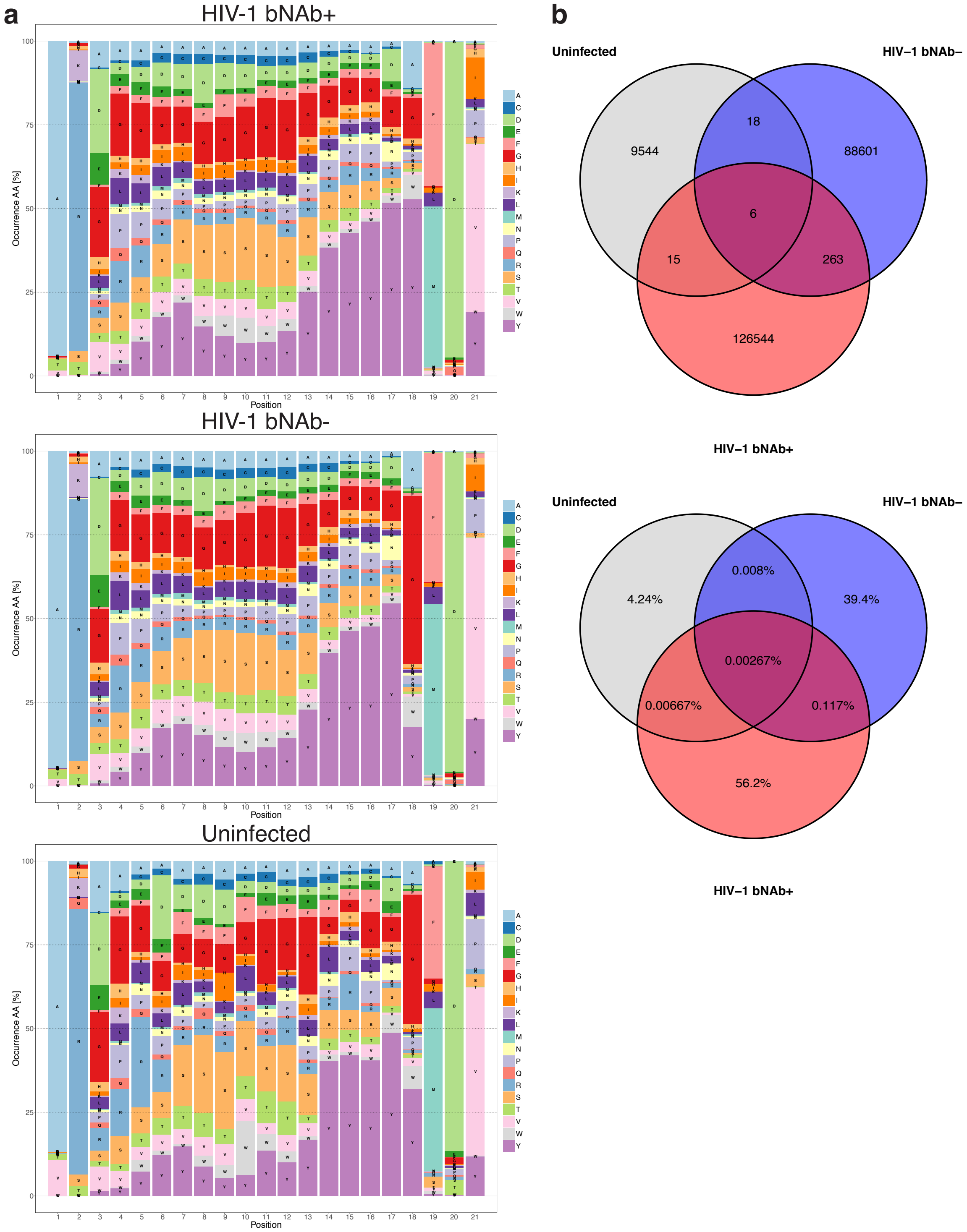


Fig. 5.4 Sequence composition and overlap (A) Quantification of a.a. composition in cohorts in sequences with CDR3 length = 21 a.a. (median bNAbs-DB). (B) Overlap of the number of sequences (top panel) and of their percentage (bottom panel) between cohorts.

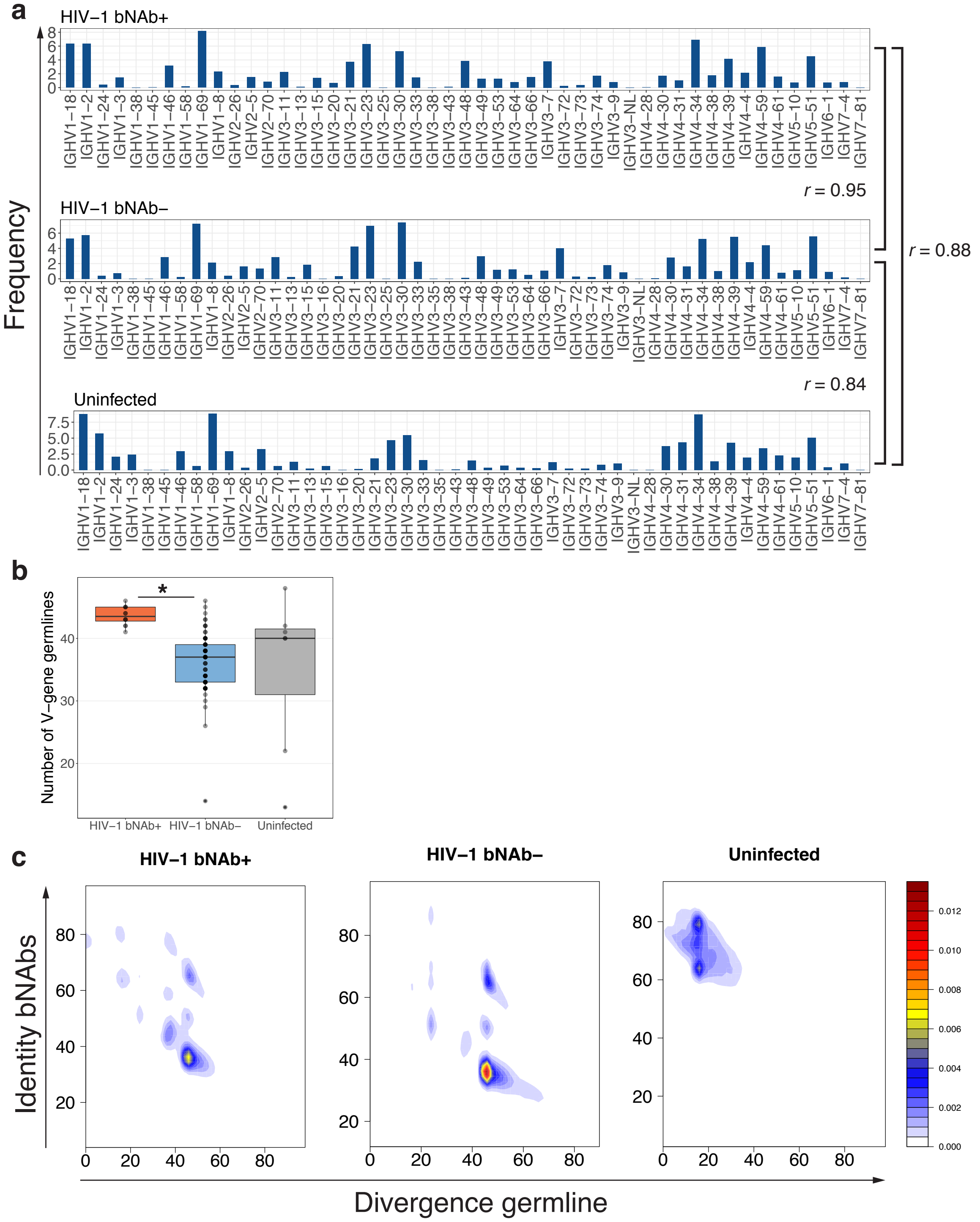


Fig. 5.5 bNAb+ are differentiated by germline V-gene characterization, and their sequence similarity to sequences in bNAbs-DB is 3–25 fold higher than in HIV-1 bNAb- and uninfected individuals.

Fig. 5.5 (A) V-gene frequencies and Pearson correlation (r) of HIV-1 infected individuals that have developed bNAbs, that have not developed bNAbs and uninfected donors. (B) Number of V-gene germlines represented in antibody repertoires. Dots represent single repertoire samples (n=83). (C) V region identity to bNAbs and V region divergence from germline shows that HIV-1-infected individuals show V regions that are more divergent from germline. Respectively, 3 and 25 fold more sequences are $>$ 90% similar to bNAbs in HIV-1 bNAb+ then in HIV-1 bNAb+ and uninfected individuals.

was considered, identity to bNAb was not a discriminating characteristic for bNAb status (Fig. 5.6b). Although a general tendency towards a higher density of sequences more diverging from germline was detected in HIV-1 infected individuals, this germline divergence did not seem to be distinctive for bNAb-status. Furthermore, the combinations frequencies of V-J germline genes were similar in the 3 classes of donors (Fig. 5.6c).

5.3.4 HIV-1-bNAb+ individuals show similar clonal expansion, and repertoire architecture differs between HIV-1 infected and uninfected individuals

The repertoire's state of clonal expansion, described by evenness profiles, can be used for detection of immunological status [68]. Therefore, evenness profiles from each donor and time point were calculated as described by Greiff et al. (Fig. 5.7a, see Methods) and subjected to hierarchical clustering. Briefly, evenness profiles describe the extent to which a given species frequency vector is distanced from the uniform distribution species frequency vector. Hierarchical clustering led to three big clusters in one of which 10 of the 12 HIV-1 bNAb+ samples were located. Thus, HIV-1 bNAb+ individuals showed similar dynamics of clonal expansion. However, clonal expansion was not sufficient to discriminate between HIV-1 bNAb- and HIV-1 bNAb+ samples.

Network CDR3 sequence similarity describe the architecture of an antibody repertoire [31]. In order to determine whether repertoire architecture differed among classes, we constructed networks from the Levenshtein distance matrix of all-against-all CDR3 sequence similarity comparisons for each sample (n=83, see Methods, Fig. 5.7b). Repertoire architecture as quantified by the network clustering coefficient was discriminative of HIV-1 infected and uninfected individuals but did not differentiate between HIV-1

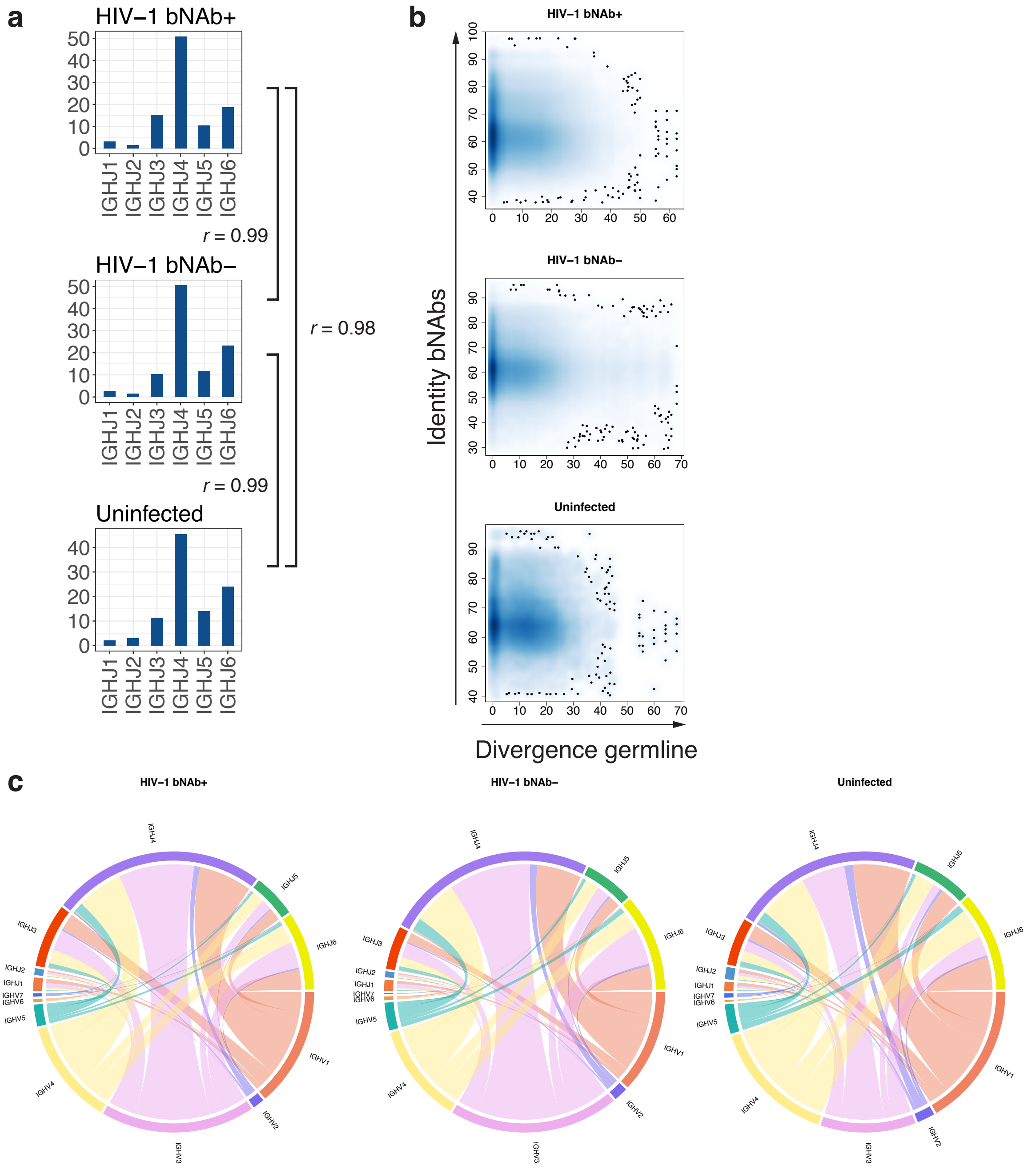


Fig. 5.6 J-gene frequency, V-J combinations and VDJ sequence identity to database bNAbs. (A) J-gene germline frequencies in HIV-1 bNAb+, HIV-1 bNAb-, and uninfected classes. (B) VDJ a.a. similarity to DB bNAbs and divergence from germline. Sequences have been aligned using Blastp. (C) Quantification of V-J gene combinations.

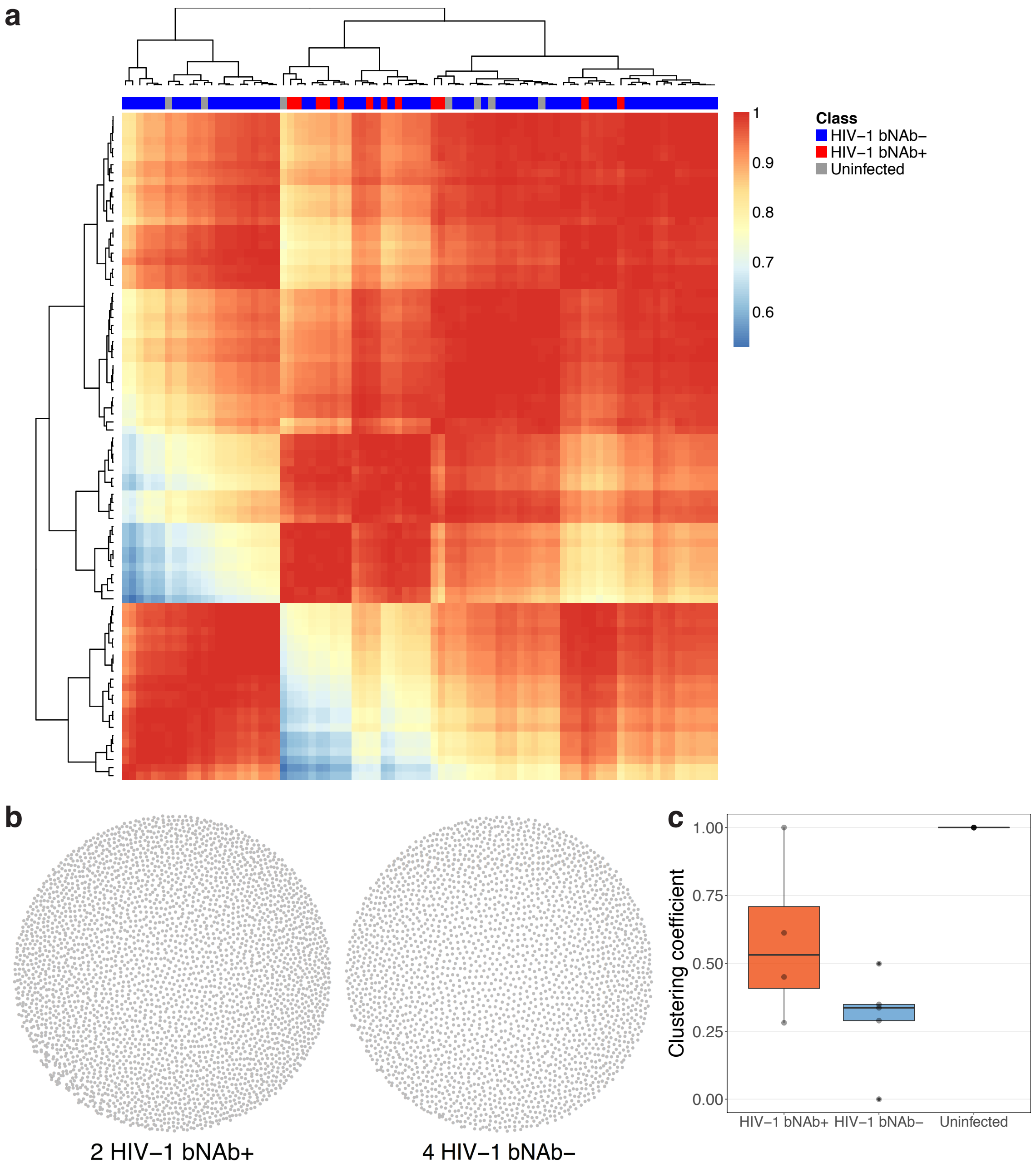


Fig. 5.7 Evenness profiles of CDR3 frequency distributions and CDR3 similarity profiles. (A) Hierarchical clustering of evenness profiles performed on Pearson correlation distance. The heatmap shows pairwise Pearson correlation coefficients of all evenness profiles (n=83 samples: HIV-1 bNAb+ n=12, HIV-1 bNAb- n=64 and uninfected n=7). Evenness profiles were calculated in a range of α=0 to α=10 with a step size of 0.2: ${}^{\alpha}\mathrm{D} = SR * {}^{\alpha}\mathrm{E}$ where SR is the species richness ($SR =^{\alpha=0}$ D), the number of unique clones in a repertoire dataset. (B) CDR3 similarity networks were constructed from the Levenshtein distance matrix of all-against-all CDR3 a.a. sequences. Exemplary networks from representative individuals. (C) Clustering coefficient of CDR3 similarity networks (n=15). Box plot indicates median±s.e.m. Differences between HIV-1 and uninfected individuals are significant (p<0.05).

bNAb- and HIV-1 bNAb+ (Fig. 5.7c).

5.3.5 Machine learning predicts bNAb-DB sequences from HIV-1 bNAb- with 95% accuracy

Some specificities like high SHM, long CDR3, and restriction to certain germlines are present amongst bNAbs but not necessarily required for broad neutralizing activity. Additionally, bNAb sequences have been reported to be rare. These factors render it challenging to relate bNAb sequences to a whole-repertoire signature. Machine learning, and in particular support vector machines (SVM), have emerged as a best practice for the classification of biological sequences [221]. In order to determine whether the bNAb-DB and HIV-1 bNAb- individuals differed on the repertoire sequence level, we leveraged sequence-based SVM analysis [218, 153]. We built the model from all the bNAbs-DB sequences and used HIV-1 bNAbs- sequences as a negative control. Because of the different size between the bNAbs-DB and HIV-1 bNAbs- sequences, we split the HIV-1 bNAbs- sequences into chunks of the same size of bNAbs-DB and used 80% of the sequences as a training dataset and 20% as a test dataset. Sequences in the bNAbs-DB versus HIV-1 bNAb- individuals were predicted with 95% accuracy (random is 50%, Fig. 5.8b). Briefly, in order to predict sequences based on features we used kernel-based analysis of biological sequences using a position-independent gappy pair kernel that divided sequences into features separated by gaps [218] and quantified prediction accuracy of class discrimination by calculating the balanced accuracy (BACC = 0.5 * (Specificity + Sensitivity)).

