Steps in data processing and analysis of adaptive immune receptor repertoires:
best practices, pitfalls, and future directions.

Victor Greiff
Laboratory for Computational and Systems Immunology
Department of Immunology
Associate Professor
University of Oslo

AIRR and TABS Webinar. April 06, 2021.

Victor Greiff

University of Oslo
Associate Professor for Systems Immunology
Focus: Machine-learning driven analysis of immune specificity

AIRR-C function
Chair-Elect

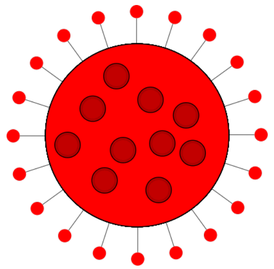
Education
Postdoc – ETH Zürich (Sai Reddy Lab), 2013–2017
Ph.D. – Humboldt University Berlin, 2012

Selected publications
Akbar et al., Cell Reports, 2021
Miho et al., Nat Comms, 2019
Greiff et al., Cell Reports, 2017
Greiff et al., Journal of Immunology, 2017
Greiff et al., Genome Medicine, 2015

Victor Greiff
victor.greiff@medisin.uio.no
@victorgreiff

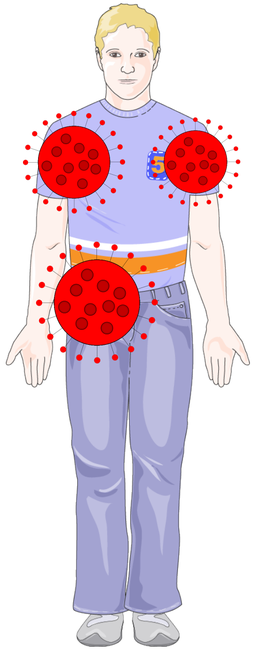
Disclaimer: this webinar is meant as a brief overview of the AIRR field. For increased nuance, please consult the cited references.

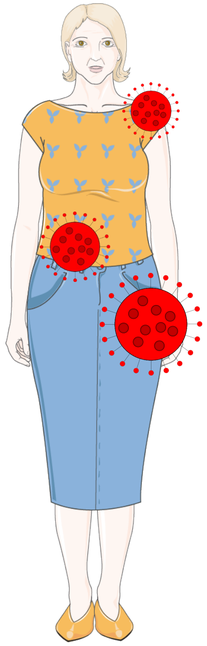
The adaptive immune system records each immune event over a lifetime



- Bacteria
- Virus
- Cancer
- Autoimmunity
- Vaccine

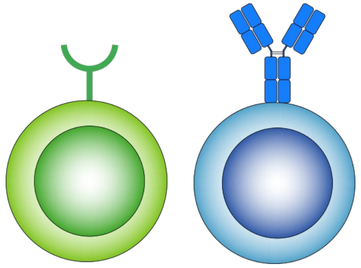
- Infection
- Disease
- Vaccination





Immune history

- T1D
- Celiac disease
- Cancer
- Infection
- …



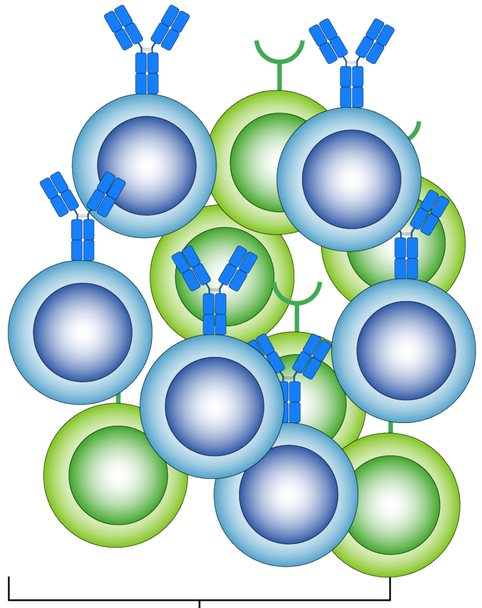
T cell B cell (antibody)

TCRs and BCR (antibodies) are natural diagnostics and therapeutics

Abs and TCRs record past and current immune status

95% of vaccines are antibody-dependent

Monoclonal Abs and CAR-T are blockbuster drugs

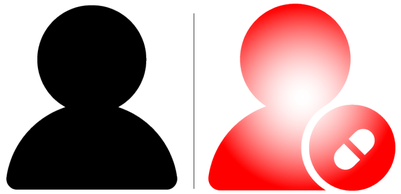


The immune repertoire (Antibody and T-cell receptors)

Diagnostics

Repertoire-level

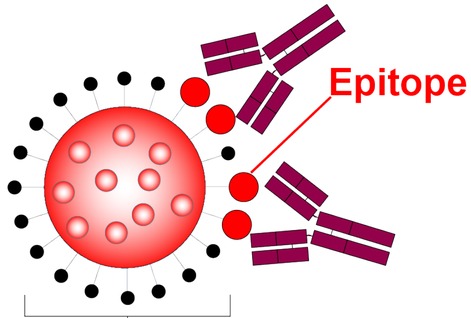
Big immune sequencing data



Healthy **Diseased**

Sequence-level

Immunosequencing
Immunoproteomics



Antigen

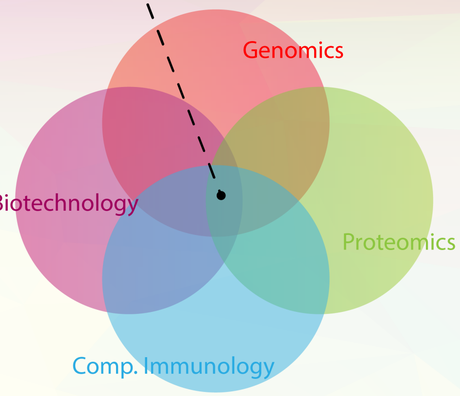
Therapeutics

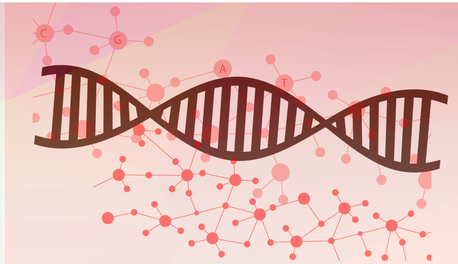
- Potential TCR/Ab diversity: >10^{13}

Key advances and challenges in adaptive immune receptor (BCR, TCR) analysis

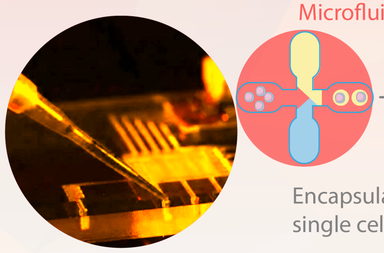
Immunoengineering

Four elements of immunoengineering

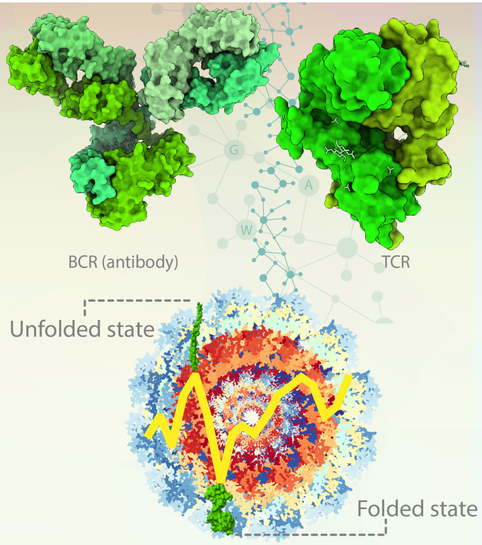




Microfluidics



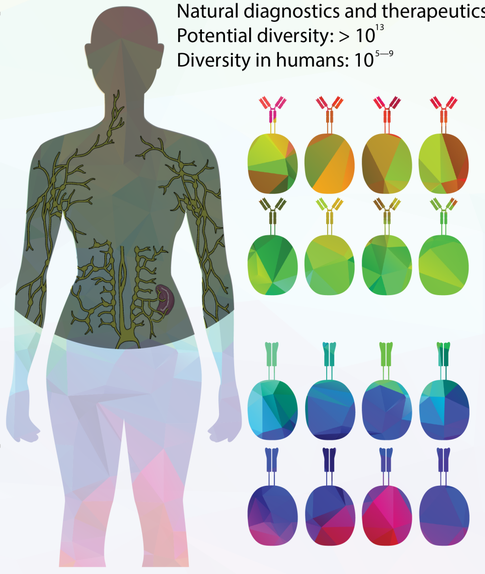
Encapsulated single cell



Immunology

The immune repertoire
Antibodies and T-cell receptors:

Natural diagnostics and therapeutics
Potential diversity: > 10^{13}
Diversity in humans: 10^{5-9}



Genomics

Key advances

- High-throughput single-cell sequencing
- Inference of immune receptor generation probabilities
- Extraction of immune receptor sequences from transcriptomics data
- Quantification of restricted repertoire diversity

Outstanding challenges

- Ultra-throughput antigen-specific receptor sequencing
- Linking receptor specificity to transcriptome
- Non-standardized computational analysis
- Reproducibility and cost

Proteomics

Key advances

- Serum antibody proteomics
- High-throughput epitope mapping
- Inference of 3D immune receptor structure from amino acid sequence
- Antibody maturation driven homotypic antibody interaction

Outstanding challenges

- Linking genomic and proteomic antibody diversity
- De novo antibody protein sequencing
- High-troughput ab initio 3D structure prediction
- High-accuracy receptor antigen docking

Comp. immunology

Key advances

- In-silico prediction of TCR antigen binding
- Multidimensional description of immune receptor landscape
- Probabilistic modeling of immune receptor data
- Large-scale open access standardization and receptor database efforts

Outstanding challenges

- Predicting receptor cross-reactivity and recognition holes
- Machine learning analysis of immune repertoires
- Identification of immune-signal in repertoires
- Scalable analysis of ultra-large datasets

Biotechnology

Key advances

- De-orphanization of TCR-ligand pairs
- Optimization of immune receptor affinity maturation
- Nanobodies and CAR-T-cells
- CRISPR/Cas9 immune cell editing