Sequences in the three classes of individuals (HIV-1 bNAb+, bNAb- and uninfected) were diverse, and only a few sequences were shared across individuals (0.002–0.117%, Fig. 5.4). Although there is a high overall sequence diversity among the different classes, we used machine learning to detect intrinsic sequence features that go beyond sequence similarity, thus allowing to detect sequences with similar features although distant in their absolute a.a. homology. In order to investigate if features of sequencing data were distinctive of the class, we investigated the sequences from each class and compared them to each other (Fig. 5.8a). The classes could not be predicted at the repertoire level (Figure Fig. 5.8a).

Given the substantial immunogenomic differences between bNAb-DB and HIV-1 bNAb-, we used the constructed SVM model to search for bNAb-like CDR3 sequences in the repertoires of HIV-1 bNAb+ individuals. We built SVM models based on equally-sized chunks of HIV- bNAb- and bNAb-DB sequences, and we applied these SVM models to the entire set of uninfected and HIV-1 bNAbs+ sequences. In these individuals, we determined that 22% out of all CDR3 were bNAb-like, which was a significantly higher percentage compared to the uninfected cohort (17%, Fig. 5.8b).

5.3.6 Discussion

Broadly neutralizing antibodies are potential therapeutics and guides to vaccine design [222, 198, 168, 149]. However, methods for the discovery of broadly neutralizing antibodies are laborious and inefficient, with around 30 bNAbs discovered over the last 30 years [223, 199]. Advances in HTS technologies have made it possible to interrogate in large-scale antibody sequences from HIV-1 individuals [27], thus harnessing the potential of computational methods for de novo discovery of bNAbs [90].

We constructed a database of 90 bNAbs sequences (https://github.com/enkelejdamiho/bNAbs-DB) in order to gather the full VDJ sequences of known bNAbs and characterize them (Tab. 5.1, 5.2). The bNAbs-DB was built drawing sequences from all available sources like bNAber [213], GenBank, Los Alamos HIV Molecular Immunology (http://www.hiv.lanl.gov/), Abysis (http://bioinf.org.uk/abysis/), scientific literature and personal communications. In addition to its advantage in comprehensiveness of sequences, bNAbs-DB allows a coherent and standard analysis of all bNAbs features (Fig. 5.1).

While persistence and continuation of SHM in contribution to lineage divergence in bNAb evolution has been confirmed to be a determining factor of the breadth and potency of bNAbs [42], our results indicate that the SHM of bNAb sequences is not reflected at the global level of the antibody repertoire. Nevertheless, bNAb+ individuals present more sequences that are largely mutated (up to 50% SHM) compared to controls. bNAb+ repertoires showed also more sequences with longer CDR3 than the average length of bNAbs-DB, however they were not discriminated by the average CDR3 length of their clones (Fig. 5.3).

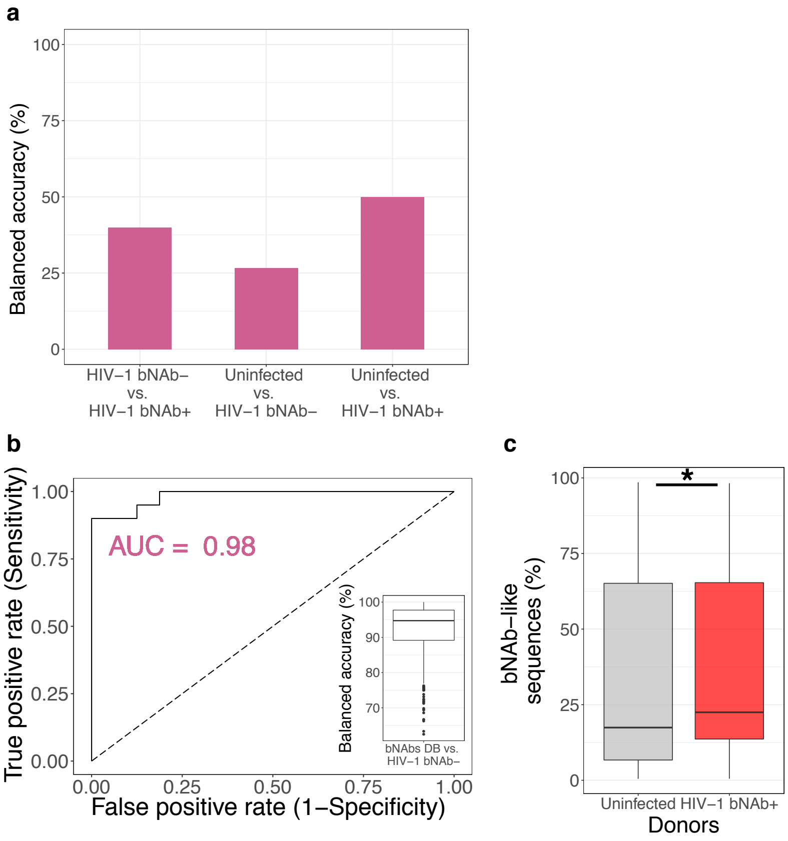


Fig. 5.8 The frequency of bNAb-like CDR3s in HIV-1 bNAb+ is significantly higher compared to HIV-1-uninfected individuals. (A) Balanced accuracy of sequence-based SVM model comparing sequences from three classes of repertoires: HIV-1 bNAb+, HIV-1 bNAb- and uninfected among each-other. Balanced accuracy is the mean of specificity and sensitivity (random classifier: 50%). (B) The ROC curve of SVM model shows the true positive rate against the false positive rate. Balanced accuracy (median±s.e.m. = 95±0.2%) of sequence-based SVM model (see Methods) classifying sequences in the bNAbs database (bNAbs-DB) vs. sequences from HIV-1 bNAb- samples. The HIV-1 bNAb- class was divided into chunks of the size of the bNAbs-DB prior to running the sequence-based SVM across all chunks (Training data: 80%, Test data: 20%). (B) Percentage of bNAb-like CDR3s in test datasets of HIV-1 bNAb+ (median±s.e.m.=22±0.9%) and uninfected individuals (median±s.e.m.= 17±0.9%). SVM models built for each chunk in (B) were applied to both uninfected and HIV-1 bNAb+ donors as test data in order enumerate bNAb-like CDR3s (displayed as percentage among all CDR3s per class). The SVM models each contribute to the variation observed. Significance was computed using Wilcoxon rank-sum test. Differences were significant ($p < 10^{-9}$).

Although a recent study of germline profiles showed no significant differences in germline IGHV repertoires between individuals who develop and who do not develop bNAbs ([220]), somatic variants of a specific bNAb lineage use unique germlines [42]. Our results showed no detectable differences in germline frequency, but significant differences in the numbers of germlines present between HIV-1 and uninfected individuals (Fig. 5.5). Nevertheless, these differences were not significant regarding bNAb status. On the other hand, the percentage of V region identity to bNAbs-DB sequences was 3–25 fold higher in bNAb+ repertoires compared to the other two control classes (HIV-1 bNAb- and uninfected), suggesting an application of this method to detect bNAb status.

Evenness profiles based on clonal frequency distributions have been previously applied to detect immunological status [68] and described a repertoire's state of clonal expansion. Evenness profiles of bNAb+ individuals showed the highest similarity out of the three classes, however it did not discern all bNAb+ and bNAb- repertoires. While repertoire architecture differed between HIV-1 and uninfected classes of individuals, it did not differentiate between HIV-1 bNAb+ and bNAb- classes (Fig. 5.7).

Antibody repertoire sequencing datasets have steadily increased over the past few years to 10^6 sequences/sample [27, 5, 190]. Meanwhile, machine learning has been applied progressively to analyze large and complex sequencing data sets [221, 224]. We leveraged HTS of antibody repertoires in order to identify bNAb features through supervised machine learning: we set to determine if the sequences from HIV-1 bNAb+ versus uninfected donors' class were distinct. Comparing repertoires based on characteristics of bNAbs (e.g., SHM, CDR3 length) could not discriminate between bNAb+/- class. The support vector machine (SVM) model was build using the bNAbs-DB as a positive control and HIV-1 bNAbs- as a negative control. Although antibody repertoire sequences from uninfected individuals could be used as a negative control, it is not known if uninfected individuals have developed bNAbs. When sequence-features at the repertoire level of the three classes of HIV-1 bNAb+, bNAb- and uninfected repertoires were compared using the SVM model, these classes could not be discriminated based on the sequence features retrieved by the model. Nevertheless, the selection of features based on known bNAbs sequences (bNAbs-DB) allowed to discriminate sequences of the bNAbs-DB versus HIV-1 bNAb- sequences with 95% prediction accuracy. Significantly more bNAb-like sequences were found in HIV-1 bNAb+ repertoires by applying the model to discriminate between HIV-1 bNAb+ repertoires compared to uninfected controls (5.8). Although there is a large variance because of the difference in the number of sequences between HIV-1 bNAbs+ and bNAbs-DB, this variance can be lowered by increasing the bNAbs database and by adding bNAbs to different binding sites. The variance of bNAb-like sequence frequencies is independent of BACC, which indicates that the bNAbs traing database size needs to be increased. Our results show how SVM can be applied to detect bNAbs-like sequence signatures and although there is a large variance, the difference between uninfected and HIV-1 bNAbs+ classes is significant.

This work suggests that low-dimensional features such as SHM, CDR3 length, germline gene usage, etc. are insufficient to detect bNAb-status differences at the repertoire level. High-dimensional methods are needed to detect shifts in repertoire architecture and composition. Because new and more potent bNAbs could be different in sequence from known bNAbs, but still retaining intrinsic sequence features, methods like V region identity to bNAbs-DB may prove not optimal for de novo bNAb discovery. Machine learning captures high-dimensional sequence features, allowing for a less biased selection of bNAb-like sequences and allows to detect bNAb-like sequences in the bNAbs-DB versus HIV-1 bNAb- with 95% accuracy. Additionally, better experimental data is needed: higher coverage, sorted populations, and less errors [5] in order to increase biological conclusiveness of high-throughput large-scale studies.

Construction of a bigger bNAbs-DB might be advantageous for an in-depth analysis of the differences in sequence composition between bNAb-like and non-bNAb-like CDR3 sequences in order to engineer in silico bNAb-like CDR3 sequences in the future. The construction of a catalog of bNAb sequences, all potentially with slightly different therapeutic function, and synthetic bNAb-biased repertoire libraries to screen for more affinity matured and potent bNAbs, might be a near future. In addition to bNAb-like sequence detection against HIV-1, this work can support the research of other broadly neutralizing antibodies e.g., against influenza [54] and dengue [138].

5.3.7 Conclusions

Our results show that bNAb-like sequences can be predicted from high-throughput antibody repertoire sequencing data. While sequences within the bNAbs' reference range of SHM and CDR3 length can be detected at higher numbers in HIV-1 bNAb+ repertoires than in controls (HIV-1 bNAb- and uninfected), these characteristics are not reflected at the repertoire level. Thus, bNAb status could not be predicted by the average SHM or CDR3 length in a repertoire. Although a few bNAbs use specific germlines, V and J germline frequencies were highly correlated in HIV-1 bNAb+, HIV-1 bNAb- and uninfected cohorts and did not assist the discrimination of bNAb status. While sequence identity to bNAbs-DB detected a larger fraction of sequences with bNAbs characteristics in bNAb+ cohorts, by applying machine learning we could predict bNAb-like sequences in bNAbs-DB versus HIV-1 bNAb- with 95% prediction accuracy and detect significantly more bNAb-like sequences in HIV-1 bNAb+ than in uninfected repertoires. Systems analysis may thus advance the information quality and knowledge from large-scale data, supporting the future de novo discovery of bNAbs.

Chapter 6

SystimsDB: a database of immune repertoires

6.1 Abstract

Immune repertoires are large and diverse collections of B and T cell receptors, each clone defined by the CDR3 region which is responsible for antigen binding. High-throughput sequencing has enabled the detection, deposition and analysis of millions of sequences from immune repertoires. However, immune sequences have not been aggregated into a database, thus hindering the cross-analysis and benchmarking of immune sequences and repertoires from public data. Therefore, we constructed an immune repertoire database (systimsDB) to incorporate the collection of 80 public datasets of 7,592 publicly available sequencing samples from 5 different species resulting in a total of $\approx$ 6.5 billion sequences, and 142 sequencing sample resulting in $\approx$ 0.2 billion sequences generated in-house. Sequences have been VDJ annotated in systimsDB, thus providing immediate usability for downstream analysis. Specifically, we provide users with the options of searching, exporting results and analyzing entire annotated immune repertoires and partial sets of sequences according to user-defined selection criteria. The database is available as a web server at https://www.systimsdb.ethz.ch/index.html.

6.2 Introduction

Immune repertoires are collections of B-cell receptors (BCR, antibody) and T-cell receptors (TCR). BCRs and TCRs are formed by two chains of similar domains: heavy and light chains in B cells and and chains in T cells. Naïve B or T cells are generated through the somatic recombination of Variable (V), Diversity (D), and Joining (J) gene segments, and further addition of nucleotides at the junctions, which encode the variable domains of the chains. Clonality and antigen-binding is mainly determined by the complementarity determining region 3 (CDR3) in the variable domains. Further somatic hypermutations add to the diversity in the case of BCR repertoires. Thus, because of the underlying mechanism of their generation, immune repertoires are characterized by an enormous diversity in their CDR3 clonal sequences (Fig. 6.1).

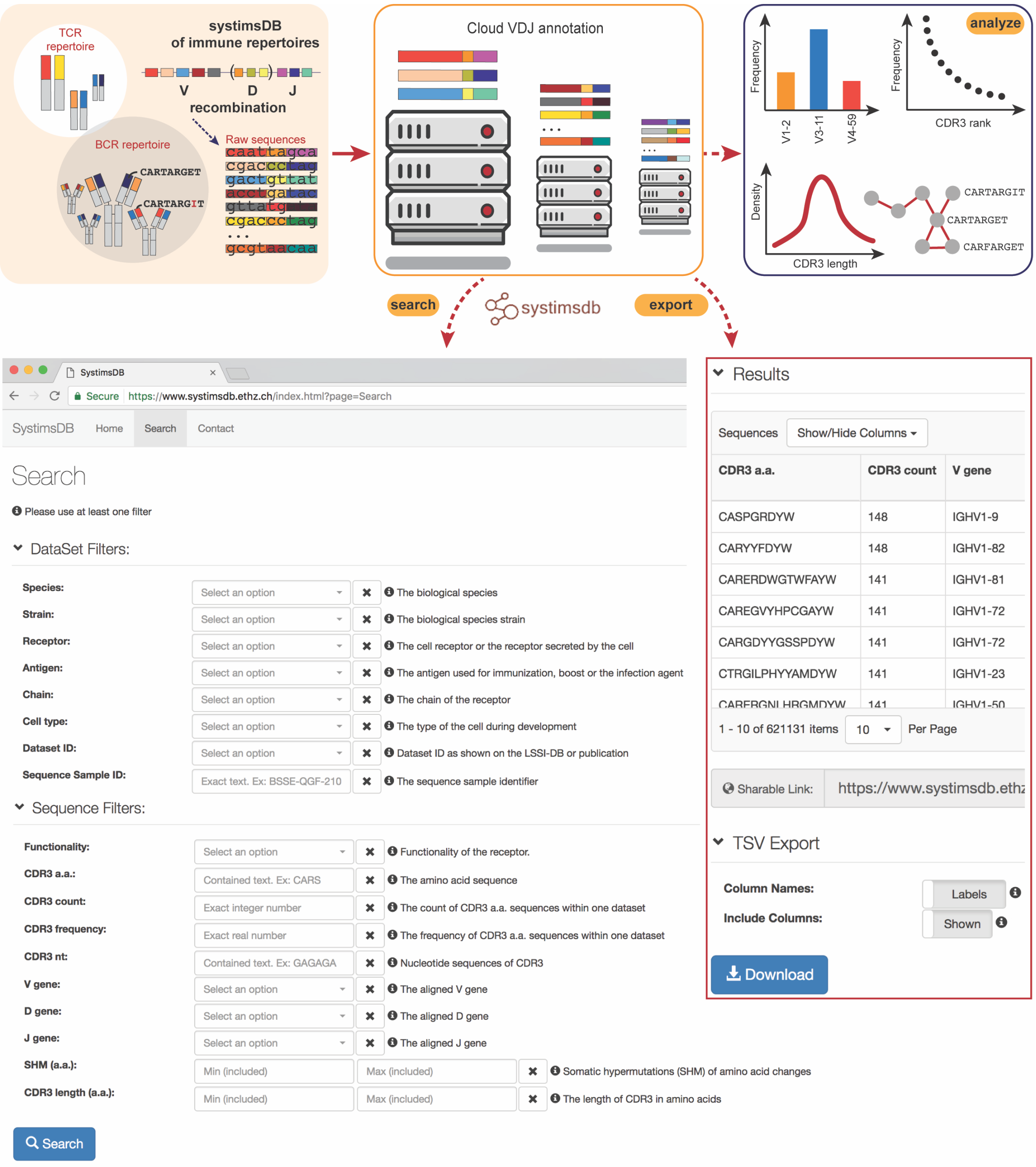
High-throughput sequencing has enabled the detection, deposition and analysis of millions of BCR and TCR sequences [27, 26, 225, 68], capturing the diversity of natural immune repertoires in order to answer fundamental immunological questions [31] and investigate immune repertoires as biomarkers [226–228]. Furthermore, in the last decade, the biopharma industry has extensively sequenced synthetic immune systems (e.g., antibody display libraries) for the discovery of therapeutic antibodies with improved affinity, specificity and stability [229–232]. Thus, there is a large amount of natural and synthetic immune repertoire sequencing data.

Here we provide systimsDB as a collection raw sequences generated in-house and downloaded from repositories through accession numbers indicated in the literature [233, 94, 234–236, 132, 48, 237, 152, 238, 16, 239–244, 190, 245–248, 118, 249, 95, 250, 251, 54, 252, 110, 253, 142, 254–257, 92, 64, 258, 43, 259–264, 138, 265–267, 133, 268, 36, 269, 93, 270, 271, 97, 272, 44, 273–276, 60, 277–279, 41, 99, 280, 55, 15, 281, 40, 42, 282, 283, 206, 98, 91, 90, 147]. In addition, we provide a ready-to-use set of clones that have been annotated with a standardized bioinformatic pipeline and sample analyses that can serve as a control to immune repertoire sequencing data. The web server supports a user-friendly search function, selection of processed CDR3 clones, download of data of interest and repertoire analysis across clones and datasets.

6.3 Material and methods

6.3.1 Data collection

We have compiled systimsDB from BCR repertoire sequences generated in-house and from public data of BCR/TCR immune repertoires. Accession codes were identified from peer-reviewed scientific publications from literature search and on-line sources of refer-



SystimsDB

Secure https://www.systimsdb.ethz.ch/index.html?page=Search

SystimsDB Home Search Contact

Search

Please use at least one filter

DataSet Filters:

Species:	Select an option	The biological species
Strain:	Select an option	The biological species strain
Receptor:	Select an option	The cell receptor or the receptor secreted by the cell
Antigen:	Select an option	The antigen used for immunization, boost or the infection agent
Chain:	Select an option	The chain of the receptor
Cell type:	Select an option	The type of the cell during development
Dataset ID:	Select an option	Dataset ID as shown on the LSSI-DB or publication
Sequence Sample ID:	Exact text. Ex: BSSE-QGF-210	The sequence sample identifier

Sequence Filters:

Functionality:	Select an option		Functionality of the receptor.
CDR3 a.a.:	Contained text. Ex: CARS		The amino acid sequence
CDR3 count:	Exact integer number		The count of CDR3 a.a. sequences within one dataset
CDR3 frequency:	Exact real number		The frequency of CDR3 a.a. sequences within one dataset
CDR3 nt:	Contained text. Ex: GAGAGA		Nucleotide sequences of CDR3
V gene:	Select an option		The aligned V gene
D gene:	Select an option		The aligned D gene
J gene:	Select an option		The aligned J gene
SHM (a.a.):	Min (included)	Max (included)	Somatic hypermutations (SHM) of amino acid changes
CDR3 length (a.a.):	Min (included)	Max (included)	The length of CDR3 in amino acids

Search

Results

Sequences Show/Hide Columns

CDR3 a.a.	CDR3 count	V gene
CASPGRDYW	148	IGHV1-9
CARYYFDYW	148	IGHV1-82
CARERDWGTWFAYW	141	IGHV1-81
CAREGVYHPCGAYW	141	IGHV1-72
CARGDYYGSSPDYW	141	IGHV1-72
CTRGILPHYYAMDYW	141	IGHV1-23
CAREFRGNLHRGMDYW	141	IGHV1-50

1 - 10 of 621131 items 10 Per Page

Sharable Link: https://www.systimsdb.ethz

TSV Export

Column Names: Labels

Include Columns: Shown

Download

Fig. 6.1 systimsDB. Publicly available immune repertoire sequences are collected and sequentially VDJ annotated using parallel computing. systimsDB is populated with CDR3 clones and is searchable. The database is searchable at a sequence level and at a dataset level. Selected results can be downloaded. Post-processing analysis can be performed or downloaded for a dataset.

ence for public available datasets (http://b-t.cr/t/publicly-available-airr-seq-data-sets/317). Sample Accession numbers were used to identify codes of sequencing runs from the NCBI Sequence Read Archive (SRA, https://www.ncbi.nlm.nih.gov/sra/). Raw reads from publicly available datasets were downloaded from the SRA, European Nucleotide Archive (ENA, http://www.ebi.ac.uk/ena), the database of Genotypes and Phenotypes (dbGaP, https://www.ncbi.nlm.nih.gov/gap), the public functional genomics data repository Gene Expression Omnibus (GEO), and collected in fastq files using the EMBL-EBI FTP server (ftp://ftp.sra.ebi.ac.uk). Public data was retrieved through literature and was annotated with BCR/TCR information, species, antigen (or infection/disease/vaccination status) and cell populations specifications where applicable.

6.3.2 Data annotation

Raw reads downloaded from the source were annotated using MiXCR version 2.1.5 [17]. B and T cell clones were assembled based on CDR3 a.a. sequence and were exported filter out of frame sequences, and sequences containing stop codons.

6.3.3 Database design

Three base data structures were used to store the data in the SQL database (Fig. 6.2):

1. *Dataset table* that contains the Dataset ID, Sample name, Species, Strain, Cell Type (Naïve B cell, Plasmacell, etc), Antigen (OVA, HIV-1, etc), Chain (IGH, IGK, TRA, etc), Status (confidential/public).

2. *Clone table* that contains the clone ID, Dataset ID, Read ID, Isotype, CDR3 AA, CDR3 NT, V gene, D gene, J gene, SHM AA, SHM NT, CDR3 Length, Functionality (true/false), CDR3 Frequency.

3. *Status table* that contains Dataset ID, Info (indicates the stage of the analysis: e.g., fetching data, assembling clones), Type (begin/end), Timestamp (time when data was processed and included in database).

The dataset table is populated with all the information about the datasets stored in-house, as well as from the public datasets. To each dataset is assigned/extracted a Dataset ID and were preprocessed with MiXCR [17] parallel-wise using Snakemake

[284] on the ETH computational cluster (Euler). Additionally, a status table is updated with information on the progress of the processing. The output of the preprocessing is exported as clones and inserted into the sequence table.

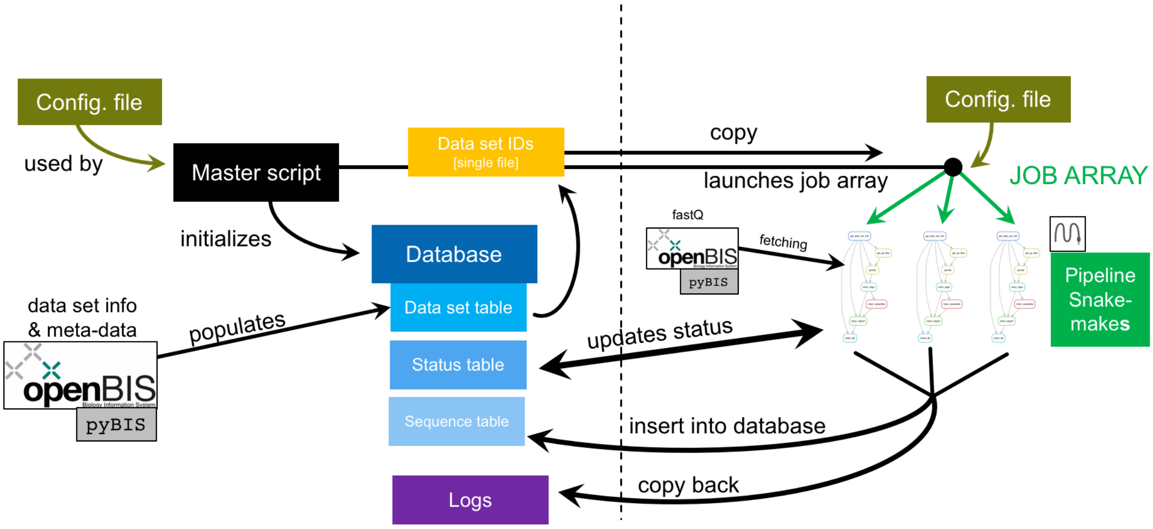


Fig. 6.2 Database design. Database is structured in *Dataset table*, *Status table* and *Sequence table*. The *Sequence table* is populated after CDR3 are annotated. (Euler: ETH computational cluster; openBIS: ETH raw sequences storage).

6.3.4 Database construction

systimsDB was constructed by collecting of raw sequences generated in-house and downloading from repositories through accession numbers as indicated in the literature. Publicly available datasets were downloaded using the accession numbers (http://b-t.cr/t/publicly-available-airr-seq-data-sets/317). The database was populated by indexing clones and extracting the CDR3 a.a. sequence, the CDR3 nt sequence, somatic hypermutations (a.a.), CDR3 length, CDR3 frequency and count. For each CDR3 clone, alleles (e.g., *00) were removed from the best aligned V gene, D gene, and J gene and the isotype information was retained in the database.

6.3.5 Web application

SystemsDB application has been developed using a standard web technology stack and following a standard three-tier architecture:

- *Tier 1, Database*: SystemsDB data is stored in a standard SQL database.
- *Tier 2, Backend*: Data is then queried by the backend, on this case the backend is written using a Python framework and a JSON-RPC 2.0 API exposes the query functionality. The backend returns the results as standard JSON objects that can be used by any application.
- *Tier 3, Frontend*: A single-page web application is served by an Apache web server, this application has been developed using standard HTML5 and JavaScript technologies and communicates with the backend using the JSON-RPC 2.0 API.

This architecture will allow creating further applications reusing the same backend and JSON-RPC 2.0 API when further research or collaborations requires it.

6.3.6 Data analysis

On-the-fly and standard analyses are provided. Analysis of selected clones is performed on-the-fly for CDR3 length, SHM, and V/J gene frequencies. Standard analysis across datasets like the percentage of public clones (shared CDR3s in different datasets) and network analysis (Miho et al., 2017) has been performed a priori and is provided as requested.

6.4 Results

6.4.1 Raw datasets

systimsDB (Fig. 6.1) was constructed from 142 in-house generated sequencing experiments resulted in $\approx$ 0.2 billion reads (n=260,529,741) and it is to incorporate 80 public datasets with a total of 7,592 sequencing experiments resulting in $\approx$ 6.5 billion reads (n>6,462,738,099) from 5 species: *Homo sapiens* (n= 65), *Mus musculus* (n=13), *Macaca mulatta* n=(2), *Danio rerio* (n=2) and *Oryctolagus cuniculus* (n=1), of which 58 BCR and 23 TCR datasets.

6.4.2 Preprocessing

Clones were assembled on CDR3 given the importance of this region in the variable domain to determine clonality and antigen-binding [11, 12]. systimsDB has approximately 3 billion CDR3 clones. The standardized preprocessing framework ensured reproducibility of the results.

6.4.3 Web application

The web application is publicly accessible and allows a non-expert user to search the database specifying predetermined filters. Results are shown in a tabular format and columns to be shown can be selected or hidden. Finally, the user can decide to export the results as a tab-separated text file (TSV). This format allows data to be easily loaded in standard spreadsheet applications for further analysis or imported in other databases.

6.4.4 Analysis

Analysis of CDR3 length, SHM, and V/J gene frequencies, public clones and diversity networks is provided for each dataset. Further analysis comparing selected CDR3 clones according to the input provided by the user is performed on-the-fly.

6.5 Discussion

The determination of immune repertoires using high-throughput sequencing has become a crucial tool to understand the immune response, in library preparation and screening, and in clinical practice [27, 26, 226, 229]. The immuno-sequences deluge has created in-house and global unmet needs in terms of data management, storage and analysis. Up-to-date, no database exists and there is not one standard framework for immune repertoire sequencing data. systimsDB, is a database of immune repertoire sequences. It supports the search of CDR3 sequences and characteristics across datasets through a web server, and uses a standardized framework to analyze publicly available immune repertoire data.

With sequencing experiments becoming faster and cost-effective, research laboratories and biopharmaceutical industries are expected to produce increasing volumes of data and only part of this data will be deposited in public repositories. By using the systimsDB structure, research groups and private companies can introduce standardization, limit redundancy of stored data, establish reproducibility in downstream preprocessing and analysis, and increase the reusability and the cross-validation of the data.

Publicly available data from immune repertoires data has revealed to be arduous to access because of the intricate retrieval process (through single scientific articles and the data accession codes) and the necessary computational skills. Although several bioinformatic tools have been developed to analyze and visualize immune repertoires [5] in order to make this data more accessible to the laboratory immunologist, the access and use public data (e.g., for benchmarking purposes or as an in silico control) has been limited to a few (2–3) datasets exploited principally by computational immunologists [68].

Furthermore, if to be used, the data necessitates *de novo* bioinformatic processing and analysis. While accession numbers are now almost regularly stated in literature, the correspondent metadata often suffers from dynamic modification of the experiments and a lack of standardization in reporting experiment and protocol related information. Often, data correspondence to the samples is not self-evident and further computational manipulation or personal contact with the authors of the original publication is necessary. Additionally, raw data are preprocessed with a variety of bioinformatic frameworks, giving rise to a potentially large variation in the analysis results of clones as preprocessed with the different methods. Thus, there is a gap between the generation (deposition) of large amounts of immune repertoire sequencing data, its retrieval and downstream management and analysis.

systimsDB is a specialized database of preprocessed immune repertoire sequencing data (TCR, BCR), that is searchable through a web-interface and allows for the download of selected sub-repertoires in tabular output. The available downstream analysis in systemsDB ensures consistency of methods and reproducibility of results.

Chapter 7

Conclusion

7.1 Relevance of this work

A systems view of the immune systems was first proposed by Nobel laureate Niels K. Jerne more than 40 years ago [29]. Jerne anticipated the transition of immunological research to a systems prospective: “Though the search for mechanisms at sub-cellular, cellular and inter-cellular levels will of course continue, [. . .] emphasis will shift to a structural analysis of the entire immune system.” Nevertheless, systems immunology has developed as a field only recently. Overall, it represents the study of the immune system from a comprehensive and holistic perspective, through integrated mathematical and computational approaches. Specifically, systems analysis of the adaptive immune responses, such as B and T cell repertoires, leverages information-rich data obtained from high-throughput sequencing in order to detect and relate structure and function in complex immune repertoires [24, 27, 25, 225, 5, 285].

Thus, a system-level survey of the adaptive responses has been enabled by the rapid recent advances in high-throughput sequencing technologies, which have resulted in massive amounts of data. The immunosequencing data deluge has introduced significant unresolved computational challenges. Therefore, the transition to a systems analysis of B-cell and T-cell receptor repertoires has implied the unification of immunology with informatics [30]. The development of computational tools and analysis for scalable repertoire data analytics has become a necessity, and so far has frequently constituted the bottleneck to immunological insights at the authentic repertoire-level. For example, the millions of sequences representing an antibody repertoire have limited the practice of network analysis in immunology and the fundamental architecture of the antibody repertoire, defined by the network similarity landscape of its sequences which reflects the spectrum of antigen binding thereby determining immunological protection and function, has long remained unknown. In addition to the indispensable computational advances and the development of scalable tools in order to perform the analysis of large-scale data, new integrative approaches have become necessary to synthesize and transform data into biological insights.

The systems immunology computational tools and analyses developed in this work enable novel insights on the structure and function of antibody repertoires, and generation of new hypothesis (Fig. 2.2, Chapter 2). Specifically:

- Chapter 3: The novel large-scale network analysis and statistical framework revealed the clonal architecture of antibody repertoires from a systems prospective and the structural function of public clones (B-cell receptors that are shared among different individuals). Jerne proposed “the possibility of viewing the immune system as a formal network [. . .]” [29]. This work brings to fruition his vision and opens the way to the science of *network systems immunology*, where nodes can be cells of an immune repertoire, or extend to represent antibodies and antigens in the same network.

- Chapter 4: An open-source novel computational tool, imNet, was developed to construct unprecedentedly large-scale antibody repertoire networks. It enables the deconstruction of the antibody repertoire architecture in entire antibody repertoires of millions of sequences by generating the networks, and thus empowering the development of novel statistical analyses (as shown in Chapter 3). imNet allows to extend the application of network analysis to all sequence-based fields of medical research, from microbiology and virology to proteomics, enabling the generation of new hypothesis and hypothesis testing. Furthermore, imNet is cross-disciplinary, being also useful in social and business sciences for the construction of large-scale string-based networks.

- Chapter 5: Systems analyses (i.e., global sequence characterization, germline gene frequency analysis, evenness profiles of B cell expansion, network analysis, and sequence-based support vector machine learning) were applied to human datasets from uninfected and HIV-1 infected individuals that had or had not developed broadly neutralizing antibodies (bNAbs). The reference open-source database of bNAbs (bNAbs-DB) constructed in order to detect bNAb-specific signatures at a repertoire level and can be analyzed to extract analysis-based other parameters, like the frequency of positioned-based amino acids. The compilation of diverse systems methods allowed not only to compare repertoires in health and disease, but also to detect components within the disease category like the presence of bNAb-like sequences in HIV-1 infection. The collection of various types of systems analyses enabled relating to the power of the different methods in gaining immunological insights at the antibody repertoire level.

- Chapter 6: Considering the existing challenges in generating and collecting immunosequencing data of antibody repertoires (Chapter 3, 5), systimsDB was constructed to enable faster future research of immune repertoires. systimsDB, a database of complementarity determining regions 3 (CDR3) clones from sequenced B and T cell repertoires, enables the immediate testing of new hypothesis building from this large-scale existing data and *a posteriori* integration of systems analyses within and across datasets.

In summary, (i) imNet constructs large-scale networks from biological sequences, (ii) large-scale network analysis reveals that the architecture of antibody repertoires is reproducible, robust and redundant, (iii) machine learning enables the detection of signatures of broadly neutralizing antibodies in HIV-1 individuals, and (iv) systimsDB provides access to systematically annotated immune repertoires.

7.2 Perspective

Systems and network analysis developed and applied here to the antibody repertoire are powerful approaches to answer basic immunology questions and to investigate potential applications in diagnostics, vaccine design and antibody discovery.

7.2.1 Systematically annotated immune repertoire data for benchmarking and repurposing

One immediate application of systimsDB database is to search, filter and subset the database in order to benchmark new data comparing their analysis to the statistics from database samples and experiments (Fig. 7.1). For example, the clonal expansion of B cells in disease is reflected in their uneven profiles [68], where a few clones are very expanded and dominate the repertoire (e.g., HIV-1 infected individuals). In comparison, B cell clonal expansion profiles in health are uniform and clonal frequencies are evenly distributed. systimsDB datasets and samples can serve to benchmark new data, and compare clonal expansion profiles to different healthy individuals, and individuals with diseases (HIV-1 infection, Myasthenia Gravis, etc) or after vaccination. In this regard, systimsDB can represent a first reference and could provide initial cues on the immune repertoires analyzed.

Another example of the future use of the database is as a reference to disease/health associated clones. It is known that sequence-similar immune clones present similar sequence-binding properties. systimsDB presents not only a collection of datasets, but also a collection of clones which can be correlated with the health or disease status.

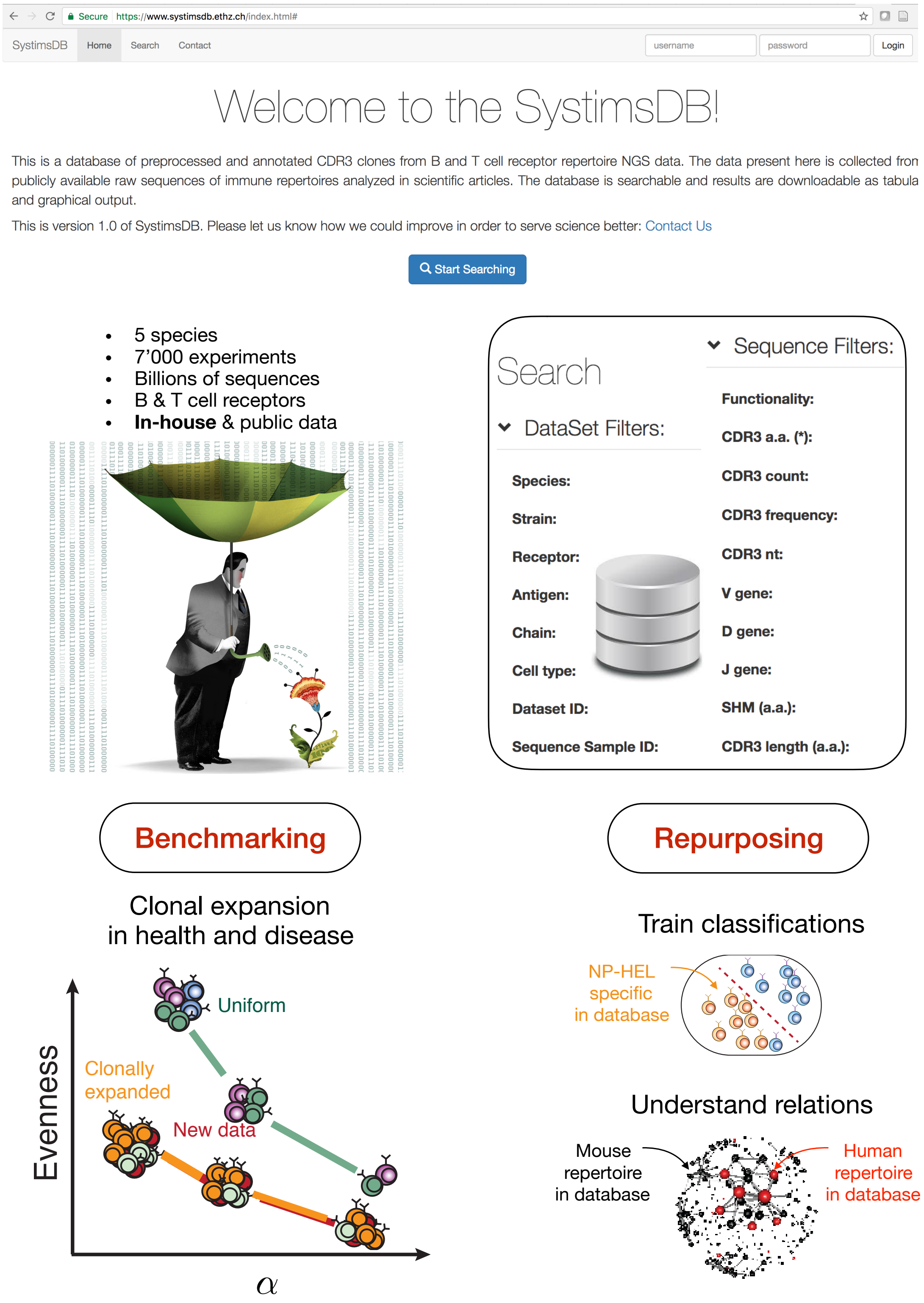
On the other hand, systimsDB opens new research questions. The standard annotation of all datasets reduces preprocessing and analysis bias, and renders the datasets comparable. Thus, machine learning can be applied to classify clones in health or diseases and to detect new features. Another possibility is to detect features of clones specific to a particular sequencing platform, thus providing future ways to predict and correct for platform-specific sequencing errors.

Clonal relations in different species can be further researched in the future. This would bring a further understanding to evolution of clonal differences from a systems-species prospective. Additionally, understanding and characterizing clonal sequence differences, for example between mice and humans, would be useful in clinical research where the development of monoclonal antibodies has developed from murine, to chimeric, humanized and human only sequences because of the dangerous side-effects that non-human monoclonals can exert.

systimsDB provides a basis for new hypothesis generation and enables answering to questions such as: (i) What are the clonal features and signatures of health and disease? How do we compare the network repertoire architecture across individuals beyond condensed network indices, thus without losing information? (iii) What are the cross-species relations of immune clones and repertoires? In prospective, a collection of individual data can support personalized diagnostics and therapeutics.