Outstanding challenges

- High-throughput generation of monoclonal antibodies
- Library-on-library receptor-ligand pair selection
- Efficient immune cell editing in vivo
- Transcriptional control of engineered cellular therapeutics in vivo

Brown et al., MSDE, 2019

Outline

Introduction to Adaptive immune receptor repertoire sequencing (AIRR-seq)
- Generation of immune repertoire diversity
- Workflow and applications of AIRR-seq

Error correction and Standardization of AIRR-seq data
- Experimental design and considerations
- Error and bias correction
- Standardization

Single-cell AIRR-seq
- Pairing by targeted amplification
- Single-cell sequencing

Computational strategies for immune repertoire analysis
- Diversity and convergence analysis
- Network analysis
- Machine learning

Antibody and T cell diversity is generated by VDJ recombination

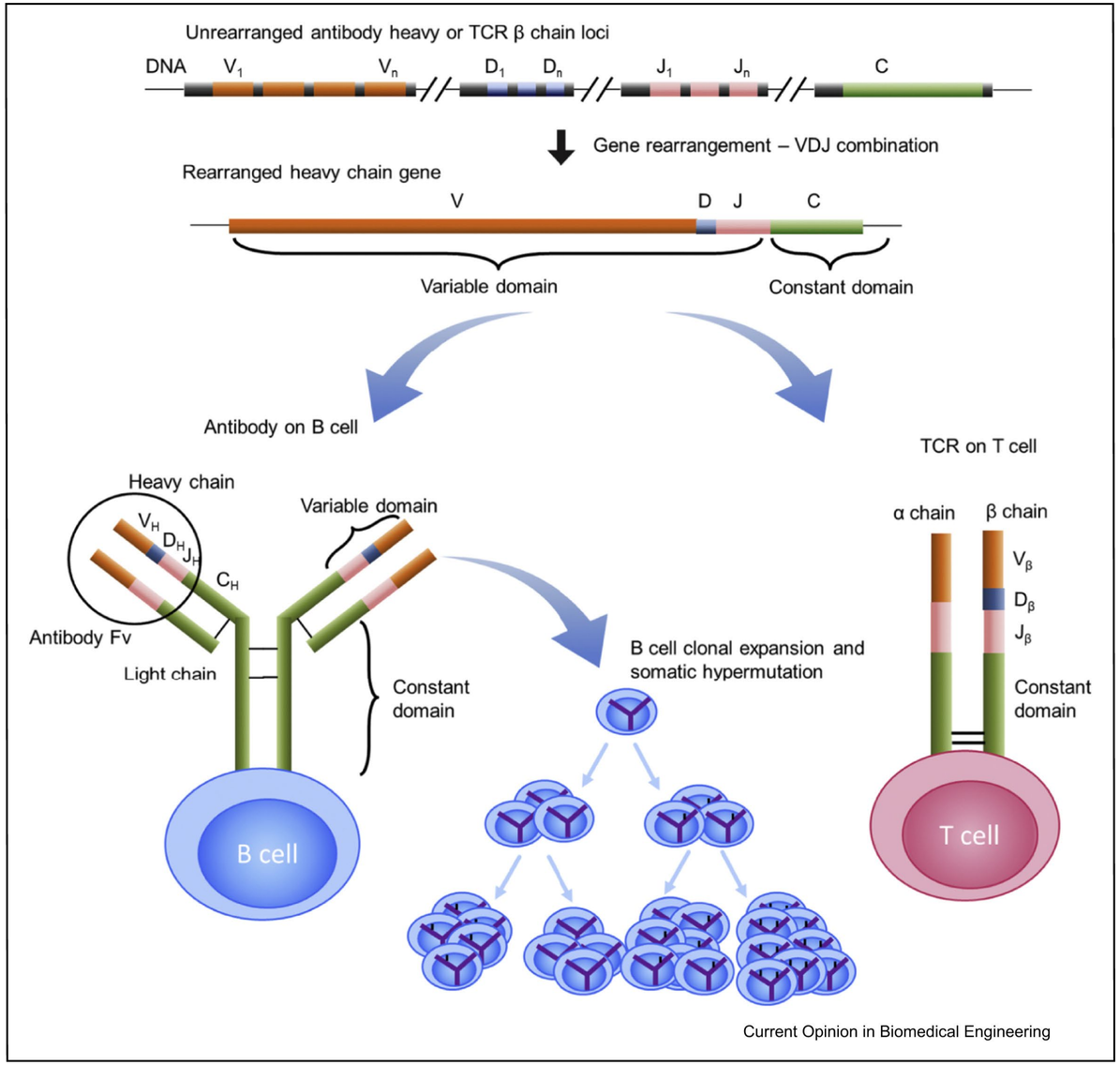
V D, J genes encode the variable domains of the antibody heavy and light chains and the T cell receptor (TCR) α and β chains

One gene segment from each of the three groups of gene segments (V, D, and J) are randomly recombined to form new antibody or TCR sequences (**VDJ recombination**)

There are also random nucleotides introduced at the junction of V, D and J genes. This creates an **enormous potential diversity** of antigen receptor sequences.

Unique to antibody or B cell receptor, its gene sequences can also change itself by introducing random mutations (**somatic hypermutation, 10^{-3}/bp/generation**)

Progeny B cells become a mixture of sub-species (clones, clonal lineage), each expresses a different antibody sequence and is represented by different number of cells.



Jiang, 2017, Curr Op in Biomed. Eng.

Immunogenomic architecture of antibodies and TCRs

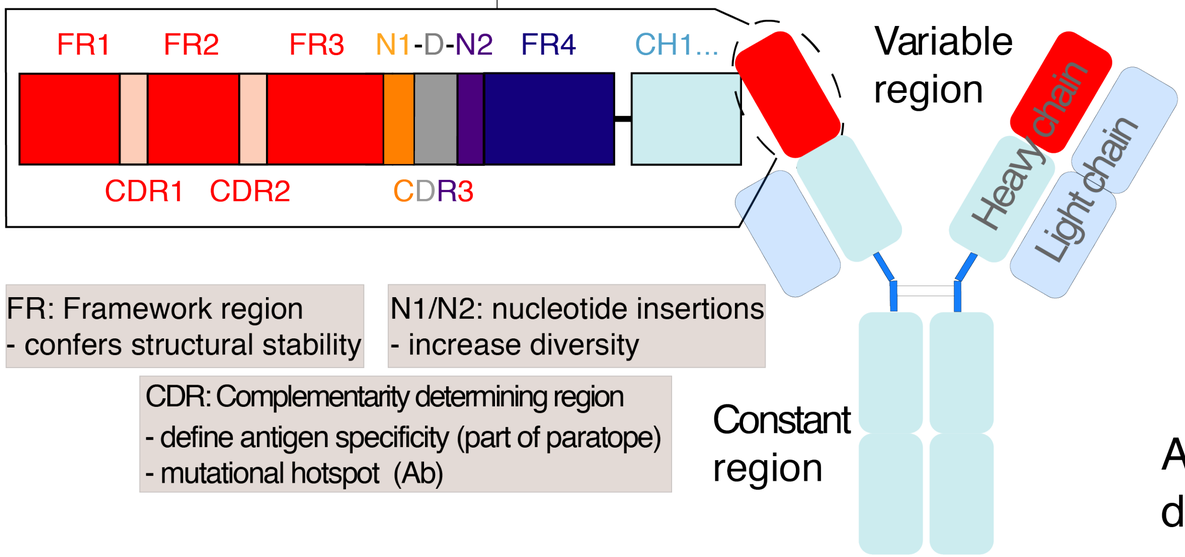
Nucleotides

V, D, J genes
(Diversity: ≈10^2)

V_H segments (55) D_H (27) J_H (9)

VDJ Recombination

Ab/TCR sequence
(Diversity: >10^{13})

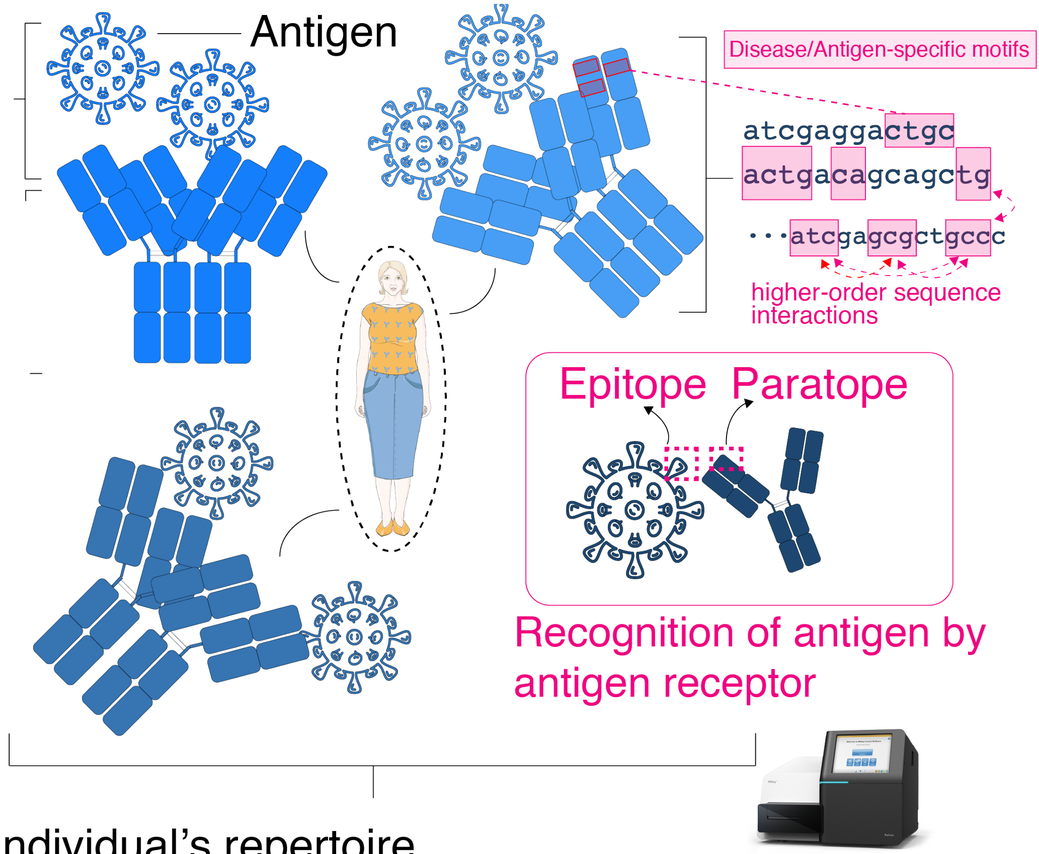


Ab/TCR Sequence structure

FR: Framework region
- confers structural stability

N1/N2: nucleotide insertions
- increase diversity

CDR: Complementarity determining region
- define antigen specificity (part of paratope)
- mutational hotspot (Ab)