7.2.2 Fundamental immunology

In this work, novel analysis tools have been developed (imNet) for high-dimensional data and they have been applied to immunology. In particular, large-scale data provides



Welcome to the SystimsDB!

This is a database of preprocessed and annotated CDR3 clones from B and T cell receptor repertoire NGS data. The data present here is collected fron publicly available raw sequences of immune repertoires analyzed in scientific articles. The database is searchable and results are downloadable as tabula and graphical output.

This is version 1.0 of SystimsDB. Please let us know how we could improve in order to serve science better: Contact Us

Fig. 7.1 SystimsDB. Overview and perspective applications. (Note: The *data deluge* illustration first appeared in the Economist on Feb 25th 2010, is part of this figure with permission from the author Brett Ryder).

a means to uncover fundamental principles of the adaptive response, to monitor and probe entire immune repertoires and their specificities. Some fundamental research questions can be answered with the novel systems and network approaches developed here. For example, it is yet unknown what is the architecture and the principles of the T cell repertoire. Although there are studies that represent clusters of related T cells sequences as networks [249, 286], these are constructed on hundreds to thousands of T cells, representing partial repertoires of 2-3 orders of magnitude lower to the number of clones that would robustly represent a repertoire as shown in this work. Thus, the global similarity landscape of the T cell repertoires and their architecture remains unknown. Furthermore, studies regarding the tumor-infiltrating T cell repertoire have become more and more important, as understanding the relations between cancer antigens and adaptive immunity has become critical in discovering and developing effective therapeutics [287]. The methods developed in this work can be of immediate use in understanding the similarity landscape and architecture of T cell repertoires in tumors, thus identifying structurally critical clones or characteristic tumor-specific clonal and network motifs. Because the population of infiltrating T cells has been shown to maintain a surprising large fraction of public clonotypes [287], knowing that public clones are pillars of the architecture of B cell repertoires, the potential structural role of public T cells in the tumor environment could be elucidate by future network studies.

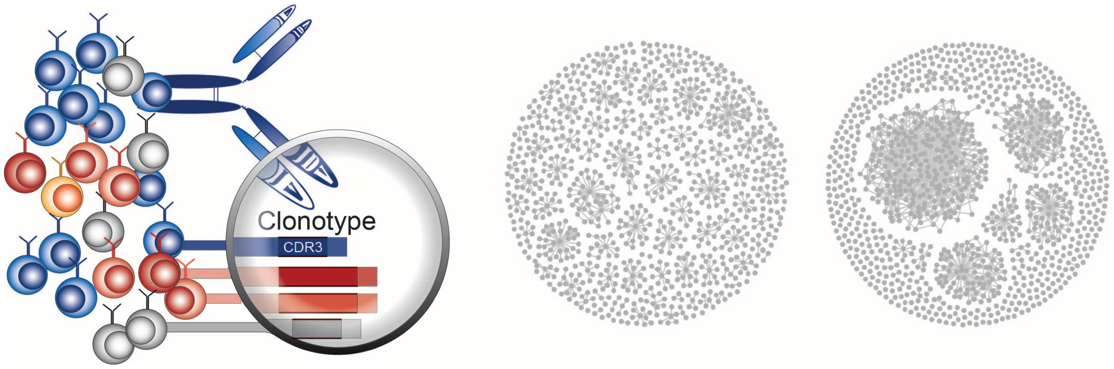


Fig. 7.2 Hypothesis of the architecture of the human antibody repertoire. Is the naïve B cell repertoire evenly distributed? Is the structure of the antibody repertoire star-like?

The direct follow-up of this study, making use also of the public datasets available, can answer the question regarding the architecture of B cell repertoires in humans (Fig. 7.2). At this stage, it is imperative to inquire if human antibody repertoires have the same structure as mouse repertoires. Is the tree-like structure of naïve cells and the star-like structure of antigen-experienced compartments evolutionary conserved in different species? What are the advantages beyond robustness and redundancy to function of these structures? These and similar questions can be answered by leveraging large-scale network construction and applying networks analysis developed here.

Another fundamental immunological question that has not been resolved yet is: Does the antibody (immune) repertoire organize in epitope-specific clusters? Are these clusters reflected in the global architecture of antibody repertoires? Although antigen- and epitope-specificity is one of the major questions in immunology, network analysis and the sequence-space relations and combinations of antibodies and epitopes have not been investigated yet. This work enables the investigation of the adaptive immune response side of the virus-host interaction and can characterize, for the first time, the epitope-specific and virus-specific antibody repertoires from a sequence perspective. This study opens the possibility to study the dynamics of the immune response at an unprecedented detail and allows for further investigation of the prediction of a personalized immune response based on the knowledge of epitope-specific repertoires.

Given a potential experimental pipeline using ([8, 288], Fig. 7.3 A), the following research questions could be addressed using the framework of the network analysis described here:

- What fraction of the epitope-specific antibody repertoire is shared (within a virus)? (Fig. 7.2 A, B)
- How do different virus-specific antibody repertoires relate? (Fig. 7.2 B)
- Do specific repertoires converge in a population? Can we predict specificity and its potential? (Fig. 7.2 C)

7.2.3 Applied Immunology

Antibodies have become therapeutics of choice. There was a recent 100% increase (from 26 mAbs in early 2010 to 52 mAbs in early 2017) in the number of mAbs in Phase 3 clinical studies. And with over 230 mAbs currently in Phase II clinical studies, the commercial mAb therapeutics pipeline should continue to support a steady flow of mAbs from Phase II into Phase III, as those in Phase III advance to regulatory review

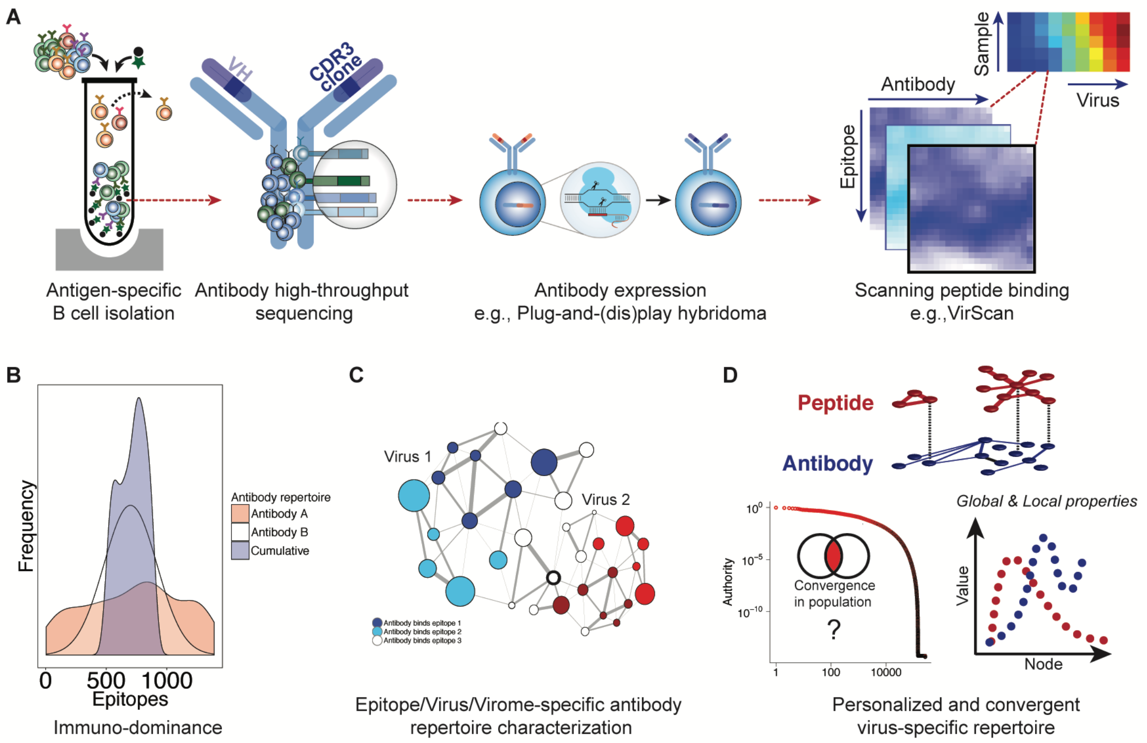

A
VH
CDR3 clone
Sample
Antibody
Virus
Epitope
Antigen-specific B cell isolation
Antibody high-throughput sequencing
Antibody expression e.g., Plug-and-(dis)play hybridoma
Scanning peptide binding e.g.,VirScan
B
Frequency
Antibody repertoire
Antibody A
Antibody B
Cumulative
0
500
1000
Epitopes
Immuno-dominance
C
Virus 1
Virus 2
Antibody binds epitope 1
Antibody binds epitope 2
Antibody binds epitope 3
Epitope/Virus/Virome-specific antibody repertoire characterization
D
Peptide
Antibody
10^{0}
10^{-5}
10^{-10}
Authority
Convergence in population
?
1
100
10000
Global & Local properties
Value
Node
Personalized and convergent virus-specific repertoire

Fig. 7.3 Perspective characterization of epitope-specific antibody structures. (A) Briefly, the potential experimental design for a new study to investigate epitope/antigen-specificity: i) isolation of antigen-specific B cells, ii) high-throughput sequencing of the variable region of single B-cell receptors and iii) expression of the antigen-specific antibody repertoire in a plug-and-(dis)play hybridoma platform [8] in order to secrete the epitope-specific antibodies which will serve as known samples for iv) epitope screening on VirScan [288]. (B,C,D) Systems methods and high-performance computing tools can determine i) the epitope-specific binding profile of each antibody and the cumulative binding profile of the specific antibody repertoire, constructing a framework to detect epitope immuno-dominance, ii) to characterize the epitope/virus/viriome-specific antibody repertoire and iii) to interrogate the correlation of peptide and antibody repertoire network diversity, structure and properties in order to quantify the degree of convergence of these repertoires in the population and determine if specific-repertoires are predictable or can be predicted from the potential repertoire.

and to the market [289].

This increased interest in antibody therapeutics, have given rise to innovative experimental tools in antibody discovery. Over the last 15 years, powerful combinatorial technologies have allowed for the development of in vitro immune repertoires and selection methodologies that can be used to derive antibodies without the need for direct immunization of a living host but by generating antibody diversity from synthetic V genes or cloned from B cells. These synthetic libraries provide information regarding the total number of diverse clones but not not the structure of this diversity and how it is distributed within a library.

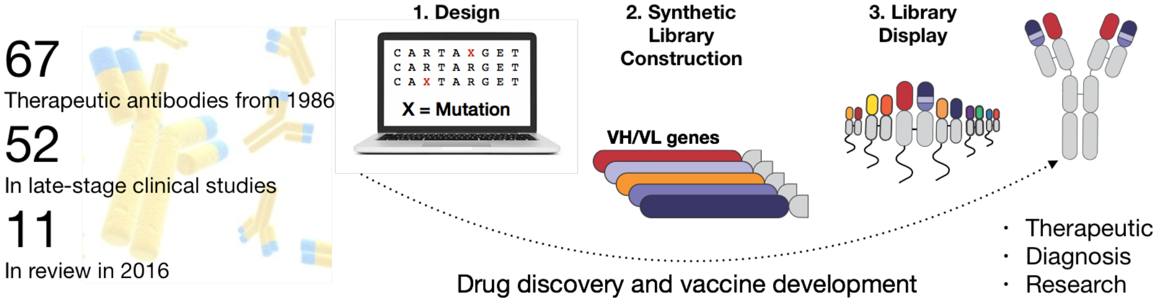


Fig. 7.4 Synthetic repertoires simulating natural repertoires. Synthetic repertoires can be constructed on the basis of the principles of antibody repertoires, thus simulating natural antibody repertoires.

- Can the principles of the architecture of antibody repertoires serve as a blueprint for the construction of synthetic antibody repertoires that simulate humoral immunity for monoclonal antibody discovery and vaccine development?

- The three fundamental principles of the architecture of antibody repertoires uncovered here through network analysis may serve as a blueprint for the construction of synthetic antibody repertoires, which may be used to simulate natural humoral immunity for monoclonal antibody drug discovery and vaccine development.

- In conclusion, network and systems analysis are powerful and adaptive computational tools that can be further leveraged in fundamental and applied immunological research.

References

[1] Victor Greiff, Ulrike Menzel, Enkelejda Miho, Cedric R. Weber, Cook Riedel, Ren, Skylar, Atijeh Valai, Telma Lopes, Andreas Radbruch, Thomas H. Winkler, and Sai T. Reddy. Systems analysis reveals high genetic and antigen-driven predetermination of antibody repertoires throughout b-cell development. *Cell Reports*, 2017.

[2] Enkelejda Miho et al. Detection of broadly neutralizing antibody sequence signatures in hiv-1. *in submission*, 2017.

[3] Jacqueline Parkin and Bryony Cohen. An overview of the immune system. *The Lancet*, 357(9270):1777–1789, 2001.

[4] C. Janeway, M. J. Shlomchik, and Walport. *Immunobiology.* Garland Science, 8 edition, June 2012.

[5] Victor Greiff, Enkelejda Miho, Ulrike Menzel, and Sai T. Reddy. Bioinformatic and statistical analysis of adaptive immune repertoires. *Trends in Immunology*, 36(11), 2015.

[6] David Furman and Mark M Davis. New approaches to understanding the immune response to vaccination and infection. *Vaccine*, 33(40):5271–5281, 2015.

[7] Julie G Burel, Simon H Apte, and Denise L Doolan. Systems approaches towards molecular profiling of human immunity. *Trends in immunology*, 37(1):53–67, 2016.

[8] Mark Pogson, William Kelton, and Sai T Reddy. Microscale technologies for high-throughput analysis of immune cells. In *Microscale Technologies for Cell Engineering*, pages 219–230. Springer, 2016.

[9] Martin F Flajnik and Masanori Kasahara. Origin and evolution of the adaptive immune system: genetic events and selective pressures. *Nature Reviews Genetics*, 11(1):47–59, 2010.

[10] Nobumichi Hozumi and Susumu Tonegawa. Evidence for somatic rearrangement of immunoglobulin genes coding for variable and constant regions. *Proceedings of the National Academy of Sciences*, 73(10):3628–3632, 1976.

[11] Susumu Tonegawa. Somatic generation of antibody diversity. *Nature*, 302(5909):575–581, April 1983.

[12] John L Xu and Mark M Davis. Diversity in the cdr3 region of v h is sufficient for most antibody specificities. *Immunity*, 13(1):37–45, 2000.

[13] Bryan S Briney and James E Crowe Jr. Secondary mechanisms of diversification in the human antibody repertoire. 2013.

[14] Marie-Paule Lefranc, Christelle Pommié, Manuel Ruiz, Véronique Giudicelli, Elodie Foulquier, Lisa Truong, Valérie Thouvenin-Contet, and Gérard Lefranc. Imgt unique numbering for immunoglobulin and t cell receptor variable domains and ig superfamily v-like domains. *Developmental & Comparative Immunology*, 27(1):55–77, 2003.

[15] Joshua A. Weinstein, Ning Jiang, Richard A. White, Daniel S. Fisher, and Stephen R. Quake. High-throughput sequencing of the zebrafish antibody repertoire. *Science*, 324(5928):807–810, May 2009.

[16] Andrew M Collins, Yan Wang, Krishna M Roskin, Christopher P Marquis, and Katherine JL Jackson. The mouse antibody heavy chain repertoire is germline-focused and highly variable between inbred strains. *Phil. Trans. R. Soc. B*, 370(1676):20140236, 2015.

[17] Dmitriy A Bolotin, Stanislav Poslavsky, Igor Mitrophanov, Mikhail Shugay, Ilgar Z Mamedov, Ekaterina V Putintseva, and Dmitriy M Chudakov. Mixcr: software for comprehensive adaptive immunity profiling. *Nat. Methods*, 12(5):380–381, 2015.

[18] Victor Greiff, Ulrike Menzel, Ulrike Haessler, Skylar C Cook, Simon Friedensohn, Tarik A Khan, Mark Pogson, Ina Hellmann, and Sai T Reddy. Quantitative assessment of the robustness of next-generation sequencing of antibody variable gene repertoires from immunized mice. *BMC immunology*, 15(1):40, 2014.

[19] Ulrike Menzel, Victor Greiff, Tarik A Khan, Ulrike Haessler, Ina Hellmann, Simon Friedensohn, Skylar C Cook, Mark Pogson, and Sai T Reddy. Comprehensive evaluation and optimization of amplicon library preparation methods for high-throughput antibody sequencing. *PloS one*, 9(5):e96727, 2014.

[20] Alan S. Perelson. Immune network theory. *Immunological Reviews*, 110(1):5–36, August 1989.

[21] Max D Cooper. The early history of b cells. *Nature Reviews Immunology*, 15(3):191–197, 2015.

[22] Dongni Hou, Cuicui Chen, Eric John Seely, Shujing Chen, and Yuanlin Song. High-throughput sequencing-based immune repertoire study during infectious disease. *Frontiers in Immunology*, 7, 2016.

[23] Jonathan R McDaniel, Brandon J DeKosky, Hidetaka Tanno, Andrew D Ellington, and George Georgiou. Ultra-high-throughput sequencing of the immune receptor repertoire from millions of lymphocytes. *Nature protocols*, 11(3):429–442, 2016.

[24] Jennifer Benichou, Rotem Ben-Hamo, Yoram Louzoun, and Sol Efroni. Rep-seq: uncovering the immunological repertoire through next-generation sequencing. *Immunology*, 135(3):183–191, March 2012.

[25] Harlan Robins. Immunosequencing: applications of immune repertoire deep sequencing. *Current Opinion in Immunology*, 2013.

[26] William H Robinson. Sequencing the functional antibody repertoire [mdash] diagnostic and therapeutic discovery. *Nature Reviews Rheumatology*, 11(3):171–182, 2014.

[27] George Georgiou, Gregory C. Ippolito, John Beausang, Christian E. Busse, Hedda Wardemann, and Stephen R. Quake. The promise and challenge of high-throughput sequencing of the antibody repertoire. *Nature Biotechnology*, advance online publication, January 2014.

[28] Jorg JA Calis and Brad R Rosenberg. Characterizing immune repertoires by high throughput sequencing: strategies and applications. *Trends in immunology*, 35(12):581–590, 2014.

[29] Niels K Jerne. Towards a network theory of the immune system. In *Annales d'immunologie*, volume 125, page 373, 1974.

[30] Brian A. Kidd, Lauren A. Peters, Eric E. Schadt, and Joel T. Dudley. Unifying immunology with informatics and multiscale biology. *Nature Immunology*, 15(2):118–127, February 2014.

[31] Enkelejda Miho, Victor Greiff, Rok Roskar, and Sai T Reddy. The fundamental principles of antibody repertoire architecture revealed by large-scale network analysis. *bioRxiv*, page 124578, 2017.

[32] Enkelejda Miho et al. imnet: software for the generation and analysis of large-scale sequence networks. *in submission*, 2017.

[33] Enkelejda Miho et al. Systimsdb: a database of immune repertoires. *in submission*, 2017.

[34] Victor Greiff, Cédric R Weber, Johannes Palme, Ulrich Bodenhofer, Enkelejda Miho, Ulrike Menzel, and Sai T Reddy. Learning the high-dimensional immunogenomic features that predict public and private antibody repertoires. *The Journal of Immunology*, 2017.

[35] A Yermanos, V Greiff, N Krutler, U Menzel, A Dounas, E Miho, A Oxenius, T Stadler, and ST Reddy. Comparison of methods for phylogenetic b-cell lineage inference using time-resolved antibody repertoire simulations (absim). *Bioinformatics*, 2017.

[36] Merle Schanz, Thomas Liechti, Osvaldo Zagordi, Enkelejda Miho, Sai T Reddy, Huldrych F Günthard, Alexandra Trkola, and Michael Huber. High-throughput sequencing of human immunoglobulin variable regions with subtype identification. *PloS one*, 9(11):e111726, 2014.

[37] Jacob Glanville, Wenwu Zhai, Jan Berka, Dilduz Telman, Gabriella Huerta, Gautam R. Mehta, Irene Ni, Li Mei, Purnima D. Sundar, Giles M. R. Day, David Cox, Arvind Rajpal, and Jaume Pons. Precise determination of the diversity of a combinatorial antibody library gives insight into the human immunoglobulin repertoire. *Proceedings of the National Academy of Sciences*, 106(48):20216–20221, December 2009.

[38] Markus G Rudolph, Robyn L Stanfield, and Ian A Wilson. How tcrs bind mhcs, peptides, and coreceptors. *Annu. Rev. Immunol.*, 24:419–466, 2006.

[39] Uri Hershberg and Eline T Luning Prak. The analysis of clonal expansions in normal and autoimmune b cell repertoires. *Phil. Trans. R. Soc. B*, 370(1676):20140239, 2015.

[40] Xueling Wu, Tongqing Zhou, Jiang Zhu, Baoshan Zhang, Ivelin Georgiev, Charlene Wang, Xuejun Chen, Nancy S. Longo, Mark Louder, Krisha McKee, Sijy O'Dell, Stephen Perfetto, Stephen D. Schmidt, Wei Shi, Lan Wu, Yongping Yang, Zhi-Yong Yang, Zhongjia Yang, Zhenhai Zhang, Mattia Bonsignori, John A. Crump, Saidi H. Kapiga, Noel E. Sam, Barton F. Haynes, Melissa Simek, Dennis R. Burton, Wayne C. Koff, Nicole A. Doria-Rose, Mark Connors, James C. Mullikin, Gary J. Nabel, Mario Roederer, Lawrence Shapiro, Peter D. Kwong, and John R. Mascola. Focused evolution of HIV-1 neutralizing antibodies revealed by structures and deep sequencing. *Science*, 333(6049):1593–1602, September 2011.

[41] Chen Wang, Yi Liu, Lan T Xu, Katherine JL Jackson, Krishna M Roskin, Tho D Pham, Jonathan Laserson, Eleanor L Marshall, Katie Seo, Ji-Yeun Lee, et al. Effects of aging, cytomegalovirus infection, and ebv infection on human b cell repertoires. *The Journal of Immunology*, 192(2):603–611, 2014.

[42] Xueling Wu, Zhenhai Zhang, Chaim A Schramm, M Gordon Joyce, Young Do Kwon, Tongqing Zhou, Zizhang Sheng, Baoshan Zhang, Sijy Oell, Krisha McKee, et al. Maturation and diversity of the vrc01-antibody lineage over 15 years of chronic hiv-1 infection. *Cell*, 161(3):470–485, 2015.

[43] Asaf Madi, Eric Shifrut, Shlomit Reich-Zeliger, Hilah Gal, Katharine Best, Wilfred Ndifon, Benjamin Chain, Irun R Cohen, and Nir 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(10):1603–1612, 2014.

[44] Christopher M Tipton, Christopher F Fucile, Jaime Darce, Asiya Chida, Travis Ichikawa, Ivan Gregoretti, Sandra Schieferl, Jennifer Hom, Scott Jenks, Ron J Feldman, et al. Diversity, cellular origin and autoreactivity of antibody-secreting cell population expansions in acute systemic lupus erythematosus. *Nature immunology*, 16(7):755–765, 2015.

[45] Verena Hehle, Louise D Fraser, Romeeza Tahir, David Kipling, Yu-Chang Wu, Pamela MK Lutalo, John Cason, LeeMeng Choong, David P Druz, Andrew P Cope, et al. Immunoglobulin kappa variable region gene selection during early

human b cell development in health and systemic lupus erythematosus. *Molecular immunology*, 65(2):215–223, 2015.

[46] Paolo A Muraro, Harlan Robins, Sachin Malhotra, Michael Howell, Deborah Phippard, Cindy Desmarais, Alessandra de Paula Alves Sousa, Linda M Griffith, Noha Lim, Richard A Nash, et al. T cell repertoire following autologous stem cell transplantation for multiple sclerosis. *The Journal of clinical investigation*, 124(3):1168–1172, 2014.

[47] Deborah K Dunn-Walters and Alexander A Ademokun. B cell repertoire and ageing. *Current Opinion in Immunology*, 22(4):514–520, August 2010.

[48] Olga V Britanova, Ekaterina V Putintseva, Mikhail Shugay, Ekaterina M Merzlyak, Maria A Turchaninova, Dmitriy B Staroverov, Dmitriy A Bolotin, Sergey Lukyanov, Ekaterina A Bogdanova, Ilgar Z Mamedov, et al. Age-related decrease in tcr repertoire diversity measured with deep and normalized sequence profiling. *The Journal of Immunology*, 192(6):2689–2698, 2014.

[49] H Morbach, EM Eichhorn, JG Liese, and HJ Girschick. Reference values for b cell subpopulations from infancy to adulthood. *Clinical & Experimental Immunology*, 162(2):271–279, 2010.

[50] Vitaly V. Ganusov and Rob J. De Boer. Do most lymphocytes in humans really reside in the gut? *Trends in Immunology*, 28(12):514–518, December 2007.

[51] Donna L Farber, Naomi A Yudanin, and Nicholas P Restifo. Human memory t cells: generation, compartmentalization and homeostasis. *Nature Reviews Immunology*, 14(1):24–35, 2014.

[52] F. Trepel. Number and distribution of lymphocytes in man. a critical analysis. *Journal of Molecular Medicine*, 52(11):511–515, 1974.

[53] Kara M Johnson, Kevin Owen, and Pamela L Witte. Aging and developmental transitions in the b cell lineage. *International immunology*, 14(11):1313–1323, 2002.

[54] Katherine JL Jackson, Yi Liu, Krishna M Roskin, Jacob Glanville, Ramona A Hoh, Katie Seo, Eleanor L Marshall, Thaddeus C Gurley, M Anthony Moody, Barton F Haynes, et al. Human responses to influenza vaccination show seroconversion signatures and convergent antibody rearrangements. *Cell host & microbe*, 16(1):105–114, 2014.

[55] René L Warren, J Douglas Freeman, Thomas Zeng, Gina Choe, Sarah Munro, Richard Moore, John R Webb, and Robert A Holt. Exhaustive t-cell repertoire sequencing of human peripheral blood samples reveals signatures of antigen selection and a directly measured repertoire size of at least 1 million clonotypes. *Genome research*, 21(5):790–797, 2011.

[56] Katherine J. L. Jackson, Marie J. Kidd, Yan Wang, and Andrew M. Collins. The shape of the lymphocyte receptor repertoire: lessons from the b cell receptor. *Frontiers in B Cell Biology*, 4:263, 2013.

[57] Vanessa Venturi, Brian D Rudd, and Miles P Davenport. Specificity, promiscuity, and precursor frequency in immunoreceptors. *Current Opinion in Immunology*, 2013.

[58] M Shugay, DA Bolotin, EV Putintseva, MV Pogorelyy, IZ Mamedov, and DM Chudakov. Huge overlap of individual tcr beta repertoires. *Frontiers in immunology*, 4:466–466, 2012.

[59] Harlan S Robins, Santosh K Srivastava, Paulo V Campregher, Cameron J Turtle, Jessica Andriesen, Stanley R Riddell, Christopher S Carlson, and Edus H Warren. Overlap and effective size of the human cd8+ t cell receptor repertoire. *Science translational medicine*, 2(47):47ra64–47ra64, 2010.

[60] C. Vollmers, R. V. Sit, J. A. Weinstein, C. L. Dekker, and S. R. Quake. Genetic measurement of memory b-cell recall using antibody repertoire sequencing. *Proceedings of the National Academy of Sciences*, July 2013.

[61] Simone Becattini, Daniela Latorre, Federico Mele, Mathilde Foglierini, Corinne De Gregorio, Antonino Cassotta, Blanca Fernandez, Sander Kelderman, Ton N Schumacher, Davide Corti, et al. Functional heterogeneity of human memory cd4+ t cell clones primed by pathogens or vaccines. *Science*, 347(6220):400–406, 2015.

[62] Rachael JM Bashford-Rogers, Anne L Palser, Saad F Idris, Lisa Carter, Michael Epstein, Robin E Callard, Daniel C Douek, George S Vassiliou, George A Follows, Mike Hubank, et al. Capturing needles in haystacks: a comparison of b-cell receptor sequencing methods. *BMC immunology*, 15(1):29, 2014.

[63] Nicholas J Gotelli and Robert K Colwell. Quantifying biodiversity: procedures and pitfalls in the measurement and comparison of species richness. *Ecology letters*, 4(4):379–391, 2001.

[64] Cornelia Lindner, Irene Thomsen, Benjamin Wahl, Milas Ugur, Maya K Sethi, Michaela Friedrichsen, Anna Smoczek, Stephan Ott, Ulrich Baumann, Sebastian Suerbaum, et al. Diversification of memory b cells drives the continuous adaptation of secretory antibodies to gut microbiota. *Nature immunology*, 16(8):880–888, 2015.

[65] Nitya Nair, Evan W Newell, Christopher Vollmers, Stephen R Quake, John M Morton, Mark M Davis, Xiao-Song He, and Harry B Greenberg. High-dimensional immune profiling of total and rotavirus vp6-specific intestinal and circulating b cells by mass cytometry. *Mucosal immunology*, 9(1):68–82, 2015.

[66] Megan Estorninho, Vivienne B. Gibson, Deborah Kronenberg-Versteeg, Yuk-Fun Liu, Chester Ni, Karen Cerosaletti, and Mark Peakman. A novel approach to tracking antigen-experienced CD4 t cells into functional compartments via tandem deep and shallow TCR clonotyping. *The Journal of Immunology*, page 1300622, October 2013.

[67] Vanessa Venturi, Máire F Quigley, Hui Yee Greenaway, Pauline C Ng, Zachary S Ende, Tina McIntosh, Tedi E Asher, Jorge R Almeida, Samuel Levy, David A Price, et al. A mechanism for tcr sharing between t cell subsets and individuals revealed by pyrosequencing. *The Journal of Immunology*, 186(7):4285–4294, 2011.

[68] Victor Greiff, Pooja Bhat, Skylar C Cook, Ulrike Menzel, Wenjing Kang, and Sai T Reddy. A bioinformatic framework for immune repertoire diversity profiling enables detection of immunological status. *Genome medicine*, 7(1):49, 2015.

[69] Nicholas J. Loman, Raju V. Misra, Timothy J. Dallman, Chrystala Constantinidou, Saheer E. Gharbia, John Wain, and Mark J. Pallen. Performance comparison of benchtop high-throughput sequencing platforms. *Nature Biotechnology*, 30(5):434–439, 2012.

[70] Cassandra B. Jabara, Corbin D. Jones, Jeffrey Roach, Jeffrey A. Anderson, and Ronald Swanstrom. Accurate sampling and deep sequencing of the HIV-1 protease gene using a primer ID. *Proceedings of the National Academy of Sciences*, 108(50):20166–20171, December 2011.

[71] Mikhail Shugay, Olga V Britanova, Ekaterina M Merzlyak, Maria A Turchaninova, Ilgar Z Mamedov, Timur R Tuganbaev, Dmitriy A Bolotin, Dmitry B Staroverov, Ekaterina V Putintseva, Karla Plevova, et al. Towards error-free profiling of immune repertoires. *Nature methods*, 11(6):653–655, 2014.

[72] Isaac Kinde, Jian Wu, Nick Papadopoulos, Kenneth W. Kinzler, and Bert Vogelstein. Detection and quantification of rare mutations with massively parallel sequencing. *Proceedings of the National Academy of Sciences*, 108(23):9530–9535, June 2011.

[73] Jason A Vander Heiden, Gur Yaari, Mohamed Uduman, Joel NH Stern, Kevin C Oonnor, David A Hafler, Francois Vigneault, and Steven H Kleinstein. presto: a toolkit for processing high-throughput sequencing raw reads of lymphocyte receptor repertoires. *Bioinformatics*, page btu138, 2014.

[74] Evgeny S Egorov, Ekaterina M Merzlyak, Andrew A Shelenkov, Olga V Britanova, George V Sharonov, Dmitriy B Staroverov, Dmitriy A Bolotin, Alexey N Davydov, Ekaterina Barsova, Yuriy B Lebedev, et al. Quantitative profiling of immune repertoires for minor lymphocyte counts using unique molecular identifiers. *The Journal of Immunology*, 194(12):6155–6163, 2015.

[75] Bryan Howie, Anna M Sherwood, Ashley D Berkebile, Jan Berka, Ryan O Emerson, David W Williamson, Ilan Kirsch, Marissa Vignali, Mark J Rieder, Christopher S Carlson, et al. High-throughput pairing of t cell receptor α and β sequences. *Science Translational Medicine*, 7(301):301ra131–301ra131, 2015.

[76] Christian E. Busse, Irina Czogiel, Peter Braun, Peter F. Arndt, and Hedda Wardemann. Single-cell based high-throughput sequencing of full-length immunoglobulin heavy and light chain genes. *European Journal of Immunology*, page n/a–n/a, 2013.

[77] Claire T Deakin, Jeffrey J Deakin, Samantha L Ginn, Paul Young, David Humphreys, Catherine M Suter, Ian E Alexander, and Claus V Hallwirth. Impact of next-generation sequencing error on analysis of barcoded plasmid libraries of known complexity and sequence. *Nucleic acids research*, 42(16):e129–e129, 2014.

[78] Long V Nguyen, Maisam Makarem, Annaick Carles, Michelle Moksa, Nagarajan Kannan, Pawan Pandoh, Peter Eirew, Tomo Osako, Melanie Kardel, Alice MS Cheung, et al. Clonal analysis via barcoding reveals diverse growth and differentiation of transplanted mouse and human mammary stem cells. *Cell Stem Cell*, 14(2):253–263, 2014.

[79] Simon Andrews et al. Fastqc: a quality control tool for high throughput sequence data, 2010.

[80] Dmitry A. Bolotin, Ilgar Z. Mamedov, Olga V. Britanova, Ivan V. Zvyagin, Dmitriy Shagin, Svetlana V. Ustyugova, Maria A. Turchaninova, Sergey Lukyanov, Yury B. Lebedev, and Dmitriy M. Chudakov. Next generation sequencing for TCR repertoire profiling: Platform-specific features and correction algorithms. *European Journal of Immunology*, 42(11):3073–3083, 2012.

[81] Andre P. Masella, Andrea K. Bartram, Jakub M. Truszkowski, Daniel G. Brown, and Josh D. Neufeld. PANDAseq: paired-end assembler for illumina sequences. *BMC Bioinformatics*, 13(1):31, February 2012.

[82] Dmitriy A. Bolotin, Mikhail Shugay, Ilgar Z. Mamedov, Ekaterina V. Putintseva, Maria A. Turchaninova, Ivan V. Zvyagin, Olga V. Britanova, and Dmitriy M. Chudakov. MiTCR: software for t-cell receptor sequencing data analysis. *Nature methods*, 2013.

[83] Sheng Li, Paweł P Łabaj, Paul Zumbo, Peter Sykacek, Wei Shi, Leming Shi, John Phan, Po-Yen Wu, May Wang, Charles Wang, et al. Detecting and correcting systematic variation in large-scale rna sequencing data. *Nature biotechnology*, 32(9):888–895, 2014.

[84] Seqc/Maqc-Iii Consortium et al. A comprehensive assessment of rna-seq accuracy, reproducibility and information content by the sequencing quality control consortium. *Nature biotechnology*, 32(9):903–914, 2014.

[85] Maryam B Yassai, Yuri N Naumov, Elena N Naumova, and Jack Gorski. A clonotype nomenclature for t cell receptors. *Immunogenetics*, 61(7):493–502, 2009.

[86] Rui Chen and Michael Snyder. Yeast proteomics and protein microarrays. *Journal of proteomics*, 73(11):2147–2157, October 2010.

[87] Wentian Li, Franak Batliwalla, and Thomas L. Rothstein. Human b-1 cells are not preplasmablasts: analysis of microarray data and other issues. *Blood*, 122(22):3691–3693, November 2013.

[88] Robert C Edgar. Search and clustering orders of magnitude faster than blast. *Bioinformatics*, 26(19):2460–2461, 2010.

[89] Limin Fu, Beifang Niu, Zhengwei Zhu, Sitao Wu, and Weizhong Li. Cd-hit: accelerated for clustering the next-generation sequencing data. *Bioinformatics*, 28(23):3150–3152, 2012.

[90] Jiang Zhu, Xueling Wu, Baoshan Zhang, Krisha McKee, Sijy O'Dell, Cinque Soto, Tongqing Zhou, Joseph P. Casazza, James C. Mullikin, Peter D. Kwong, John R. Mascola, Lawrence Shapiro, Jesse Becker, Betty Benjamin, Robert Blakesley, Gerry Bouffard, Shelise Brooks, Holly Coleman, Mila Dekhtyar, Michael Gregory, Xiaobin Guan, Jyoti Gupta, Joel Han, April Hargrove, Shi-ling Ho, Taccara Johnson, Richelle Legaspi, Sean Lovett, Quino Maduro, Cathy Masiello, Baishali Maskeri, Jenny McDowell, Casandra Montemayor, James Mullikin, Morgan Park, Nancy Riebow, Karen Schandler, Brian Schmidt, Christina Sison, Mal Stantripop, James Thomas, Pam Thomas, Meg Vemulapalli, and Alice Young. De novo identification of VRC01 class HIV-1–neutralizing antibodies by next-generation sequencing of b-cell transcripts. *Proceedings of the National Academy of Sciences*, page 201306262, October 2013.

[91] J. Zhu, G. Ofek, Y. Yang, B. Zhang, M. K. Louder, G. Lu, K. McKee, M. Pancera, J. Skinner, Z. Zhang, R. Parks, J. Eudailey, K. E. Lloyd, J. Blinn, S. M. Alam, B. F. Haynes, M. Simek, D. R. Burton, W. C. Koff, NISC Comparative Sequencing Program, J. C. Mullikin, J. R. Mascola, L. Shapiro, P. D. Kwong, J. Becker, B. Benjamin, R. Blakesley, G. Bouffard, S. Brooks, H. Coleman, M. Dekhtyar, M. Gregory, X. Guan, J. Gupta, J. Han, A. Hargrove, S.-l. Ho, T. Johnson, R. Legaspi, S. Lovett, Q. Maduro, C. Masiello, B. Maskeri, J. McDowell, C. Montemayor, J. Mullikin, M. Park, N. Riebow, K. Schandler, B. Schmidt, C. Sison, M. Stantripop, J. Thomas, P. Thomas, M. Vemulapalli, and A. Young. Mining the antibodyome for HIV-1-neutralizing antibodies with next-generation sequencing and phylogenetic pairing of heavy/light chains. *Proceedings of the National Academy of Sciences*, March 2013.

[92] Hua-Xin Liao, Rebecca Lynch, Tongqing Zhou, Feng Gao, S. Munir Alam, Scott D. Boyd, Andrew Z. Fire, Krishna M. Roskin, Chaim A. Schramm, Zhenhai Zhang, Jiang Zhu, Lawrence Shapiro, Jesse Becker, Betty Benjamin, Robert Blakesley, Gerry Bouffard, Shelise Brooks, Holly Coleman, Mila Dekhtyar, Michael Gregory, Xiaobin Guan, Jyoti Gupta, Joel Han, April Hargrove, Shi-ling Ho, Taccara Johnson, Richelle Legaspi, Sean Lovett, Quino Maduro, Cathy Masiello, Baishali Maskeri, Jenny McDowell, Casandra Montemayor, James Mullikin, Morgan Park, Nancy Riebow, Karen Schandler, Brian Schmidt, Christina Sison, Mal Stantripop, James Thomas, Pam Thomas, Meg Vemulapalli, Alice Young, James C. Mullikin, S. Gnanakaran, Peter Hraber, Kevin Wiehe, Garnett Kelsoe, Guang Yang, Shi-Mao Xia, David C. Montefiori, Robert Parks, Krissey E. Lloyd, Richard M. Scearce, Kelly A. Soderberg, Myron Cohen, Gift Kamanga, Mark K. Louder, Lillian M. Tran, Yue Chen, Fangping Cai, Sheri Chen, Stephanie Moquin, Xiulian Du, M. Gordon Joyce, Sanjay Srivatsan, Baoshan Zhang, Anqi Zheng, George M. Shaw, Beatrice H. Hahn, Thomas B. Kepler, Bette T. M. Korber, Peter D. Kwong, John R. Mascola, and Barton F. Haynes. Co-evolution of a broadly neutralizing HIV-1 antibody and founder virus. *Nature*, April 2013.

[93] Devin Sok, Uri Laserson, Jonathan Laserson, Yi Liu, Francois Vigneault, Jean-Philippe Julien, Bryan Briney, Alejandra Ramos, Karen F. Saye, Khoa Le, Alison

Mahan, Shenshen Wang, Mehran Kardar, Gur Yaari, Laura M. Walker, Birgitte B. Simen, Elizabeth P. St. John, Po-Ying Chan-Hui, Kristine Swiderek, Stephen H. Kleinstein, Galit Alter, Michael S. Seaman, Arup K. Chakraborty, Daphne Koller, Ian A. Wilson, George M. Church, Dennis R. Burton, and Pascal Poignard. The effects of somatic hypermutation on neutralization and binding in the PGT121 family of broadly neutralizing HIV antibodies. *PLoS Pathog*, 9(11):e1003754, November 2013.