An individual's repertoire diversity (10^5 – 10^9) resolved by NGS and proteomics

TCRβ rearrangements without D-segment are common, abundant and public

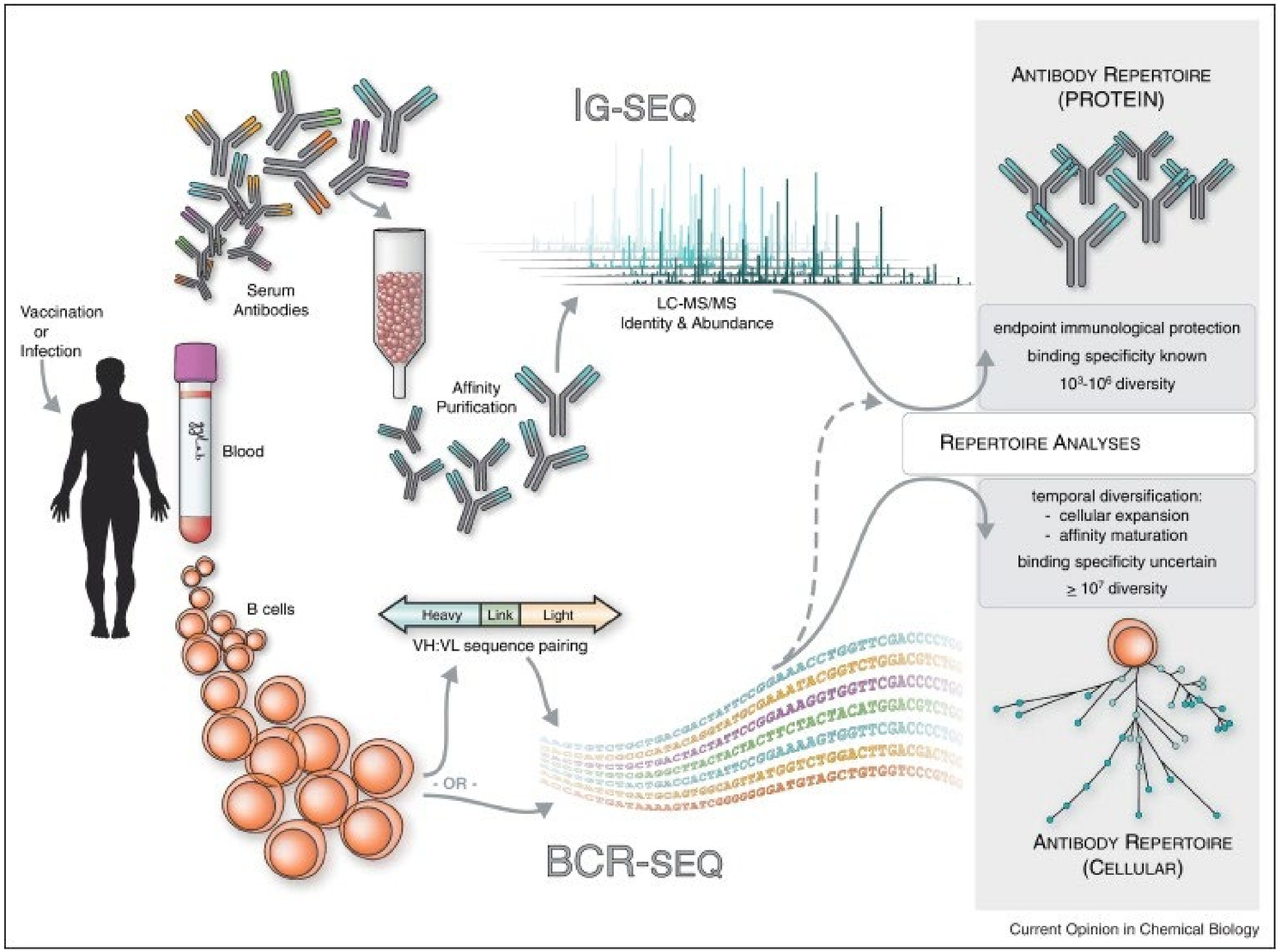
Peter C. de Greef, Rob J. de Boer

doi: https://doi.org/10.1101/2021.03.05.434088

V(DD)J recombination is an important and evolutionarily conserved mechanism for generating antibodies with unusually long CDR3s

Yana Safonova and Pavel A. Pevzner

Genetic and proteomic analysis of the antibody repertoire



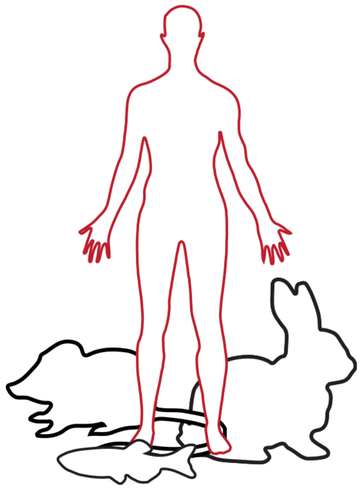
The functional antibody repertoire consists of two major components:

- **the total set of BCRs** expressed on the surface of B cells (genetic analysis)
- **the collection of soluble gut and serum antibody** circulating in the blood (proteomic analysis)

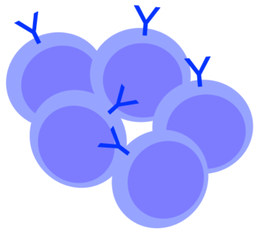
→ Both genomic and proteomic AIRR-seq lead to sequence data. Thus, all downstream computational analytic methods can be applied to both kinds of datasets

Lavinder, Curr Op in Chem Bio, 2015

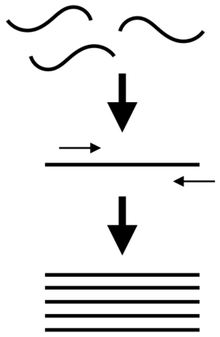
Adaptive immune receptor repertoire sequencing (AIRR-seq)



Identification of suitable study organism and health status
(healthy, diseased, immunized)

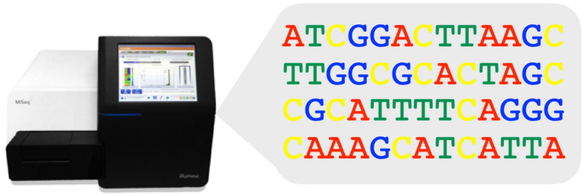


Isolation of T/B-cell population of interest

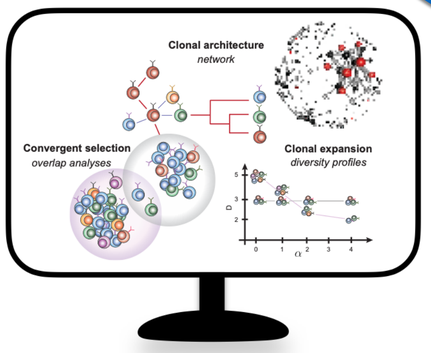


RNA isolation and RT-PCR library preparation

Immune Repertoire Sequencing



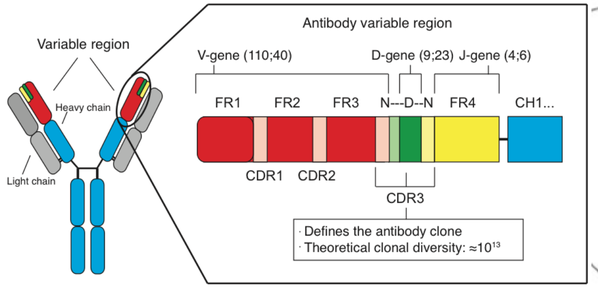
High-throughput sequencing and data preprocessing

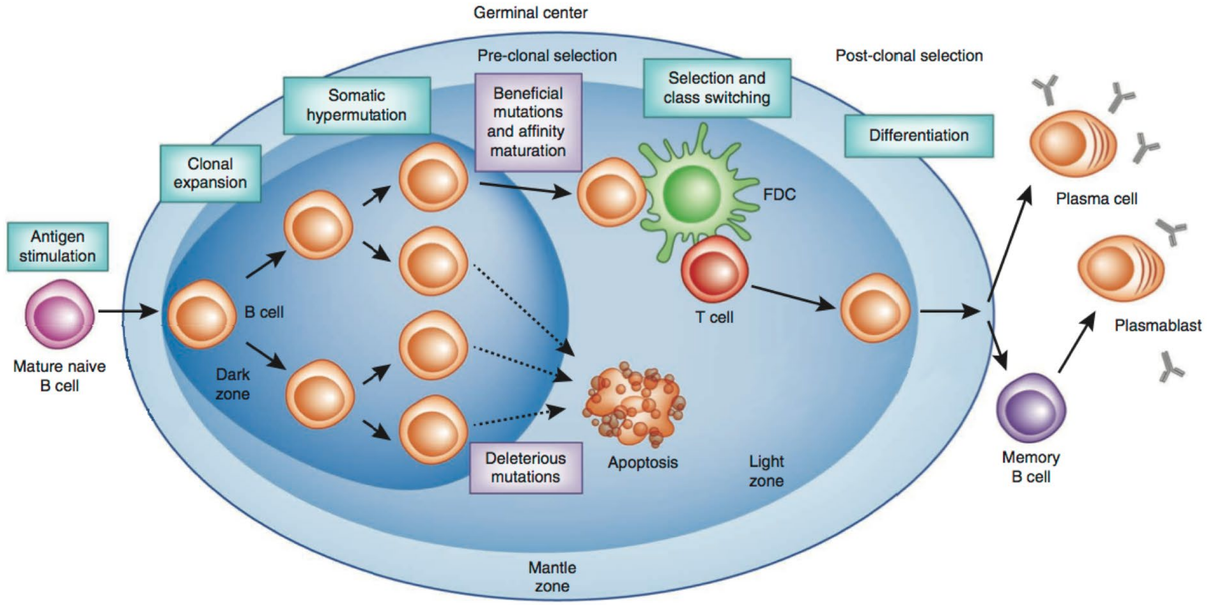


Bioinformatics and statistical data analysis

AIRR-seq = Adaptive immune receptor repertoire sequencing

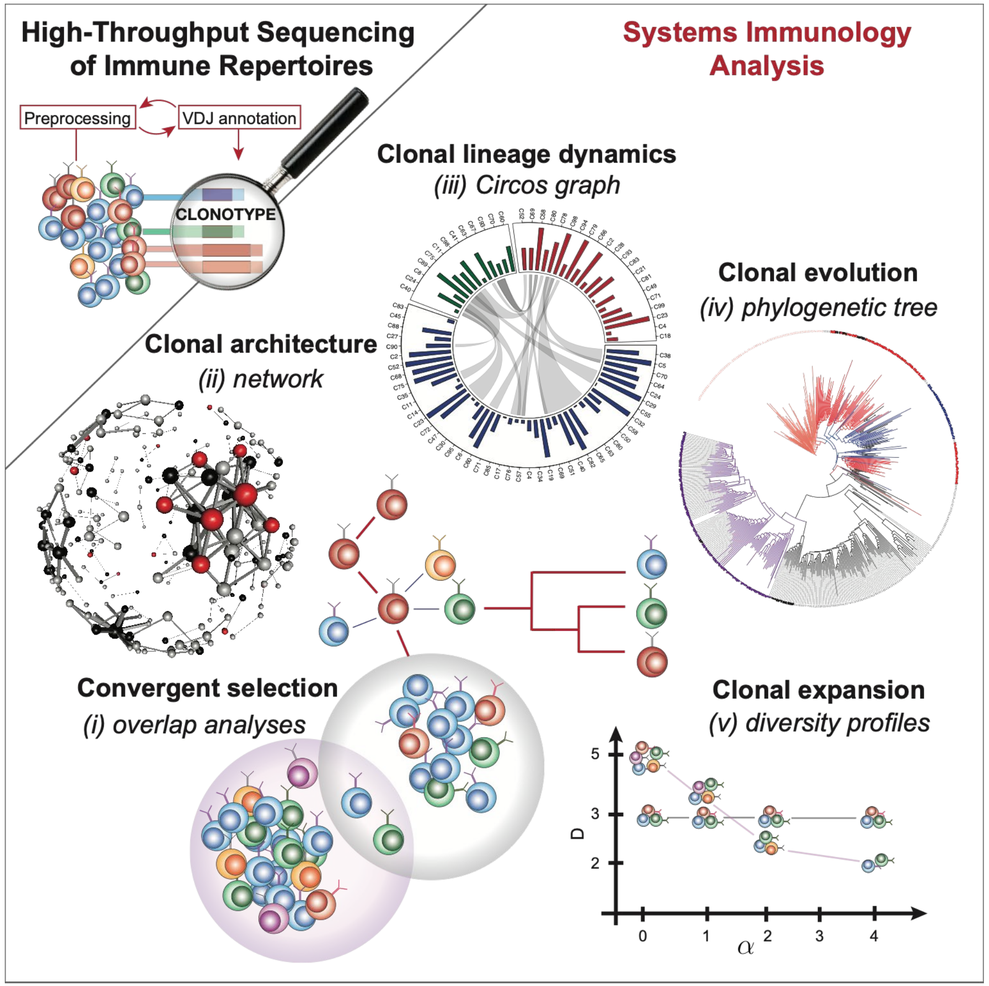
AIRR-seq measures central principles of adaptive immunity





Antigen-specific clonal selection and expansion (evolution)

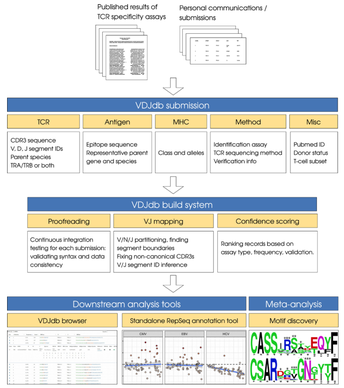
Georgiou, NBT, 2014



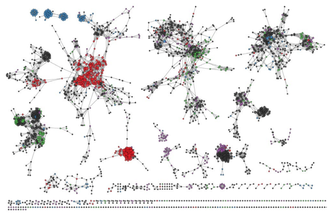
Greiff, Trends Immunol, 2015

Public immune receptor databases (DB)

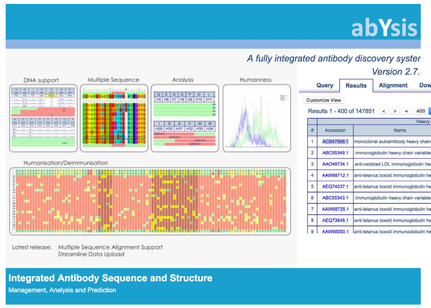
Antigen-specific DBs



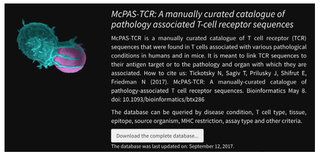
Species	Chain	Records	Paired records	Unique epitopes	Publications
HomoSapiens	TRA	1955	1121	84	79
HomoSapiens	TRB	7614	1112	132	101
MacacaMulatta	TRA	74	0	1	1
MacacaMulatta	TRB	1312	0	3	2
MusMusculus	TRA	1286	1286	22	15
MusMusculus	TRB	1481	1309	29	17



Shugay et al., NAR, 2019, **VDJDB**



abysis, Swindells, JMB, 2017



Tickotsky, Bioinformatics, 2017, McPAS-TCR

Epitope Specific Antibodies and T Cell Receptors in the Immune Epitope Database

Mahajan Front Imm, 2018, **IEDB**

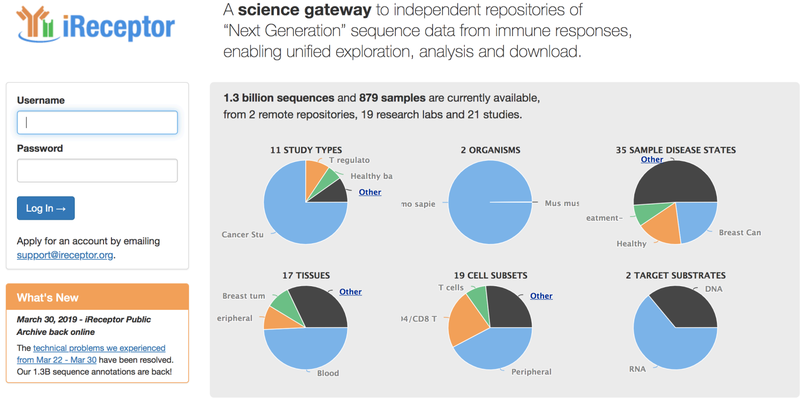
*Swapnil Mahajan[1], Randi Vita[1], Deborah Shackelford[1], Jerome Lane[1], Veronique Schulten[1], Laura Zarebski[1], Martin Closter Jespersen[2], Paolo Marcatili[2], Morten Nielsen[2,3], Alessandro Sette[1,4] and Bjoern Peters[1,4]**

[1] *Center for Infectious Disease, La Jolla Institute for Allergy and Immunology, La Jolla, CA, United States,* [2] *Department of Bio and Health Informatics, Technical University of Denmark, Kongens Lyngby, Denmark,* [3] *Instituto de Investigaciones Biotecnológicas, Universidad Nacional de San Martín, Buenos Aires, Argentina,* [4] *University of California San Diego, La Jolla, CA, United States*