[94] Rachael Bashford-Rogers, Anne Palser, Brian Huntly, Richard Rance, George Vassiliou, George Follows, and Paul Kellam. Network properties derived from deep sequencing of the human b-cell receptor repertoires delineates b-cell populations. *Genome Research*, June 2013.

[95] Kenneth B Hoehn, Astrid Gall, Rachael Bashford-Rogers, SJ Fidler, S Kaye, JN Weber, MO McClure, Paul Kellam, Oliver G Pybus, SPARTAC Trial Investigators, et al. Dynamics of immunoglobulin sequence diversity in hiv-1 infected individuals. *Phil. Trans. R. Soc. B*, 370(1676):20140241, 2015.

[96] Uri Laserson, Francois Vigneault, Daniel Gadala-Maria, Gur Yaari, Mohamed Uduman, Jason A. Vander Heiden, William Kelton, Sang Taek Jung, Yi Liu, Jonathan Laserson, Raj Chari, Je-Hyuk Lee, Ido Bachelet, Brendan Hickey, Erez Lieberman-Aiden, Bozena Hanczaruk, Birgitte B. Simen, Michael Egholm, Daphne Koller, George Georgiou, Steven H. Kleinstein, and George M. Church. High-resolution antibody dynamics of vaccine-induced immune responses. *Proceedings of the National Academy of Sciences*, page 201323862, March 2014.

[97] Christopher Sundling, Zhenhai Zhang, Ganesh E. Phad, Zizhang Sheng, Yimeng Wang, John R. Mascola, Yuxing Li, Richard T. Wyatt, Lawrence Shapiro, and Gunilla B. Karlsson Hedestam. Single-cell and deep sequencing of IgG-Switched macaque b cells reveal a diverse ig repertoire following immunization. *The Journal of Immunology*, page 1303334, March 2014.

[98] Jiang Zhu, Sijy Oell, Gilad Ofek, Marie Pancera, Xueling Wu, Baoshan Zhang, Zhenhai Zhang, James C Mullikin, Melissa Simek, Dennis R Burton, et al. Somatic populations of pgt135–137 hiv-1-neutralizing antibodies identified by 454 pyrosequencing and bioinformatics. *Frontiers in microbiology*, 3, 2012.

[99] Chen Wang, Yi Liu, Mary M Cavanagh, Sabine Le Saux, Qian Qi, Krishna M Roskin, Timothy J Looney, Ji-Yeun Lee, Vaishali Dixit, Cornelia L Dekker, et al. B-cell repertoire responses to varicella-zoster vaccination in human identical twins. *Proceedings of the National Academy of Sciences*, 112(2):500–505, 2015.

[100] Joseph R Francica, Zizhang Sheng, Zhenhai Zhang, Yoshiaki Nishimura, Masashi Shingai, Akshaya Ramesh, Brandon F Keele, Stephen D Schmidt, Barbara J Flynn, Sam Darko, et al. Analysis of immunoglobulin transcripts and hypermutation following shivad8 infection and protein-plus-adjuvant immunization. *Nature communications*, 6, 2015.

[101] Yana Safonova, Alla Lapidus, and Jennie Lill. Igsimulator: a versatile immunosequencing simulator. *Bioinformatics*, page btv326, 2015.

[102] Daniel Gadala-Maria, Gur Yaari, Mohamed Uduman, and Steven H Kleinstein. Automated analysis of high-throughput b-cell sequencing data reveals a high frequency of novel immunoglobulin v gene segment alleles. *Proceedings of the National Academy of Sciences*, 112(8):E862–E870, 2015.

[103] Jason J. Lavinder, Yariv Wine, Claudia Giesecke, Gregory C. Ippolito, Andrew P. Horton, Oana I. Lungu, Kam Hon Hoi, Brandon J. DeKosky, Ellen M. Murrin, Megan M. Wirth, Andrew D. Ellington, Thomas Dörner, Edward M. Marcotte, Daniel R. Boutz, and George Georgiou. Identification and characterization of the constituent human serum antibodies elicited by vaccination. *Proceedings of the National Academy of Sciences*, page 201317793, January 2014.

[104] Henry S. Horn. Measurement of "Overlap" in comparative ecological studies. *The American Naturalist*, 100(914):419–424, September 1966.

[105] Grzegorz A. Rempala and Michal Seweryn. Methods for diversity and overlap analysis in t-cell receptor populations. *Journal of Mathematical Biology*, pages 1–30, 2013.

[106] Martin Krzywinski, Jacqueline Schein, Inanc Birol, Joseph Connors, Randy Gascoyne, Doug Horsman, Steven J Jones, and Marco A Marra. Circos: an information aesthetic for comparative genomics. *Genome research*, 19(9):1639–1645, 2009.

[107] Gabor Csardi and Tamas Nepusz. The igraph software package for complex network research. *InterJournal, Complex Systems*, 1695(5):1–9, 2006.

[108] Michal Barak, Neta S. Zuckerman, Hanna Edelman, Ron Unger, and Ramit Mehr. IgTree©: creating immunoglobulin variable region gene lineage trees. *Journal of Immunological Methods*, 338(1–2):67–74, September 2008.

[109] Mathieu Bastian, Sebastien Heymann, Mathieu Jacomy, et al. Gephi: an open source software for exploring and manipulating networks. *ICWSM*, 8:361–362, 2009.

[110] Ning Jiang, Jiankui He, Joshua A Weinstein, Lolita Penland, Sanae Sasaki, Xiao-Song He, Cornelia L Dekker, Nai-Ying Zheng, Min Huang, Meghan Sullivan, et al. Lineage structure of the human antibody repertoire in response to influenza vaccination. *Science translational medicine*, 5(171):171ra19–171ra19, 2013.

[111] Roberto Di Niro, Seung-Joo Lee, Jason A Vander Heiden, Rebecca A Elsner, Nikita Trivedi, Jason M Bannock, Namita T Gupta, Steven H Kleinstein, Francois Vigneault, Tamara J Gilbert, et al. Salmonella infection drives promiscuous b cell activation followed by extrafollicular affinity maturation. *Immunity*, 43(1):120–131, 2015.

[112] J Felsenstein. Phylip, version 3.6: phylogeny inference package. *Cladistics*, 5:164–166, 1989.

[113] Liam J Revell and Scott A Chamberlain. Rphylip: an r interface for phylip. *Methods in Ecology and Evolution*, 5(9):976–981, 2014.

[114] Daniel H Huson, Daniel C Richter, Christian Rausch, Tobias Dezulian, Markus Franz, and Regula Rupp. Dendroscope: An interactive viewer for large phylogenetic trees. *BMC bioinformatics*, 8(1):460, 2007.

[115] Meriem Attaf, Eric Huseby, and Andrew K Sewell. $\alpha\beta$ t cell receptors as predictors of health and disease. *Cellular & molecular immunology*, 12(4):391–399, 2015.

[116] Anne E. Magurran. *Ecological Diversity and Its Measurement.* Princeton University Press, November 1988.

[117] Bernardo Cortina-Ceballos, Elizabeth Ernestina Godoy-Lozano, Hugo Sámano-Sánchez, Andrés Aguilar-Salgado, Martín Del Castillo Velasco-Herrera, Carlos Vargas-Chávez, Daniel Velázquez-Ramírez, Guillermo Romero, José Moreno, Juan Téllez-Sosa, et al. Reconstructing and mining the b cell repertoire with immunediversity. In *MAbs*, volume 7, pages 516–524. Taylor & Francis, 2015.

[118] Namita T Gupta, Jason A Vander Heiden, Mohamed Uduman, Daniel Gadala-Maria, Gur Yaari, and Steven H Kleinstein. Change-o: a toolkit for analyzing large-scale b cell immunoglobulin repertoire sequencing data. *Bioinformatics*, 31(20):3356–3358, 2015.

[119] Philip Dixon and MW Palmer. Vegan, a package of r functions for community ecology. *Journal of Vegetation Science*, 14(6):927–930, 2003.

[120] Vadim I Nazarov, Mikhail V Pogorelyy, Ekaterina A Komech, Ivan V Zvyagin, Dmitry A Bolotin, Mikhail Shugay, Dmitry M Chudakov, Yury B Lebedev, and Ilgar Z Mamedov. tcr: an r package for t cell receptor repertoire advanced data analysis. *BMC bioinformatics*, 16(1):175, 2015.

[121] M. O. Hill. Diversity and evenness: A unifying notation and its consequences. *Ecology*, 54(2):427–432, March 1973.

[122] Daniel L. Solomon, Cornell University Biometrics Unit, Cornell University Dept of Biometrics, Cornell University Dept of Biological Statistics Biology, and Computational. Biometrics unit technical reports: Number BU-573-M: a comparative approach to species diversity. 1975.

[123] Omri Snir, Luka Mesin, Moriah Gidoni, Knut EA Lundin, Gur Yaari, and Ludvig M Sollid. Analysis of celiac disease autoreactive gut plasma cells and their corresponding memory compartment in peripheral blood using high-throughput sequencing. *The Journal of Immunology*, 194(12):5703–5712, 2015.

[124] Anne Chao and Tsung-Jen Shen. Nonparametric estimation of shannon's index of diversity when there are unseen species in sample. *Environmental and ecological statistics*, 10(4):429–443, 2003.

[125] Lou Jost. The relation between evenness and diversity. *Diversity*, 2(2):207–232, February 2010.

[126] Thierry Mora, Aleksandra M. Walczak, William Bialek, and Curtis G. Callan. Maximum entropy models for antibody diversity. *Proceedings of the National Academy of Sciences*, 107(12):5405–5410, March 2010.

[127] David J Schwab, Ilya Nemenman, and Pankaj Mehta. Zipf law and criticality in multivariate data without fine-tuning. *Physical review letters*, 113(6):068102, 2014.

[128] Olesya V Bolkhovskaya, Daniil Yu Zorin, and Mikhail V Ivanchenko. Assessing t cell clonal size distribution: a non-parametric approach. *PloS one*, 9(10):e108658, 2014.

[129] Daniel J Laydon, Anat Melamed, Aaron Sim, Nicolas A Gillet, Kathleen Sim, Sam Darko, J Simon Kroll, Daniel C Douek, David A Price, Charles RM Bangham, et al. Quantification of htlv-1 clonality and tcr diversity. *PLoS Comput Biol*, 10(6):e1003646, 2014.

[130] Daniel J Laydon, Charles RM Bangham, and Becca Asquith. Estimating t-cell repertoire diversity: limitations of classical estimators and a new approach. *Phil. Trans. R. Soc. B*, 370(1675):20140291, 2015.

[131] I. J. Good. The population frequencies of species and the estimation of population parameters. *Biometrika*, 40(3-4):237–264, December 1953.

[132] Scott D. Boyd, Eleanor L. Marshall, Jason D. Merker, Jay M. Maniar, Lyndon N. Zhang, Bita Sahaf, Carol D. Jones, Birgitte B. Simen, Bozena Hanczaruk, Khoa D. Nguyen, Kari C. Nadeau, Michael Egholm, David B. Miklos, James L. Zehnder, and Andrew Z. Fire. Measurement and clinical monitoring of human lymphocyte clonality by massively parallel v-d-j pyrosequencing. *Science Translational Medicine*, 1(12):12ra23, December 2009.

[133] Krishna M Roskin, Noa Simchoni, Yi Liu, Ji-Yeun Lee, Katie Seo, Ramona A Hoh, Tho Pham, Joon H Park, David Furman, Cornelia L Dekker, et al. Igh sequences in common variable immune deficiency reveal altered b cell development and selection. *Science translational medicine*, 7(302):302ra135–302ra135, 2015.

[134] Veronika I Zarnitsyna, Brian D Evavold, Louis N Schoettle, Joseph N Blattman, and Rustom Antia. Estimating the diversity, completeness, and cross-reactivity of the t cell repertoire. 2013.

[135] Gur Yaari, Jason Vander Heiden, Mohamed Uduman, Daniel Gadala-Maria, Namita Gupta, Joel N. Stern, Kevin O'Connor, David Hafler, Uri Laserson, Francois Vigneault, and Steven Kleinstein. Models of somatic hypermutation targeting and substitution based on synonymous mutations from high-throughput immunoglobulin sequencing data. *Frontiers in B Cell Biology*, 4::358, 2013.

[136] Florian Klein, Ron Diskin, JohannesF. Scheid, Christian Gaebler, Hugo Mouquet, Ivelin S. Georgiev, Marie Pancera, Tongqing Zhou, Reha-Baris Incesu, BrooksZhongzheng Fu, PriyanthiN.P. Gnanapragasam, ThiagoY. Oliveira, MichaelS. Seaman, PeterD. Kwong, PamelaJ. Bjorkman, and MichelC. Nussenzweig. Somatic mutations of the immunoglobulin framework are generally required for broad and potent HIV-1 neutralization. *Cell*, 153(1):126–138, March 2013.

[137] Damien Chaussabel. Assessment of immune status using blood transcriptomics and potential implications for global health. In *Seminars in immunology*, volume 27, pages 58–66. Elsevier, 2015.

[138] Poornima Parameswaran, Yi Liu, KrishnaM. Roskin, KatherineK.L. Jackson, VaishaliP. Dixit, Ji-Yeun Lee, Karen L. Artiles, Simona Zompi, MariaJosé Vargas, BirgitteB. Simen, Bozena Hanczaruk, KimR. McGowan, MuhammadA. Tariq, Nader Pourmand, Daphne Koller, Angel Balmaseda, ScottD. Boyd, Eva Harris, and AndrewZ. Fire. Convergent antibody signatures in human dengue. *Cell Host & Microbe*, 13(6):691–700, June 2013.

[139] Fabio Luciani, Rowena A Bull, and Andrew R Lloyd. Next generation deep sequencing and vaccine design: today and tomorrow. *Trends in biotechnology*, 30(9):443–452, 2012.

[140] Sai T Reddy, Xin Ge, Aleksandr E Miklos, Randall A Hughes, Seung Hyun Kang, Kam Hon Hoi, Constantine Chrysostomou, Scott P Hunicke-Smith, Brent L Iverson, Philip W Tucker, Andrew D Ellington, and George Georgiou. Monoclonal antibodies isolated without screening by analyzing the variable-gene repertoire of plasma cells. *Nat Biotech*, 28(9):965–969, 2010.

[141] Wan Cheung Cheung, Sean A. Beausoleil, Xiaowu Zhang, Shuji Sato, Sandra M. Schieferl, James S. Wieler, Jason G. Beaudet, Ravi K. Ramenani, Lana Popova, Michael J. Comb, John Rush, and Roberto D. Polakiewicz. A proteomics approach for the identification and cloning of monoclonal antibodies from serum. *Nature Biotechnology*, 30(5):447–452, 2012.

[142] Joseph Kaplinsky, Anthony Li, Amy Sun, Maryaline Coffre, Sergei B Koralov, and Ramy Arnaout. Antibody repertoire deep sequencing reveals antigen-independent selection in maturing b cells. *Proceedings of the National Academy of Sciences*, 111(25):E2622–E2629, 2014.

[143] Wei Shi, Yang Liao, Simon N Willis, Nadine Taubenheim, Michael Inouye, David M Tarlinton, Gordon K Smyth, Philip D Hodgkin, Stephen L Nutt, and Lynn M Corcoran. Transcriptional profiling of mouse b cell terminal differentiation defines a signature for antibody-secreting plasma cells. *Nature immunology*, 16(6):663–673, 2015.

[144] Alan S. Perelson and George F. Oster. Theoretical studies of clonal selection: Minimal antibody repertoire size and reliability of self-non-self discrimination. *Journal of Theoretical Biology*, 81(4):645–670, December 1979.

[145] Anand Murugan, Thierry Mora, Aleksandra M. Walczak, and Curtis G. Callan. Statistical inference of the generation probability of t-cell receptors from sequence repertoires. *Proceedings of the National Academy of Sciences*, 109(40):16161–16166, October 2012.

[146] Yuval Elhanati, Zachary Sethna, Quentin Marcou, Curtis G Callan, Thierry Mora, and Aleksandra M Walczak. Inferring processes underlying b-cell repertoire diversity. *Phil. Trans. R. Soc. B*, 370(1676):20140243, 2015.

[147] Ivan V. Zvyagin, Mikhail V. Pogorelyy, Marina E. Ivanova, Ekaterina A. Komech, Mikhail Shugay, Dmitry A. Bolotin, Andrey A. Shelenkov, Alexey A. Kurnosov, Dmitriy B. Staroverov, Dmitriy M. Chudakov, Yuri B. Lebedev, and Ilgar Z. Mamedov. Distinctive properties of identical twins' TCR repertoires revealed by

high-throughput sequencing. *Proceedings of the National Academy of Sciences*, page 201319389, April 2014.

[148] Yuval Elhanati, Anand Murugan, Curtis G Callan, Thierry Mora, and Aleksandra M Walczak. Quantifying selection in immune receptor repertoires. *Proceedings of the National Academy of Sciences*, 111(27):9875–9880, 2014.

[149] Mattia Bonsignori, Hua-Xin Liao, Feng Gao, Wilton B Williams, S Munir Alam, David C Montefiori, and Barton F Haynes. Antibody-virus co-evolution in hiv infection: paths for hiv vaccine development. *Immunological Reviews*, 275(1):145–160, 2017.

[150] Pedro Romero, Jacques Banchereau, Nina Bhardwaj, Mark Cockett, Mary L Disis, Glenn Dranoff, Eli Gilboa, Scott A Hammond, Robert Hershberg, Alan J Korman, et al. The human vaccines project: A roadmap for cancer vaccine development. *Science translational medicine*, 8(334):334ps9–334ps9, 2016.

[151] Rotem Ben-Hamo and Sol Efroni. The whole-organism heavy chain b cell repertoire from zebrafish self-organizes into distinct network features. *BMC systems biology*, 5(1):27, 2011.

[152] Ya-Hui Chang, Hui-Chung Kuan, TC Hsieh, KH Ma, Chung-Hsiang Yang, Wei-Bin Hsu, Shih-Feng Tsai, Anne Chao, and Hong-Hsing Liu. Network signatures of igg immune repertoires in hepatitis b associated chronic infection and vaccination responses. *Scientific reports*, 6, 2016.

[153] Victor Greiff, Ulrike Menzel, Enkelejda Miho, Cédric Weber, René Riedel, Skylar Cook, Atijeh Valai, Telma Lopes, Andreas Radbruch, Thomas H Winkler, et al. Systems analysis reveals high genetic and antigen-driven predetermination of antibody repertoires throughout b cell development. *Cell Reports*, 19(7):1467–1478, 2017.

[154] Matei Zaharia, Mosharaf Chowdhury, Michael J Franklin, Scott Shenker, and Ion Stoica. Spark: Cluster computing with working sets. *HotCloud*, 10(10-10):95, 2010.

[155] Sang Hoon Lee, Pan-Jun Kim, and Hawoong Jeong. Statistical properties of sampled networks. *Physical Review E*, 73(1):016102, 2006.

[156] Harish Sethu and Xiaoyu Chu. A new algorithm for extracting a small representative subgraph from a very large graph. *arXiv preprint arXiv:1207.4825*, 2012.

[157] Ido Amit, Manuel Garber, Nicolas Chevrier, Ana Paula Leite, Yoni Donner, Thomas Eisenhaure, Mitchell Guttman, Jennifer K Grenier, Weibo Li, Or Zuk, et al. Unbiased reconstruction of a mammalian transcriptional network mediating pathogen responses. *Science*, 326(5950):257–263, 2009.

[158] Georgios A Pavlopoulos, Maria Secrier, Charalampos N Moschopoulos, Theodoros G Soldatos, Sophia Kossida, Jan Aerts, Reinhard Schneider, and Pantelis G Bagos. Using graph theory to analyze biological networks. *BioData mining*, 4(1):10, 2011.

[159] Mark EJ Newman. Assortative mixing in networks. *Physical Review Letters*, 89(20):208701, 2002.

[160] Linton C Freeman. Centrality in social networks conceptual clarification. *Social networks*, 1(3):215–239, 1978.

[161] Phillip Bonacich. Power and centrality: A family of measures. *American journal of sociology*, 92(5):1170–1182, 1987.

[162] Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual web search engine. *Computer networks and ISDN systems*, 30(1):107–117, 1998.

[163] Stephen B Seidman. Network structure and minimum degree. *Social networks*, 5(3):269–287, 1983.