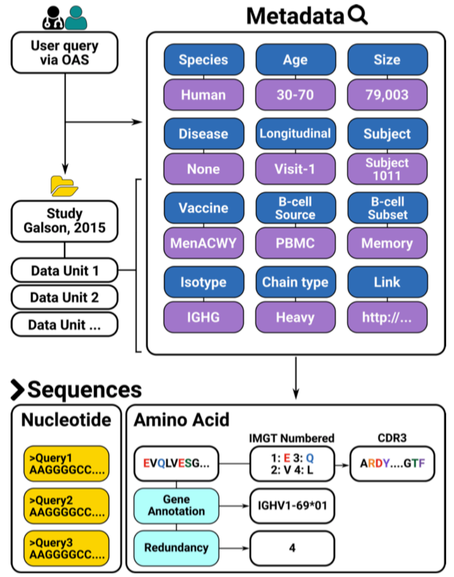
Raybould, Bioinformatics, 2020

Repertoire DBs

Corrie, Immune Rev 2018, **iReceptor**

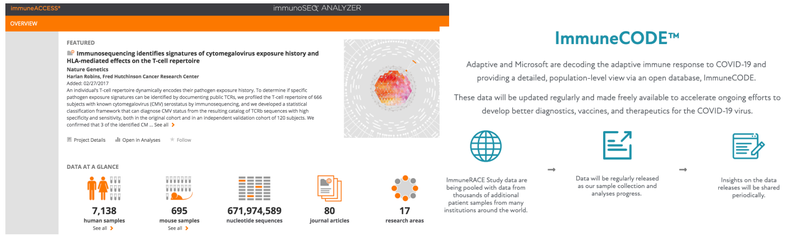


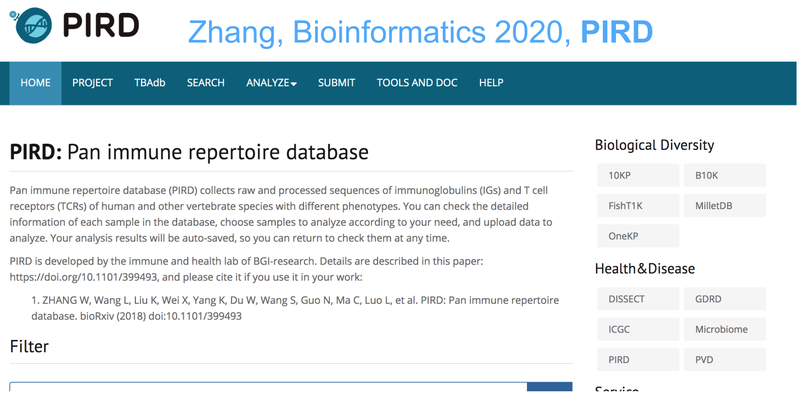
Kovaltsuk, JI 2018, **OAS**



Christley, Front Imm 2018, **VDJServer**

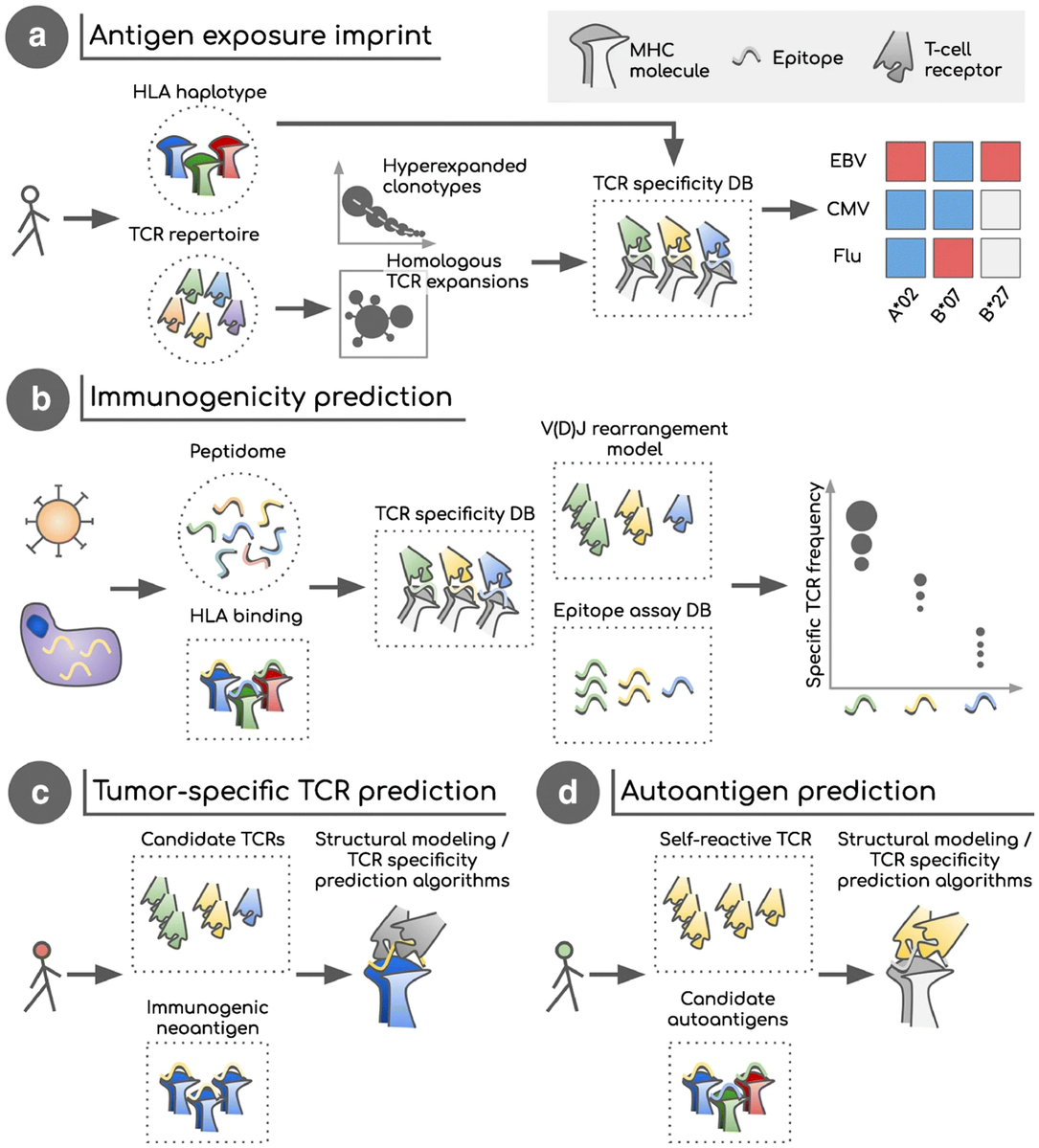
Adaptive Biotech, **immuneAccess, ImmuneCODE**





Zhang, Bioinformatics 2020, **PIRD**

Using antigen-specific public immune receptor databases in AIRR analysis



Zvyagin, Immunogenetics, 2020

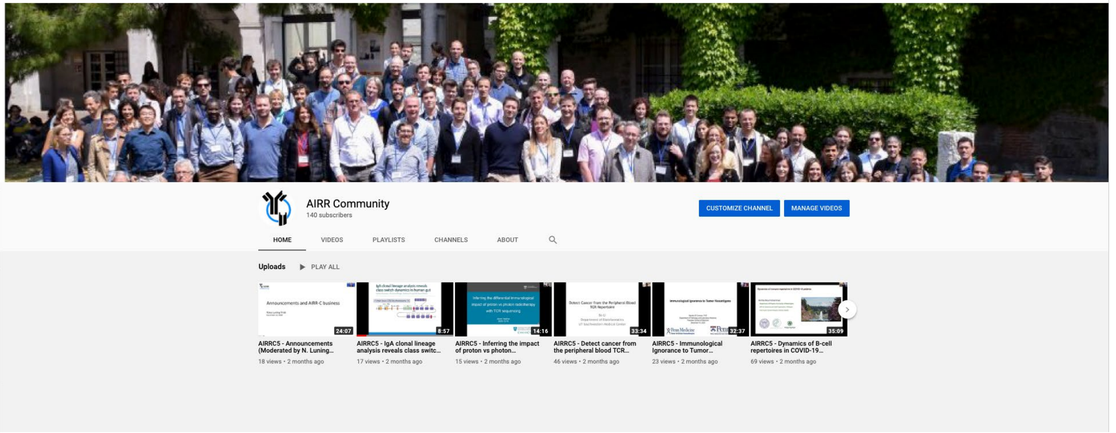
Where to ask experimental and computational AIRR-seq questions? 🤔

B-T.CR

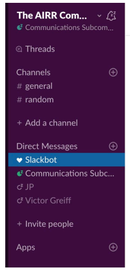
all categories | Latest | Unread (1) | Top | Categories | + New Topic

Topic	Replies	Views	Activity
Interns looking for Integrated Immunology Projects - The Antibody Society	0	52	6d
last visit			
2 Postdoc positions in Exp and Comp Immunology at University of Oslo Open Positions	0	99	27d
Postdoc at Yale School of Medicine (single cell analysis) Open Positions	0	223	Feb 3
Post-doc in Computational Immunology at University of Washington, Seattle Open Positions	1	297	Jan 6
How to incorporate clustering info in AIRR-compliant files Formats	4	144	Jan 6
How do we understand those pseudogene? such as, IGHVII, IGHVIII, and IGHVIV, etc	2	162	Jan 6
Publicly available COVID-19 AIRR-seq data sets 1 8 Wiki	24	5.3k	Jan 1

https://b-t.cr/



https://www.youtube.com/airrcommunity



slack

https://www.antibodysociety.org/airr-community/join-the-airr-community-slack-workspace/

airr_community
588 Tweets

@airr_community

Edit profile

airr_community
@airr_community

The AIRR Community of The Antibody Society is a group that is coordinating stakeholders in the use of NGS technologies to study antibody and TCR repertoires.

antibodysociety.org/the-airr-commu... Joined January 2019

731 Following **688** Followers

Summary: Introduction to AIRR-seq

- The investigation of adaptive immune repertoires requires a high-throughput sequencing approach

- AIRR-seq can be performed both on the genomic and proteomic level

- AIRR-seq measures central principles of adaptive immunity and opens the door to new applications (e.g., monoclonal antibody discovery, immunodiagnostics)

- Many AIRR-seq datasets and antigen-specific receptor sequences are publicly available (e.g., VDJDB, McPAS-TCR, iReceptor, OAS, PIRD)

Outline

Introduction to Adaptive immune receptor repertoire sequencing (AIRR-seq)

- Generation of immune repertoire diversity
- Workflow and applications of AIRR-seq

Error correction and Standardization of AIRR-seq data

- Experimental design and considerations
- Error and bias correction
- Standardization

Single-cell AIRR-seq

- Pairing by targeted amplification
- Single-cell sequencing

Computational strategies for immune repertoire analysis

- Diversity and convergence analysis
- Network analysis
- Machine learning

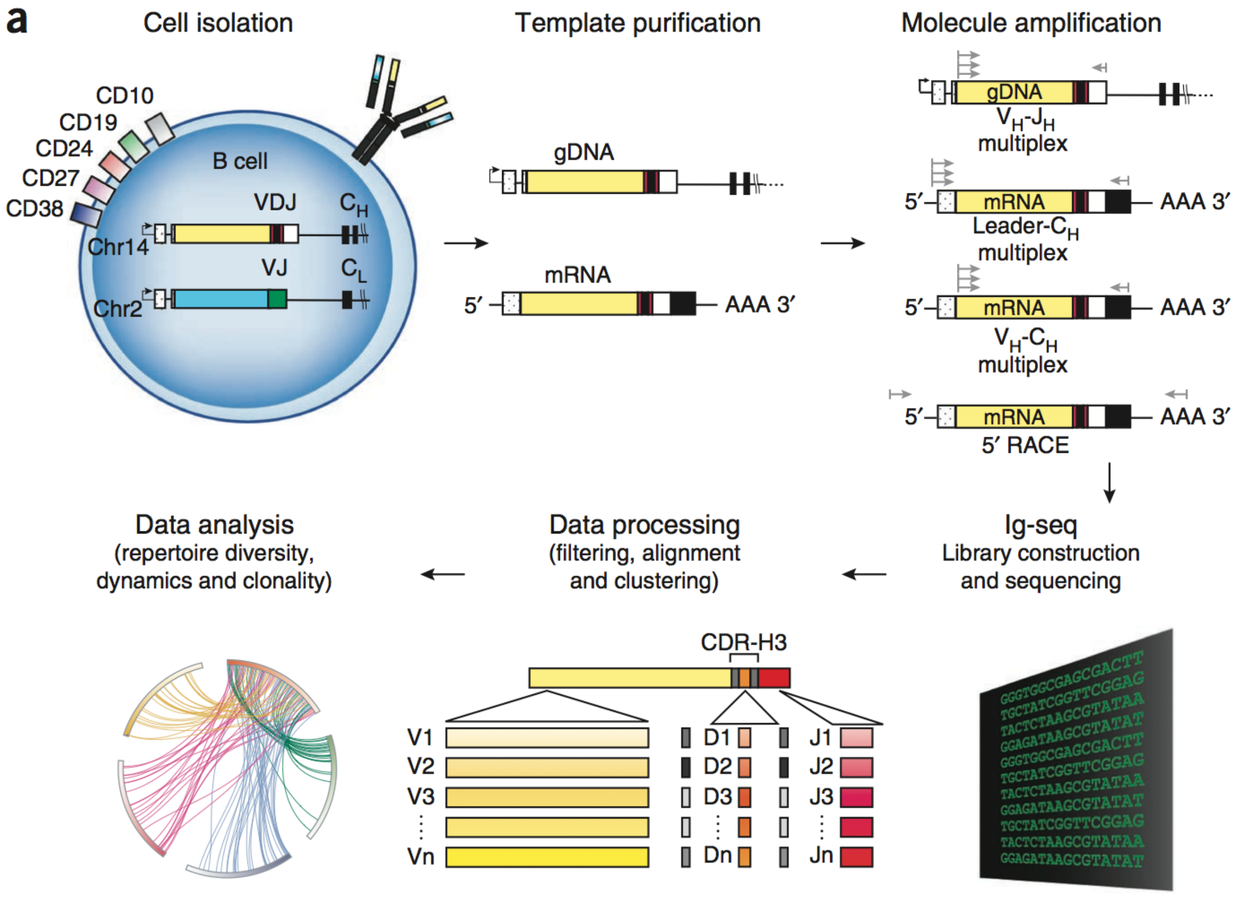
Challenges in experimental immune repertoire data generation

The promise and challenge of high-throughput sequencing of the antibody repertoire

George Georgiou[1–4], Gregory C Ippolito[3,4], John Beausang[5,6], Christian E Busse[7], Hedda Wardemann[7] & Stephen R Quake[5,6,8,9]

FEBRUARY 2014 NATURE BIOTECHNOLOGY

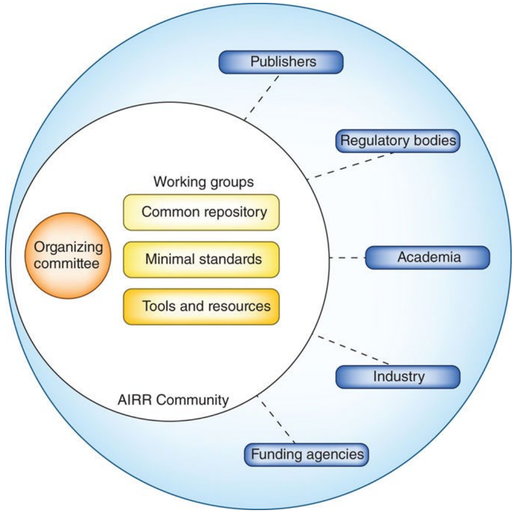
*"…broader application of Ig-seq, especially in clinical settings, will require development of **standardized experimental design framework** that will enable the sharing and meta-analysis of sequencing data generated by different laboratories."*

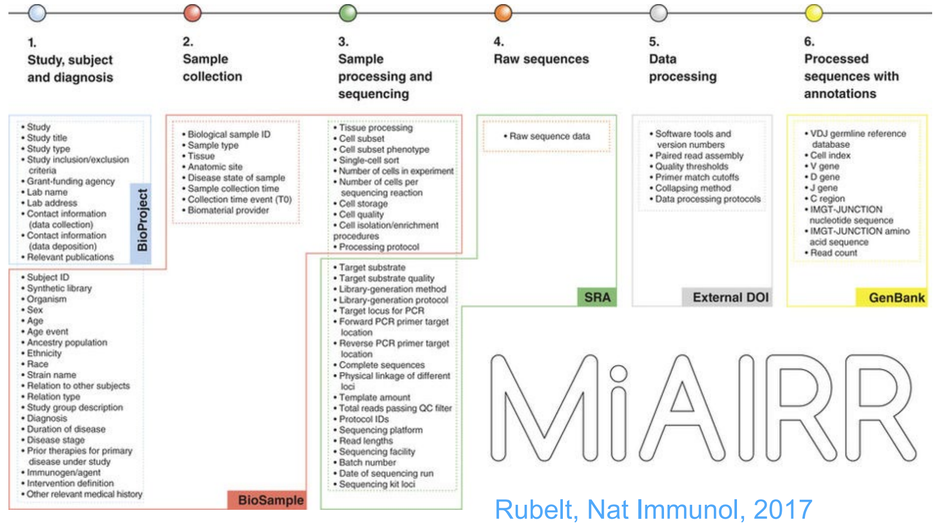


Reproducibility and Reuse of Adaptive Immune Receptor Repertoire Data

Felix Breden[1], Eline T. Luning Prak[2]*, Bjoern Peters[3], Florian Rubelt[4], Chaim A. Schramm[5], Christian E. Busse[6], Jason A. Vander Heiden[7], Scott Christley[8], Syed Ahmad Chan Bukhari[9], Adrian Thorogood[10], Frederick A. Matsen IV[11], Yariv Wine[12], Uri Laserson[13], David Klatzmann[14], Daniel C. Douek[5], Marie-Paule Lefranc[15], Andrew M. Collins[16], Tania Bubela[17], Steven H. Kleinstein[9], Corey T. Watson[18], Lindsay G. Cowell[8], Jamie K. Scott[19] and Thomas B. Kepler[20,21]*

Breden, Front Imm, 2017





MiAIRR

Rubelt, Nat Immunol, 2017

Standardization efforts of the AIRR Community

AIRR Community

frontiers in Immunology

PERSPECTIVE
published: 01 November 2017
doi: 10.3389/fimmu.2017.01418

Reproducibility and Reuse of Adaptive Immune Receptor Repertoire Data

Felix Breden[1], Eline T. Luning Prak[2]*, Bjoern Peters[3], Florian Rubelt[4], Chaim A. Schramm[5], Christian E. Busse[6], Jason A. Vander Heiden[7], Scott Christley[8], Syed Ahmad Chan Bukhari[9], Adrian Thorogood[10], Frederick A. Matsen IV[11], Yariv Wine[12], Uri Laserson[13], David Klatzmann[14], Daniel C. Douek[5], Marie-Paule Lefranc[15], Andrew M. Collins[16], Tania Bubela[17], Steven H. Kleinstein[9], Corey T. Watson[18], Lindsay G. Cowell[8], Jamie K. Scott[19] and Thomas B. Kepler[20,21]*

Inferred Allelic Variants of Immunoglobulin Receptor Genes: A System for Their Evaluation, Documentation, and Naming

Mats Ohlin[1*], Cathrine Scheepers[2,3], Martin Corcoran[4], William D. Lees[5], Christian E. Busse[6], Davide Bagnara[7], Linnea Thörnqvist[1], Jean-Philippe Bürckert[8], Katherine J. L. Jackson[9], Duncan Ralph[10], Chaim A. Schramm[11], Nishanth Marthandan[12], Felix Breden[13], Jamie Scott[14], Frederick A. Matsen IV[10], Victor Greiff[15], Gur Yaari[16], Steven H. Kleinstein[17], Scott Christley[18], Jacob S. Sherkow[19], Sofia Kossida[20], Marie-Paule Lefranc[20], Menno C. van Zelm[21], Corey T. Watson[22] and Andrew M. Collins[23*]

frontiers in Immunology

Front Immunol. 2018; 9: 2206.
Published online 2018 Sep 28. doi: 10.3389/fimmu.2018.02206

PMCID: PMC6173121
PMID: 30323809

AIRR Community Standardized Representations for Annotated Immune Repertoires

Jason Anthony Vander Heiden,[1,†] Susanna Marquez,[2] Nishanth Marthandan,[3] Syed Ahmad Chan Bukhari,[2] Christian E. Busse,[4] Brian Corrie,[5] Uri Hershberg,[6,7,8] Steven H. Kleinstein,[2,9] Frederick A. Matsen IV,[10] Duncan K. Ralph,[10] Aaron M. Rosenfeld,[6] Chaim A. Schramm,[11] The AIRR Community,[‡] Scott Christley,[12,*†] and Uri Laserson[13,*]

http://docs.airr-community.org/en/latest/swtools/airr_swtools_standard.html

frontiers in Immunology

Adaptive Immune Receptor Repertoire Community recommendations for sharing immune-repertoire sequencing data

Florian Rubelt[1,21], Christian E Busse[2,21], Syed Ahmad Chan Bukhari[3,21], Jean-Philippe Bürckert[4], Encarnita Mariotti-Ferrandiz[5], Lindsay G Cowell[6], Corey T Watson[7], Nishanth Marthandan[8], William J Faison[9], Uri Hershberg[10], Uri Laserson[11], Brian D Corrie[12,13], Mark M Davis[1,14], Bjoern Peters[15], Marie-Paule Lefranc[16], Jamie K Scott[8,12,17], Felix Breden[12,13], The AIRR Community[18], Eline T Luning Prak[19,22] & Steven H Kleinstein[3,20,22]

Front Immunol. 2018; 9: 1877.
Published online 2018 Aug 16. doi: 10.3389/fimmu.2018.01877