[164] Jacob D Galson, Johannes Trück, Anna Fowler, Márton Münz, Vincenzo Cerundolo, Andrew J Pollard, Gerton Lunter, and Dominic F Kelly. In-depth assessment of within-individual and inter-individual variation in the b cell receptor repertoire. *Frontiers in immunology*, 6:531, 2015.

[165] Felix Horns, Christopher Vollmers, Derek Croote, Sally F Mackey, Gary E Swan, Cornelia L Dekker, Mark M Davis, and Stephen R Quake. Lineage tracing of human b cells reveals the in vivo landscape of human antibody class switching. *eLife*, 5:e16578, 2016.

[166] Andreas Wagner. Robustness against mutations in genetic networks of yeast. *Nature genetics*, 24(4):355–361, 2000.

[167] Dennis R Burton and Lars Hangartner. Broadly neutralizing antibodies to hiv and their role in vaccine design. *Annual review of immunology*, 34:635–659, 2016.

[168] Bryan Briney, Devin Sok, Joseph G Jardine, Daniel W Kulp, Patrick Skog, Sergey Menis, Ronald Jacak, Oleksandr Kalyuzhniy, Natalia De Val, Fabian Sesterhenn, et al. Tailored immunogens direct affinity maturation toward hiv neutralizing antibodies. *Cell*, 166(6):1459–1470, 2016.

[169] Joshua L Payne and Andreas Wagner. The robustness and evolvability of transcription factor binding sites. *Science*, 343(6173):875–877, 2014.

[170] Manuel Ruiz, Véronique Giudicelli, Chantal Ginestoux, Peter Stoehr, James Robinson, Julia Bodmer, Steven GE Marsh, Ronald Bontrop, Marc Lemaitre, Gérard Lefranc, et al. Imgt, the international immunogenetics database. *Nucleic acids research*, 28(1):219–221, 2000.

[171] Ankur Dave, Alekh Jindal, Li Erran Li, Reynold Xin, Joseph Gonzalez, and Matei Zaharia. Graphframes: an integrated api for mixing graph and relational queries. In *Proceedings of the Fourth International Workshop on Graph Data Management Experiences and Systems*, page 2. ACM, 2016.

[172] Daniel A Schult and P Swart. Exploring network structure, dynamics, and function using networkx. In *Proceedings of the 7th Python in Science Conferences (SciPy 2008)*, volume 2008, pages 11–16, 2008.

[173] Tiago P Peixoto. The graph-tool python library. *figshare*, 2014.

[174] Stéfan van der Walt, S Chris Colbert, and Gael Varoquaux. The numpy array: a structure for efficient numerical computation. *Computing in Science & Engineering*, 13(2):22–30, 2011.

[175] Aaron Clauset, Cosma Rohilla Shalizi, and Mark EJ Newman. Power-law distributions in empirical data. *SIAM review*, 51(4):661–703, 2009.

[176] CS Gillespie. Fitting heavy tailed distributions: The powerlaw package. r package version 0.20. 5, 2015.

[177] Yogesh Virkar, Aaron Clauset, et al. Power-law distributions in binned empirical data. *The Annals of Applied Statistics*, 8(1):89–119, 2014.

[178] Mark EJ Newman. The structure and function of complex networks. *SIAM review*, 45(2):167–256, 2003.

[179] Mark EJ Newman. Random graphs as models of networks. *arXiv preprint cond-mat/0202208*, 2002.

[180] Jon M Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM (JACM)*, 46(5):604–632, 1999.

[181] Victor Greiff, Henning Redestig, Juliane Luck, Nicole Bruni, Atijeh Valai, Susanne Hartmann, Sebastian Rausch, Johannes Schuchhardt, and Michal Or-Guil. A minimal model of peptide binding predicts ensemble properties of serum antibodies. *BMC Genomics*, 13(1):79, February 2012.

[182] Rob J Hyndman, Yeasmin Khandakar, et al. Automatic time series for forecasting: the forecast package for r. Technical report, Monash University, Department of Econometrics and Business Statistics, 2007.

[183] R Core Team. A language and environment for statistical computing. r foundation for statistical computing, 2015; vienna, austria, 2016.

[184] Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, 1999.

[185] Hadley Wickham. *ggplot2: elegant graphics for data analysis.* Springer, 2009.

[186] Renaud Gaujoux and Cathal Seoighe. A flexible r package for nonnegative matrix factorization. *BMC bioinformatics*, 11(1):367, 2010.

[187] Donghyeon Yu, MinSoo Kim, Guanghua Xiao, and Tae Hyun Hwang. Review of biological network data and its applications. *Genomics & informatics*, 11(4):200–210, 2013.

[188] CB Anfinsen. Principles that govern the protein folding chains. *Science*, 181:233–230, 1973.

[189] Min-Sik Kim, Sneha M Pinto, Derese Getnet, Raja Sekhar Nirujogi, Srikanth S Manda, Raghothama Chaerkady, Anil K Madugundu, Dhanashree S Kelkar, Ruth Isserlin, Shobhit Jain, et al. A draft map of the human proteome. *Nature*, 509(7502):575–581, 2014.

[190] William S DeWitt, Paul Lindau, Thomas M Snyder, Anna M Sherwood, Marissa Vignali, Christopher S Carlson, Philip D Greenberg, Natalie Duerkopp, Ryan O Emerson, and Harlan S Robins. A public database of memory and naive b-cell receptor sequences. *PloS one*, 11(8):e0160853, 2016.

[191] Randi Vita, James A Overton, Jason A Greenbaum, Julia Ponomarenko, Jason D Clark, Jason R Cantrell, Daniel K Wheeler, Joseph L Gabbard, Deborah Hix, Alessandro Sette, et al. The immune epitope database (iedb) 3.0. *Nucleic acids research*, 43(D1):D405–D412, 2015.

[192] UniProt Consortium et al. Uniprot: a hub for protein information. *Nucleic acids research*, page gku989, 2014.

[193] Melissa D Simek, Wasima Rida, Frances H Priddy, Pham Pung, Emily Carrow, Dagna S Laufer, Jennifer K Lehrman, Mark Boaz, Tony Tarragona-Fiol, George Miiro, et al. Human immunodeficiency virus type 1 elite neutralizers: individuals with broad and potent neutralizing activity identified by using a high-throughput neutralization assay together with an analytical selection algorithm. *Journal of virology*, 83(14):7337–7348, 2009.

[194] Nicole A Doria-Rose, Rachel M Klein, Marcus G Daniels, Sijy O'Dell, Martha Nason, Alan Lapedes, Tanmoy Bhattacharya, Stephen A Migueles, Richard T Wyatt, Bette T Korber, et al. Breadth of human immunodeficiency virus-specific neutralizing activity in sera: clustering analysis and association with clinical variables. *Journal of virology*, 84(3):1631–1636, 2010.

[195] G Pantaleo and AS Fauci. Immunopathogenesis of hiv infection 1. *Annual Reviews in Microbiology*, 50(1):825–854, 1996.

[196] Kevin O Saunders, Lingshu Wang, M Gordon Joyce, Zhi-Yong Yang, Alejandro B Balazs, Cheng Cheng, Sung-Youl Ko, Wing-Pui Kong, Rebecca S Rudicell, Ivelin S Georgiev, et al. Broadly neutralizing human immunodeficiency virus type 1 antibody gene transfer protects nonhuman primates from mucosal simian-human immunodeficiency virus infection. *Journal of virology*, 89(16):8334–8345, 2015.

[197] Timothée Bruel, Florence Guivel-Benhassine, Sonia Amraoui, Marine Malbec, Léa Richard, Katia Bourdic, Daniel Aaron Donahue, Valérie Lorin, Nicoletta Casartelli, Nicolas Noël, et al. Elimination of hiv-1-infected cells by broadly neutralizing antibodies. *Nature communications*, 7, 2016.

[198] Joseph G Jardine, Daniel W Kulp, Colin Havenar-Daughton, Anita Sarkar, Bryan Briney, Devin Sok, Fabian Sesterhenn, June Ereño-Orbea, Oleksandr Kalyuzhniy, Isaiah Deresa, et al. Hiv-1 broadly neutralizing antibody precursor b cells revealed by germline-targeting immunogen. *Science*, 351(6280):1458–1463, 2016.

[199] Amelia Escolano, Pia Dosenovic, and Michel C Nussenzweig. Progress toward active or passive hiv-1 vaccination. *Journal of Experimental Medicine*, pages jem–20161765, 2016.

[200] L. M. Walker, S. K. Phogat, P.-Y. Chan-Hui, D. Wagner, P. Phung, J. L. Goss, T. Wrin, M. D. Simek, S. Fling, J. L. Mitcham, J. K. Lehrman, F. H. Priddy, O. A. Olsen, S. M. Frey, P. W. Hammond, Protocol G Principal Investigators, S. Kaminsky, T. Zamb, M. Moyle, W. C. Koff, P. Poignard, and D. R. Burton. Broad and potent neutralizing antibodies from an african donor reveal a new HIV-1 vaccine target. *Science*, 326(5950):285–289, September 2009.

[201] Johannes F. Scheid, Hugo Mouquet, Juliane Kofer, Sergey Yurasov, Michel C. Nussenzweig, and Hedda Wardemann. Differential regulation of self-reactivity discriminates between IgG+ human circulating memory b cells and bone marrow plasma cells. *Proceedings of the National Academy of Sciences*, 2011.

[202] Mattia Bonsignori, David C Montefiori, Xueling Wu, Xi Chen, Kwan-Ki Hwang, Chun-Yen Tsao, Daniel M Kozink, Robert J Parks, Georgia D Tomaras, John A Crump, et al. Two distinct broadly neutralizing antibody specificities of different clonal lineages in a single hiv-1-infected donor: implications for vaccine design. *Journal of virology*, 86(8):4688–4692, 2012.

[203] John R Mascola and Barton F Haynes. Hiv-1 neutralizing antibodies: understanding nature's pathways. *Immunological reviews*, 254(1):225–244, 2013.

[204] Peter D. Kwong, John R. Mascola, and Gary J. Nabel. Broadly neutralizing antibodies and the search for an HIV-1 vaccine: the end of the beginning. *Nature Reviews Immunology*, 13(9):693–701, September 2013.

[205] Xueling Wu and Xiang-Peng Kong. Antigenic landscape of the hiv-1 envelope and new immunological concepts defined by hiv-1 broadly neutralizing antibodies. *Current opinion in immunology*, 42:56–64, 2016.

[206] Tongqing Zhou, Rebecca M Lynch, Lei Chen, Priyamvada Acharya, Xueling Wu, Nicole A Doria-Rose, M Gordon Joyce, Daniel Lingwood, Cinque Soto, Robert T Bailer, et al. Structural repertoire of hiv-1-neutralizing antibodies targeting the cd4 supersite in 14 donors. *Cell*, 161(6):1280–1292, 2015.

[207] Laura M. Walker, Michael Huber, Katie J. Doores, Emilia Falkowska, Robert Pejchal, Jean-Philippe Julien, Sheng-Kai Wang, Alejandra Ramos, Po-Ying Chan-Hui, Matthew Moyle, Jennifer L. Mitcham, Phillip W. Hammond, Ole A. Olsen, Pham Phung, Steven Fling, Chi-Huey Wong, Sanjay Phogat, Terri Wrin, Melissa D. Simek, Protocol G. Principal Investigators, Wayne C. Koff, Ian A. Wilson, Dennis R. Burton, and Pascal Poignard. Broad neutralization coverage of HIV by multiple highly potent antibodies. *Nature*, 477(7365):466–470, August 2011.

[208] Linling He, Devin Sok, Parisa Azadnia, Jessica Hsueh, Elise Landais, Melissa Simek, Wayne C Koff, Pascal Poignard, Dennis R Burton, and Jiang Zhu. Toward a more accurate view of human b-cell repertoire by next-generation sequencing,

unbiased repertoire capture and single-molecule barcoding. *Scientific reports*, 4:6778, 2014.

[209] Georgia D Tomaras, James M Binley, Elin S Gray, Emma T Crooks, Keiko Osawa, Penny L Moore, Nancy Tumba, Tommy Tong, Xiaoying Shen, Nicole L Yates, et al. Polyclonal b cell responses to conserved neutralization epitopes in a subset of hiv-1-infected individuals. *Journal of virology*, 85(21):11502–11519, 2011.

[210] Xiaoying Shen, Robert J Parks, David C Montefiori, Jennifer L Kirchherr, Brandon F Keele, Julie M Decker, William A Blattner, Feng Gao, Kent J Weinhold, Charles B Hicks, et al. In vivo gp41 antibodies targeting the 2f5 monoclonal antibody epitope mediate human immunodeficiency virus type 1 neutralization breadth. *Journal of virology*, 83(8):3617–3625, 2009.

[211] Lynn Morris, Xi Chen, Munir Alam, Georgia Tomaras, Ruijun Zhang, Dawn J Marshall, Bing Chen, Robert Parks, Andrew Foulger, Frederick Jaeger, et al. Isolation of a human anti-hiv gp41 membrane proximal region neutralizing antibody by antigen-specific single b cell sorting. *PloS one*, 6(9):e23532, 2011.

[212] Thomas B Kepler, Hua-Xin Liao, S Munir Alam, Rekha Bhaskarabhatla, Ruijun Zhang, Chandri Yandava, Shelley Stewart, Kara Anasti, Garnett Kelsoe, Robert Parks, et al. Immunoglobulin gene insertions and deletions in the affinity maturation of hiv-1 broadly reactive neutralizing antibodies. *Cell host & microbe*, 16(3):304–313, 2014.

[213] A. M. Eroshkin, A. LeBlanc, D. Weekes, K. Post, Z. Li, A. Rajput, S. T. Butera, D. R. Burton, and A. Godzik. bNAber: database of broadly neutralizing HIV antibodies. *Nucleic Acids Research*, 42(D1):D1133–D1139, January 2013.

[214] Xavier Brochet, Marie-Paule Lefranc, and Véronique Giudicelli. Imgt/v-quest: the highly customized and integrated system for ig and tr standardized vj and vdj sequence analysis. *Nucleic acids research*, 36(suppl 2):W503–W508, 2008.

[215] Veronique Giudicelli, Denys Chaume, and Marie-Paule Lefranc. Imgt/gene-db: a comprehensive database for human and mouse immunoglobulin and t cell receptor genes. *Nucleic acids research*, 33(suppl 1):D256–D261, 2005.

[216] Alfréd Rényi et al. On measures of entropy and information. In *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 547–561, 1961.

[217] Lou Jost. Entropy and diversity. *Oikos*, 113(2):363–375, 2006.

[218] Johannes Palme, Sepp Hochreiter, and Ulrich Bodenhofer. Kebabs: an r package for kernel-based analysis of biological sequences. *Bioinformatics*, page btv176, 2015.

[219] Lei Yu and Yongjun Guan. Immunologic basis for long hcdr3s in broadly neutralizing antibodies against hiv-1. *Frontiers in immunology*, 5:250, 2014.

[220] Cathrine Scheepers, Ram K Shrestha, Bronwen E Lambson, Katherine JL Jackson, Imogen A Wright, Dshanta Naicker, Mark Goosen, Leigh Berrie, Arshad Ismail, Nigel Garrett, et al. Ability to develop broadly neutralizing hiv-1 antibodies is not restricted by the germline ig gene repertoire. *The Journal of Immunology*, 194(9):4371–4378, 2015.

[221] Maxwell W Libbrecht and William Stafford Noble. Machine learning applications in genetics and genomics. *Nature Reviews Genetics*, 16(6):321–332, 2015.

[222] Marina Caskey, Florian Klein, Julio CC Lorenzi, Michael S Seaman, Anthony P West Jr, Noreen Buckley, Gisela Kremer, Lilian Nogueira, Malte Braunschweig, Johannes F Scheid, et al. Viraemia suppressed in hiv-1-infected humans by broadly neutralizing antibody 3bnc117. *Nature*, 522(7557):487–491, 2015.

[223] Marina Caskey, Florian Klein, and Michel C Nussenzweig. Broadly neutralizing antibodies for hiv-1 prevention or immunotherapy. *New England Journal of Medicine*, 375(21):2019–2021, 2016.

[224] Niclas Thomas, Katharine Best, Mattia Cinelli, Shlomit Reich-Zeliger, Hilah Gal, Eric Shifrut, Asaf Madi, Nir Friedman, John Shawe-Taylor, and Benny 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*, page btu523, 2014.

[225] Gur Yaari and Steven H Kleinstein. Practical guidelines for b-cell receptor repertoire sequencing analysis. *Genome medicine*, 7(1):121, 2015.

[226] Yingxin Han, Hongmei Li, Yanfang Guan, and Jian Huang. Immune repertoire: A potential biomarker and therapeutic for hepatocellular carcinoma. *Cancer letters*, 379(2):206–212, 2016.

[227] Douglas G McNeel. Tcr diversity–a universal cancer immunotherapy biomarker? *Journal for ImmunoTherapy of Cancer*, 4(1):69, 2016.

[228] Kenji Tamura, Shoichi Hazama, Rui Yamaguchi, Seiya Imoto, Hiroko Takenouchi, Yuka Inoue, Shinsuke Kanekiyo, Yoshitaro Shindo, Satoru Miyano, Yusuke Nakamura, et al. Characterization of the t cell repertoire by deep t cell receptor sequencing in tissues and blood from patients with advanced colorectal cancer. *Oncology letters*, 11(6):3643–3649, 2016.

[229] Hennie R Hoogenboom. Selecting and screening recombinant antibody libraries. *Nature biotechnology*, 23(9):1105–1116, 2005.

[230] Richard A Lerner. Combinatorial antibody libraries: new advances, new immunological insights. *Nature Reviews Immunology*, 2016.

[231] Andrew RM Bradbury, Sachdev Sidhu, Stefan Dübel, and John McCafferty. Beyond natural antibodies: the power of in vitro display technologies. *Nature biotechnology*, 29(3):245–254, 2011.

[232] J Glanville, S Dngelo, TA Khan, ST Reddy, L Naranjo, F Ferrara, and ARM Bradbury. Deep sequencing in library selection projects: what insight does it bring? *Current opinion in structural biology*, 33:146–160, 2015.

[233] Saikat Banerjee, Heliang Shi, Marisa Banasik, Hojin Moon, William Lees, Yali Qin, Andrew Harley, Adrian Shepherd, and Michael W Cho. Evaluation of a novel multi-immunogen vaccine strategy for targeting 4e10/10e8 neutralizing epitopes on hiv-1 gp41 membrane proximal external region. *Virology*, 505:113–126, 2017.

[234] Jinal N Bhiman, Colin Anthony, Nicole A Doria-Rose, Owen Karimanzira, Chaim A Schramm, Thandeka Khoza, Dale Kitchin, Gordon Botha, Jason Gorman, Nigel J Garrett, et al. Viral variants that initiate and drive maturation of v1v2-directed hiv-1 broadly neutralizing antibodies. *Nature medicine*, 21(11):1332, 2015.

[235] Daniel J Bolland, Hashem Koohy, Andrew L Wood, Louise S Matheson, Felix Krueger, Michael JT Stubbington, Amanda Baizan-Edge, Peter Chovanec, Bryony A Stubbs, Kristina Tabbada, et al. Two mutually exclusive local chromatin states drive efficient v (d) j recombination. *Cell reports*, 15(11):2475–2487, 2016.

[236] Mattia Bonsignori, Tongqing Zhou, Zizhang Sheng, Lei Chen, Feng Gao, M Gordon Joyce, Gabriel Ozorowski, Gwo-Yu Chuang, Chaim A Schramm, Kevin Wiehe, et al. Maturation pathway from germline to broad hiv-1 neutralizer of a cd4-mimic antibody. *Cell*, 165(2):449–463, 2016.

[237] Olga V Britanova, Mikhail Shugay, Ekaterina M Merzlyak, Dmitriy B Staroverov, Ekaterina V Putintseva, Maria A Turchaninova, Ilgar Z Mamedov, Mikhail V Pogorelyy, Dmitriy A Bolotin, Mark Izraelson, et al. Dynamics of individual t cell repertoires: From cord blood to centenarians. *The Journal of Immunology*, 196(12):5005–5013, 2016.

[238] Zengchao Chen, Chaoting Zhang, Yaqi Pan, Ruiping Xu, Changqing Xu, Ziping Chen, Zheming Lu, and Yang Ke. T cell receptor β-chain repertoire analysis reveals intratumour heterogeneity of tumour-infiltrating lymphocytes in oesophageal squamous cell carcinoma. *The Journal of pathology*, 239(4):450–458, 2016.

[239] Martin M Corcoran, Ganesh E Phad, Néstor Vázquez Bernat, Christiane Stahl-Hennig, Noriyuki Sumida, Mats AA Persson, Marcel Martin, and Gunilla B Karlsson Hedestam. Production of individualized v gene databases reveals high levels of immunoglobulin genetic diversity. *Nature communications*, 7:13642, 2016.