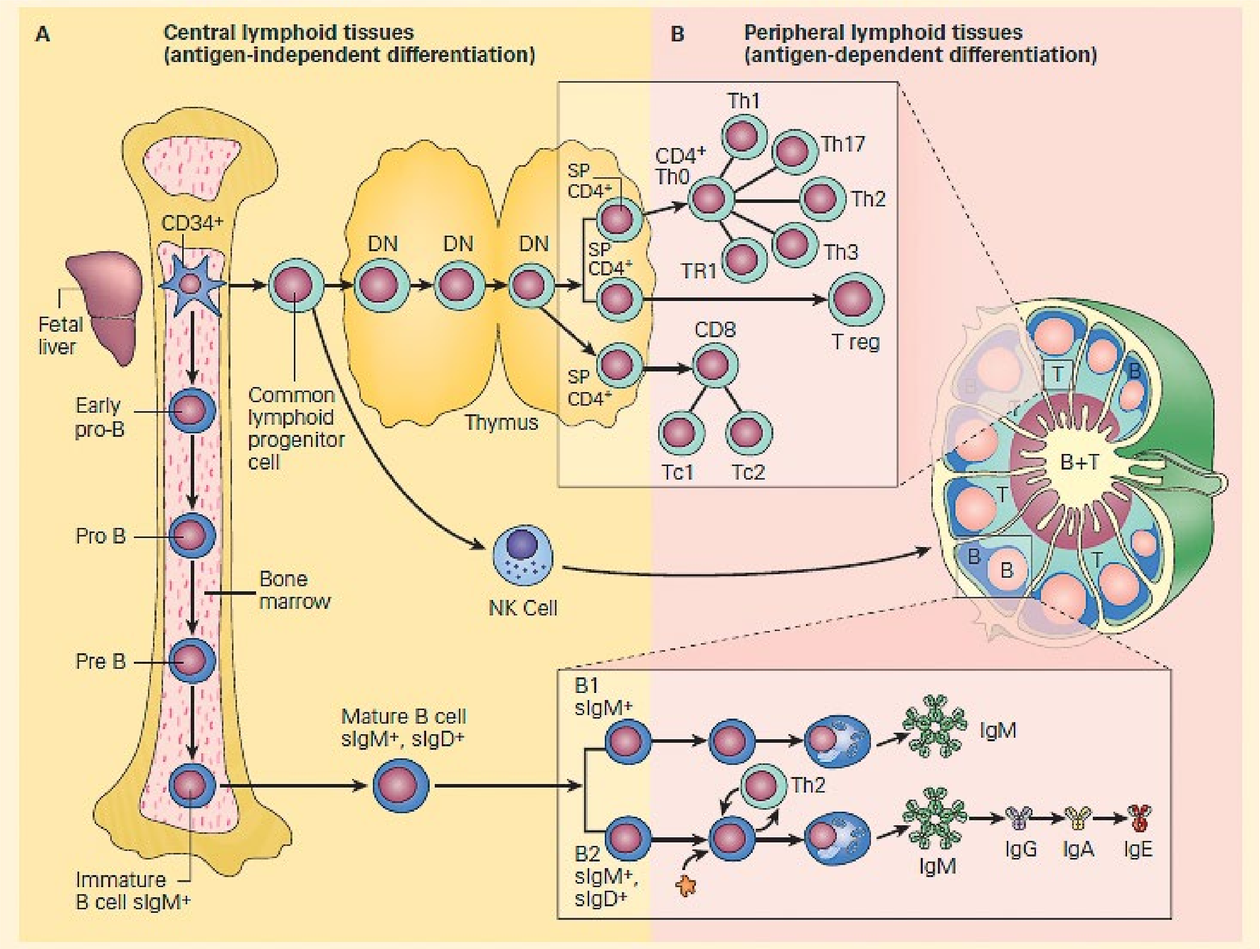
PMCID: PMC6105692
PMID: 30166985

The CAIRR Pipeline for Submitting Standards-Compliant B and T Cell Receptor Repertoire Sequencing Studies to the National Center for Biotechnology Information Repositories

Syed Ahmad Chan Bukhari,[1] Martin J. O'Connor,[2] Marcos Martínez-Romero,[2] Attila L. Egyedi,[2] Debra Willrett,[2] John Graybeal,[2] Mark A. Musen,[2] Florian Rubelt,[3] Kei-Hoi Cheung,[4,5,6,†] and Steven H. Kleinstein[1,6,*†]

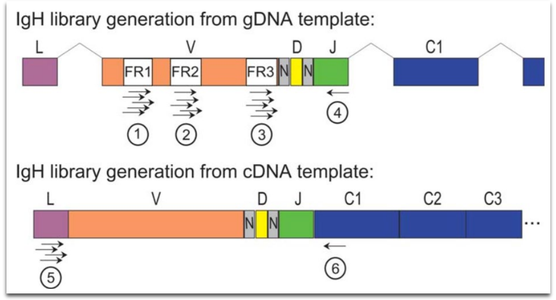
Source of B and T cells should be carefully considered

B and T cell subsets are genetically and functionally diverse



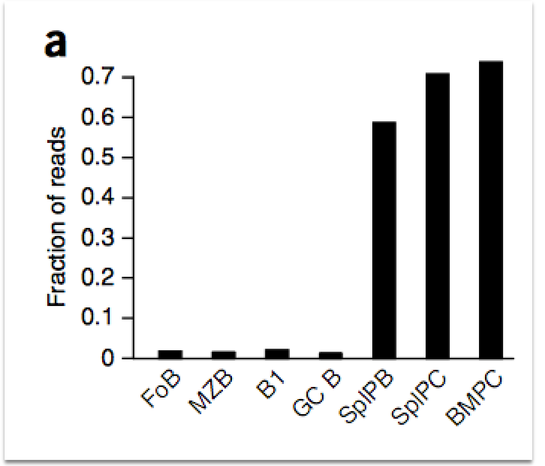
Bellanti, JA (Ed). Immunology IV: Clinical Applications in Health and Disease. I Care Press, Bethesda, MD, 2012

Genomic DNA vs. mRNA for antibody/TCR library generation



Boyd, Microbiol Spectrum, 2014.

- DNA allows easier correlation of clone and cell counts
- DNA does not allow IgH isotype analysis

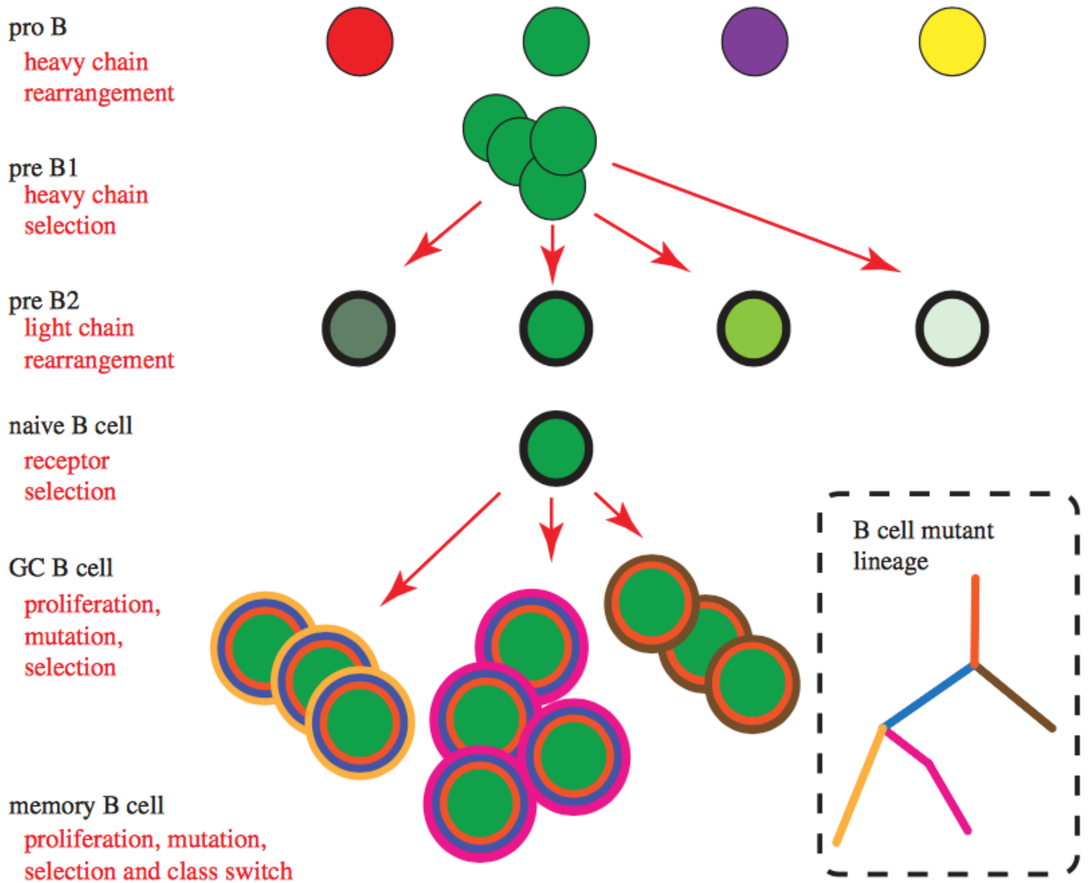


Shi, Nature Immunol, 2015.