[240] Ang Cui, Roberto Di Niro, Jason A Vander Heiden, Adrian W Briggs, Kris Adams, Tamara Gilbert, Kevin C Oonnor, Francois Vigneault, Mark J Shlomchik, and Steven H Kleinstein. A model of somatic hypermutation targeting in mice based on high-throughput ig sequencing data. *The Journal of Immunology*, 197(9):3566–3574, 2016.

[241] Martin S Davey, Carrie R Willcox, Stephen P Joyce, Kristin Ladell, Sofya A Kasatskaya, James E McLaren, Stuart Hunter, Mahboob Salim, Fiyaz Mohammed, David A Price, et al. Clonal selection in the human vδ1 t cell repertoire

indicates $\gamma\delta$ tcr-dependent adaptive immune surveillance. *Nature Communications*, 8:14760, 2017.

[242] Jared Dean, Ryan O Emerson, Marissa Vignali, Anna M Sherwood, Mark J Rieder, Christopher S Carlson, and Harlan S Robins. Annotation of pseudogenic gene segments by massively parallel sequencing of rearranged lymphocyte receptor loci. *Genome medicine*, 7(1):123, 2015.

[243] Brandon J. DeKosky, Gregory C. Ippolito, Ryan P. Deschner, Jason J. Lavinder, Yariv Wine, Brandon M. Rawlings, Navin Varadarajan, Claudia Giesecke, Thomas Dörner, Sarah F. Andrews, Patrick C. Wilson, Scott P. Hunicke-Smith, C. Grant Willson, Andrew D. Ellington, and George Georgiou. High-throughput sequencing of the paired human immunoglobulin heavy and light chain repertoire. *Nature Biotechnology*, 2013.

[244] Brandon J DeKosky, Takaaki Kojima, Alexa Rodin, Wissam Charab, Gregory C Ippolito, Andrew D Ellington, and George Georgiou. In-depth determination and analysis of the human paired heavy-and light-chain antibody repertoire. *Nature medicine*, 21(1):86–91, 2015.

[245] Nicole A. Doria-Rose, Chaim A. Schramm, Jason Gorman, Penny L. Moore, Jinal N. Bhiman, Brandon J. DeKosky, Michael J. Ernandes, Ivelin S. Georgiev, Helen J. Kim, Marie Pancera, Ryan P. Staupe, Han R. Altae-Tran, Robert T. Bailer, Ema T. Crooks, Albert Cupo, Aliaksandr Druz, Nigel J. Garrett, Kam H. Hoi, Rui Kong, Mark K. Louder, Nancy S. Longo, Krisha McKee, Molati Nonyane, Sijy O'Dell, Ryan S. Roark, Rebecca S. Rudicell, Stephen D. Schmidt, Daniel J. Sheward, Cinque Soto, Constantinos Kurt Wibmer, Yongping Yang, Zhenhai Zhang, NISC Comparative Sequencing, James C. Mullikin, James M. Binley, Rogier W. Sanders, Ian A. Wilson, John P. Moore, Andrew B. Ward, George Georgiou, Carolyn Williamson, Salim S. Abdool Karim, Lynn Morris, Peter D. Kwong, Lawrence Shapiro, and John R. Mascola. Developmental pathway for potent V1V2-directed HIV-neutralizing antibodies. *Nature*, advance online publication, March 2014.

[246] Nicole A Doria-Rose, Jinal N Bhiman, Ryan S Roark, Chaim A Schramm, Jason Gorman, Gwo-Yu Chuang, Marie Pancera, Evan M Cale, Michael J Ernandes, Mark K Louder, et al. New member of the v1v2-directed cap256-vrc26 lineage that shows increased breadth and exceptional potency. *Journal of virology*, 90(1):76–91, 2016.

[247] Yongqiang Feng, Joris Van Der Veeken, Mikhail Shugay, Ekaterina V Putintseva, Hatice U Osmanbeyoglu, Stanislav Dikiy, Beatrice E Hoyos, Bruno Moltedo, Saskia Hemmers, Piper Treuting, et al. A mechanism for expansion of regulatory t cell repertoire and its role in self tolerance. *Nature*, 528(7580):132, 2015.

[248] J Douglas Freeman, René L Warren, John R Webb, Brad H Nelson, and Robert A Holt. Profiling the t-cell receptor beta-chain repertoire by massively parallel sequencing. *Genome research*, 19(10):1817–1824, 2009.

[249] James M Heather, Katharine Best, Theres Oakes, Eleanor R Gray, Jennifer K Roe, Niclas Thomas, Nir Friedman, Mahdad Noursadeghi, and Benjamin Chain.

Dynamic perturbations of the t-cell receptor repertoire in chronic hiv infection and following antiretroviral therapy. *Frontiers in immunology*, 6, 2016.

[250] Melody S Hsu, Shaina Sedighim, Tina Wang, Joseph P Antonios, Richard G Everson, Alexander M Tucker, Lin Du, Ryan Emerson, Erik Yusko, Catherine Sanders, et al. Tcr sequencing can identify and track glioma-infiltrating t cells after dc vaccination. *Cancer immunology research*, 4(5):412–418, 2016.

[251] Jinghe Huang, Byong H Kang, Elise Ishida, Tongqing Zhou, Trevor Griesman, Zizhang Sheng, Fan Wu, Nicole A Doria-Rose, Baoshan Zhang, Krisha McKee, et al. Identification of a cd4-binding-site antibody to hiv that evolved near-pan neutralization breadth. *Immunity*, 45(5):1108–1121, 2016.

[252] Ning Jiang, Joshua A. Weinstein, Lolita Penland, Richard A. White, Daniel S. Fisher, and Stephen R. Quake. Determinism and stochasticity during maturation of the zebrafish antibody repertoire. *Proceedings of the National Academy of Sciences*, 108(13):5348 –5353, 2011.

[253] M Gordon Joyce, Adam K Wheatley, Paul V Thomas, Gwo-Yu Chuang, Cinque Soto, Robert T Bailer, Aliaksandr Druz, Ivelin S Georgiev, Rebecca A Gillespie, Masaru Kanekiyo, et al. Vaccine-induced antibodies that neutralize group 1 and group 2 influenza a viruses. *Cell*, 166(3):609–623, 2016.

[254] Julia Kargl, Stephanie E Busch, Grace HY Yang, Kyoung-Hee Kim, Mark L Hanke, Heather E Metz, Jesse J Hubbard, Sylvia M Lee, David K Madtes, Martin W McIntosh, et al. Neutrophils dominate the immune cell composition in non-small cell lung cancer. *Nature Communications*, 8, 2017.

[255] Tarik A Khan, Simon Friedensohn, Arthur R Gorter de Vries, Jakub Straszewski, Hans-Joachim Ruscheweyh, and Sai T Reddy. Accurate and predictive antibody repertoire profiling by molecular amplification fingerprinting. *Science advances*, 2(3):e1501371, 2016.

[256] Mattias Levin, Jasmine J King, Jacob Glanville, Katherine JL Jackson, Timothy J Looney, Ramona A Hoh, Adriano Mari, Morgan Andersson, Lennart Greiff, Andrew Z Fire, et al. Persistence and evolution of allergen-specific ige repertoires during subcutaneous specific immunotherapy. *Journal of Allergy and Clinical Immunology*, 137(5):1535–1544, 2016.

[257] Mattias Levin, Fredrik Levander, Robert Palmason, Lennart Greiff, and Mats Ohlin. Antibody-encoding repertoires of bone marrow and peripheral blood focus on ige. *Journal of Allergy and Clinical Immunology*, 139(3):1026–1030, 2017.

[258] Ana Raquel Maceiras, Silvia Cristina Paiva Almeida, Encarnita Mariotti-Ferrandiz, Wahiba Chaara, Fadi Jebbawi, Adrien Six, Shohei Hori, David Klatzmann, Jose Faro, and Luis Graca. T follicular helper and t follicular regulatory cells have different tcr specificity. *Nature Communications*, 8, 2017.

[259] Ilgar Z Mamedov, Olga V Britanova, Dmitriy A Bolotin, Anna V Chkalina, Dmitriy B Staroverov, Ivan V Zvyagin, Alexey A Kotlobay, Maria A Turchaninova, Denis A Fedorenko, Andrew A Novik, et al. Quantitative tracking of t cell clones

after haematopoietic stem cell transplantation. *EMBO molecular medicine*, 3(4):201–207, 2011.

[260] Spencer D Martin, Scott D Brown, Darin A Wick, Julie S Nielsen, David R Kroeger, Kwame Twumasi-Boateng, Robert A Holt, and Brad H Nelson. Low mutation burden in ovarian cancer may limit the utility of neoantigen-targeted vaccines. *PloS one*, 11(5):e0155189, 2016.

[261] Miri Michaeli, Hilla Tabibian-Keissar, Ginette Schiby, Gitit Shahaf, Yishai Pickman, Lena Hazanov, Kinneret Rosenblatt, Deborah K Dunn-Walters, Iris Barshack, and Ramit Mehr. Immunoglobulin gene repertoire diversification and selection in the stomach–from gastritis to gastric lymphomas. *Frontiers in immunology*, 5, 2014.

[262] Eva Szymanska Mroczek, Gregory C. Ippolito, Tobias Rogosch, Kam Hon Hoi, Tracy A. Hwangpo, Marsha G. Brand, Yingxin Zhuang, Cun Ren Liu, David A. Schneider, Michael Zemlin, Elizabeth E. Brown, George Georgiou, and Harry W. Jr Schroeder. Differences in the composition of the human antibody repertoire by b cell subsets in the blood. *B Cell Biology*, 5:96, 2014.

[263] Miyo Ota, Bao H Duong, Ali Torkamani, Colleen M Doyle, Amanda L Gavin, Takayuki Ota, and David Nemazee. Regulation of the b cell receptor repertoire and self-reactivity by baff. *The Journal of Immunology*, 185(7):4128–4136, 2010.

[264] Arumugam Palanichamy, Leonard Apeltsin, Tracy C Kuo, Marina Sirota, Shengzhi Wang, Steven J Pitts, Purnima D Sundar, Dilduz Telman, Lora Z Zhao, Mia Derstine, et al. Immunoglobulin class-switched b cells form an active immune axis between cns and periphery in multiple sclerosis. *Science translational medicine*, 6(248):248ra106–248ra106, 2014.

[265] Kathrin Pieper, Joshua Tan, Luca Piccoli, Mathilde Foglierini, Sonia Barbieri, Yiwei Chen, Chiara Silacci-Fregni, Tobias Wolf, David Jarrossay, Marica Anderle, Abdirahman Abdi, Francis M. Ndungu, Ogobara K. Doumbo, Boubacar Traore, Tuan M. Tran, Said Jongo, Isabelle Zenklusen, Peter D. Crompton, Claudia Daubenberger, Peter C. Bull, Federica Sallusto, and Antonio Lanzavecchia. Public antibodies to malaria antigens generated by two lair1 insertion modalities. *Nature*, 2017.

[266] Ekaterina V Putintseva, Olga V Britanova, Dmitriy B Staroverov, Ekaterina M Merzlyak, Maria A Turchaninova, Mikhail Shugay, Dmitriy A Bolotin, Mikhail V Pogorelyy, Ilgar Z Mamedov, Vlasta Bobrynina, et al. Mother and child t cell receptor repertoires: deep profiling study. *Frontiers in immunology*, 4, 2013.

[267] Qian Qi, Yi Liu, Yong Cheng, Jacob Glanville, David Zhang, Ji-Yeun Lee, Richard A Olshen, Cornelia M Weyand, Scott D Boyd, and Jörg J Goronzy. Diversity and clonal selection in the human t-cell repertoire. *Proceedings of the National Academy of Sciences*, 111(36):13139–13144, 2014.

[268] Florian Rubelt, Christopher R Bolen, Helen M McGuire, Jason A Vander Heiden, Daniel Gadala-Maria, Mikhail Levin, Ghia M Euskirchen, Murad R Mamedov, Gary E Swan, Cornelia L Dekker, et al. Individual heritable differences result

in unique cell lymphocyte receptor repertoires of naive and antigen-experienced cells. *Nature communications*, 7, 2016.

[269] Mikhail Shugay, Dmitriy V Bagaev, Maria A Turchaninova, Dmitriy A Bolotin, Olga V Britanova, Ekaterina V Putintseva, Mikhail V Pogorelyy, Vadim I Nazarov, Ivan V Zvyagin, Vitalina I Kirgizova, et al. Vdjtools: unifying post-analysis of t cell receptor repertoires. *PLoS computational biology*, 11(11):e1004503, 2015.

[270] Cinque Soto, Gilad Ofek, M Gordon Joyce, Baoshan Zhang, Krisha McKee, Nancy S Longo, Yongping Yang, Jinghe Huang, Robert Parks, Joshua Eudailey, et al. Developmental pathway of the mper-directed hiv-1-neutralizing antibody 10e8. *PloS one*, 11(6):e0157409, 2016.

[271] Joel NH Stern, Gur Yaari, Jason A Vander Heiden, George Church, William F Donahue, Rogier Q Hintzen, Anita J Huttner, Jon D Laman, Rashed M Nagra, Alyssa Nylander, et al. B cells populating the multiple sclerosis brain mature in the draining cervical lymph nodes. *Science translational medicine*, 6(248):248ra107–248ra107, 2014.

[272] Ming Tian, Cheng Cheng, Xuejun Chen, Hongying Duan, Hwei-Ling Cheng, Mai Dao, Zizhang Sheng, Michael Kimble, Lingshu Wang, Sherry Lin, et al. Induction of hiv neutralizing antibody lineages in mice with diverse precursor repertoires. *Cell*, 166(6):1471–1484, 2016.

[273] Eric Tran, Simon Turcotte, Alena Gros, Paul F Robbins, Yong-Chen Lu, Mark E Dudley, John R Wunderlich, Robert P Somerville, Katherine Hogan, Christian S Hinrichs, et al. Cancer immunotherapy based on mutation-specific cd4+ t cells in a patient with epithelial cancer. *Science*, 344(6184):641–645, 2014.

[274] Konstantinos Tsioris, Namita T Gupta, Adebola O Ogunniyi, Ross M Zimnisky, Feng Qian, Yi Yao, Xiaomei Wang, Joel NH Stern, Raj Chari, Adrian W Briggs, et al. Neutralizing antibodies against west nile virus identified directly from human b cells by single-cell analysis and next generation sequencing. *Integrative Biology*, 7(12):1587–1597, 2015.

[275] MA Turchaninova, A Davydov, OV Britanova, M Shugay, V Bikos, ES Egorov, VI Kirgizova, EM Merzlyak, DB Staroverov, DA Bolotin, et al. High-quality full-length immunoglobulin profiling with unique molecular barcoding. *Nature protocols*, 11(9):1599–1616, 2016.

[276] Jason A Vander Heiden, Panos Stathopoulos, Julian Q Zhou, Luan Chen, Tamara J Gilbert, Christopher R Bolen, Richard J Barohn, Mazen M Dimachkie, Emma Ciafaloni, Teresa J Broering, et al. Dysregulation of b cell repertoire formation in myasthenia gravis patients revealed through deep sequencing. *The Journal of Immunology*, 198(4):1460–1473, 2017.

[277] Christopher Vollmers, Lolita Penland, Jad N Kanbar, and Stephen R Quake. Novel exons and splice variants in the human antibody heavy chain identified by single cell and single molecule sequencing. *PloS one*, 10(1):e0117050, 2015.

[278] Jeffrey J Wallin, Johanna C Bendell, Roel Funke, Mario Sznol, Konstanty Korski, Suzanne Jones, Genevive Hernandez, James Mier, Xian He, F Stephen Hodi, et al. Atezolizumab in combination with bevacizumab enhances antigen-specific t-cell migration in metastatic renal cell carcinoma. *Nature communications*, 7:12624, 2016.

[279] Chunlin Wang, Catherine M. Sanders, Qunying Yang, Harry W. Schroeder, Elijah Wang, Farbod Babrzadeh, Baback Gharizadeh, Richard M. Myers, James R. Hudson, Ronald W. Davis, and Jian Han. High throughput sequencing reveals a complex pattern of dynamic interrelationships among human t cell subsets. *Proceedings of the National Academy of Sciences*, 107(4):1518–1523, January 2010.

[280] Ting Wang, Changxi Wang, Jinghua Wu, Chenyang He, Wei Zhang, Jiayun Liu, Ruifang Zhang, Yonggang Lv, Yongping Li, Xiaojing Zeng, et al. The different t-cell receptor repertoires in breast cancer tumors, draining lymph nodes, and adjacent tissues. *Cancer immunology research*, pages canimm–0107, 2016.

[281] Duane R. Wesemann, Andrew J. Portuguese, Robin M. Meyers, Michael P. Gallagher, Kendra Cluff-Jones, Jennifer M. Magee, Rohit A. Panchakshari, Scott J. Rodig, Thomas B. Kepler, and Frederick W. Alt. Microbial colonization influences early b-lineage development in the gut lamina propria. *Nature*, 501(7465):112–115, September 2013.

[282] Lawren C. Wu and Ali A. Zarrin. The production and regulation of IgE by the immune system. *Nature Reviews Immunology*, advance online publication, March 2014.

[283] Tongqing Zhou, Jiang Zhu, Xueling Wu, Stephanie Moquin, Baoshan Zhang, Priyamvada Acharya, Ivelin S. Georgiev, Han R. Altae-Tran, Gwo-Yu Chuang, M. Gordon Joyce, Young DoKwon, Nancy S. Longo, Mark K. Louder, Timothy Luongo, Krisha McKee, Chaim A. Schramm, Jeff Skinner, Yongping Yang, Zhongjia Yang, Zhenhai Zhang, Anqi Zheng, Mattia Bonsignori, Barton F. Haynes, Johannes F. Scheid, Michel C. Nussenzweig, Melissa Simek, Dennis R. Burton, Wayne C. Koff, NISC Comparative Sequencing Program, James C. Mullikin, Mark Connors, Lawrence Shapiro, Gary J. Nabel, John R. Mascola, and Peter D. Kwong. Multidonor analysis reveals structural elements, genetic determinants, and maturation pathway for HIV-1 neutralization by VRC01-Class antibodies. *Immunity*.

[284] Johannes Köster and Sven Rahmann. Snakemake scalable bioinformatics workflow engine. *Bioinformatics*, 28(19):2520–2522, 2012.

[285] Simon Friedensohn, Tarik A Khan, and Sai T Reddy. Advanced methodologies in high-throughput sequencing of immune repertoires. *Trends in Biotechnology*, 35(3):203–214, 2017.

[286] Asaf Madi, Asaf Poran, Eric Shifrut, Shlomit Reich-Zeliger, Erez Greenstein, Irena Zaretsky, Tomer Arnon, Francois Van Laethem, Alfred Singer, Jinghua Lu, et al. T cell receptor repertoires of mice and humans are clustered in similarity networks around conserved public cdr3 sequences. *eLife*, 6, 2017.

[287] Bo Li, Taiwen Li, Jean-Christophe Pignon, Binbin Wang, Jinzeng Wang, Sachet Shukla, Ruoxu Dou, Qianming Chen, F Stephen Hodi, Toni K Choueiri, et al. Landscape of tumor-infiltrating t cell repertoire of human cancers. *Nature genetics*, 48(7):725, 2016.

[288] George J Xu, Tomasz Kula, Qikai Xu, Mamie Z Li, Suzanne D Vernon, Thumbi Ndung, Kiat Ruxrungtham, Jorge Sanchez, Christian Brander, Raymond T Chung, et al. Comprehensive serological profiling of human populations using a synthetic human virome. *Science*, 348(6239):aaa0698, 2015.

[289] Janice M Reichert. Antibodies to watch in 2017. In *MAbs*, volume 9, pages 167–181. Taylor & Francis, 2017.

Appendix A

List of Abbreviations

Ab	Antibody
BACC	Balanced accuracy
BCR	B-cell receptor
bNAb	Broadly neutralizing antibody
CDR	Complementarity determining region
CV	Coefficient of variation
DB	Database
D	Diversity
FR	Framework
HIV	Human Immunodeficiency Virus
Ig	Immunoglobulin
Ig-seq	Immunoglobulin sequencing
J	Joining
LOOCV	Leave-one-out cross-validation
nBC	Naïve B cell

NGS	Next-generation sequencing
pPC	Pre-B cell
PC	Plasmacell
SHM	Somatic hypermutations
SVM	Support vector machine
TCR	T-cell receptor
V	Variable