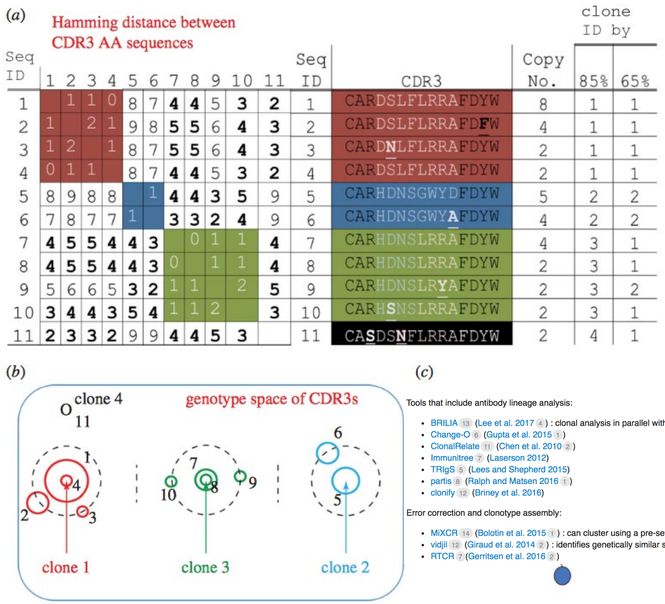
- For RNA-based amplication, antibody producing cells (PB, PC) may bias immune receptor datasets

Definition/computation of clonal (clonotype) family assignment

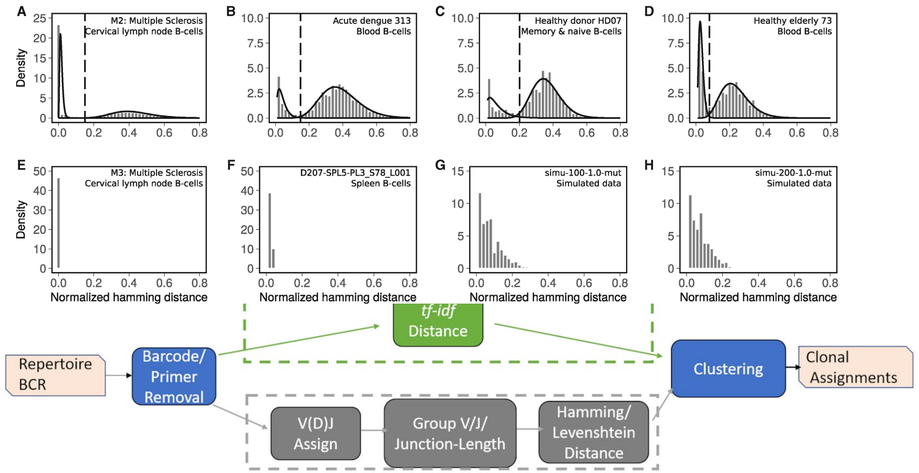
Any two sequences with the same CDR3 are presumed to be clonally related (originate from same B cell clonal lineage)



Hershberg, Phil Trans B, 2015.

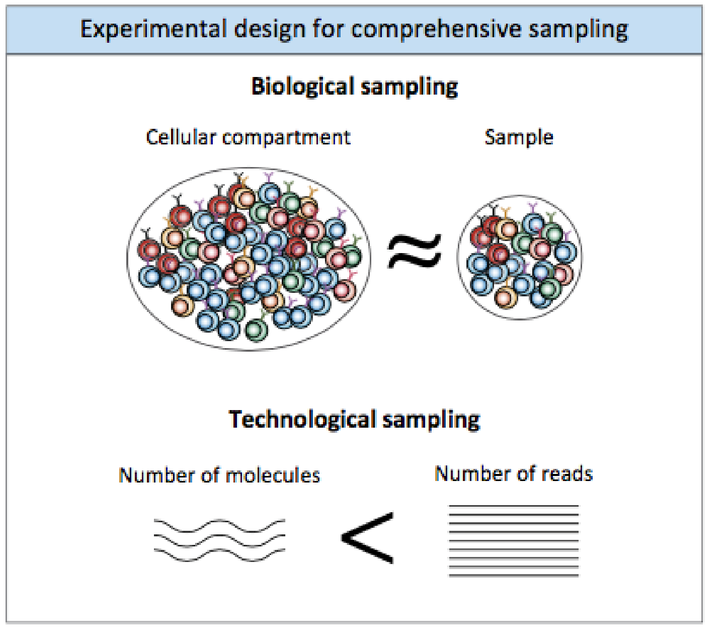


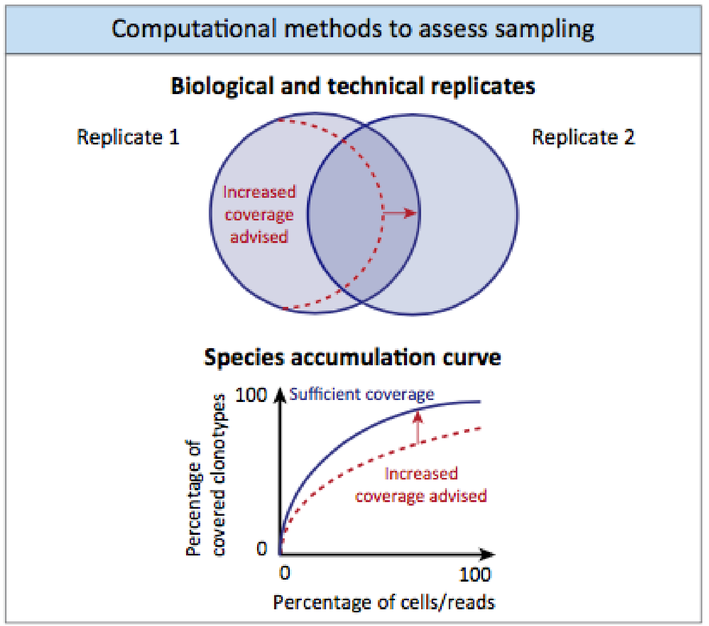
https://b-t.cr/t/list-of-b-cell-clonal-identification-software/22

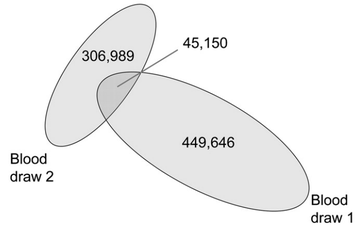


Nouri, 2018,2020, Front Imm/Bioinformatics

Sampling depth determines biological and technological coverage







Warren, Gen Res, 2011

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.

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

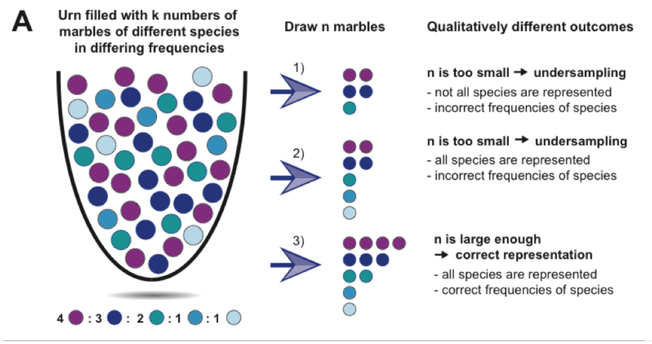
Biological replicates: HTS (high-througput sequencing) of different samples of the same underlying cell population [e.g., partitioning of PBMC (peripheral blood mononuclear cells)]. Biological replicates are used to assess biological sampling.

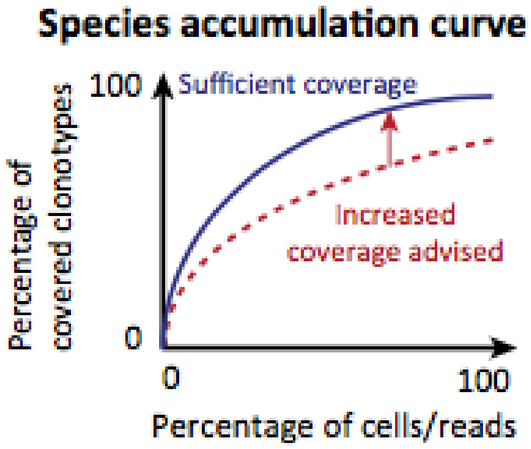
Technical replicates: replicate sequencing of the same immune repertoire library. A strict definition would be the resequencing of the same library, whereas a more lenient definition would consider also molecular replicates (separate library preparation of the same genetic material) adequate provided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

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.

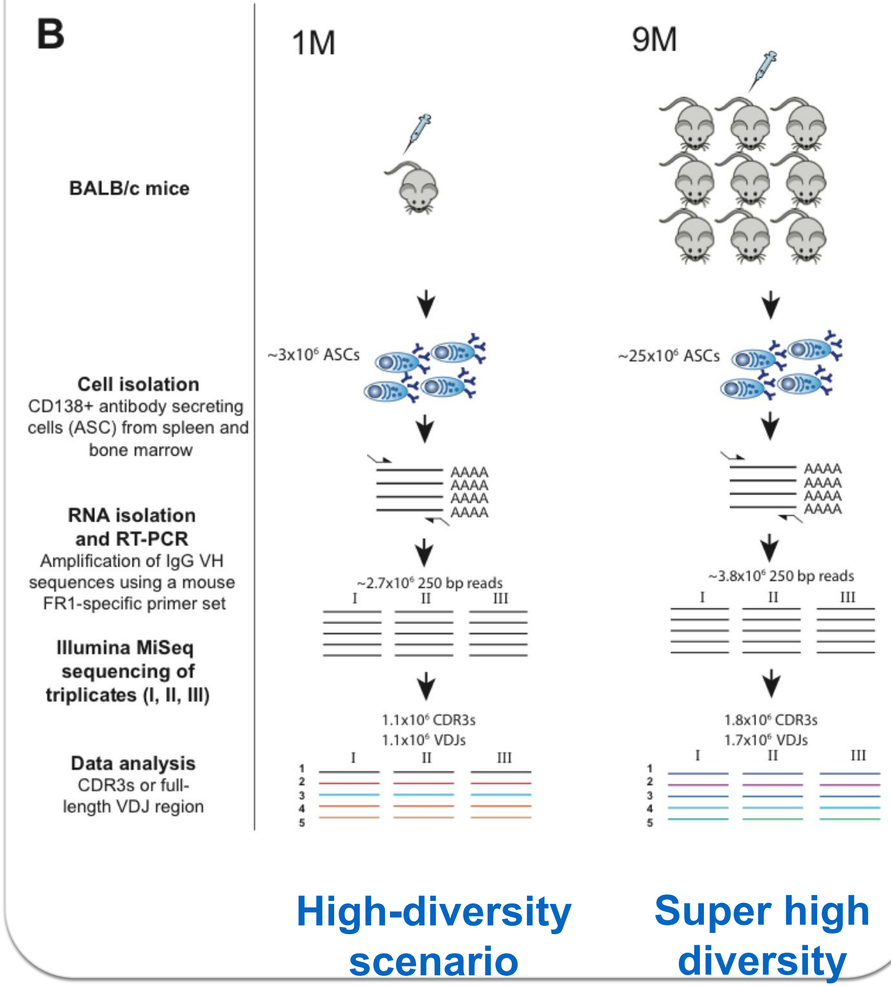
Greiff, Trends Immunol, 2015

Testing sample coverage by species accumulation curves (mouse)



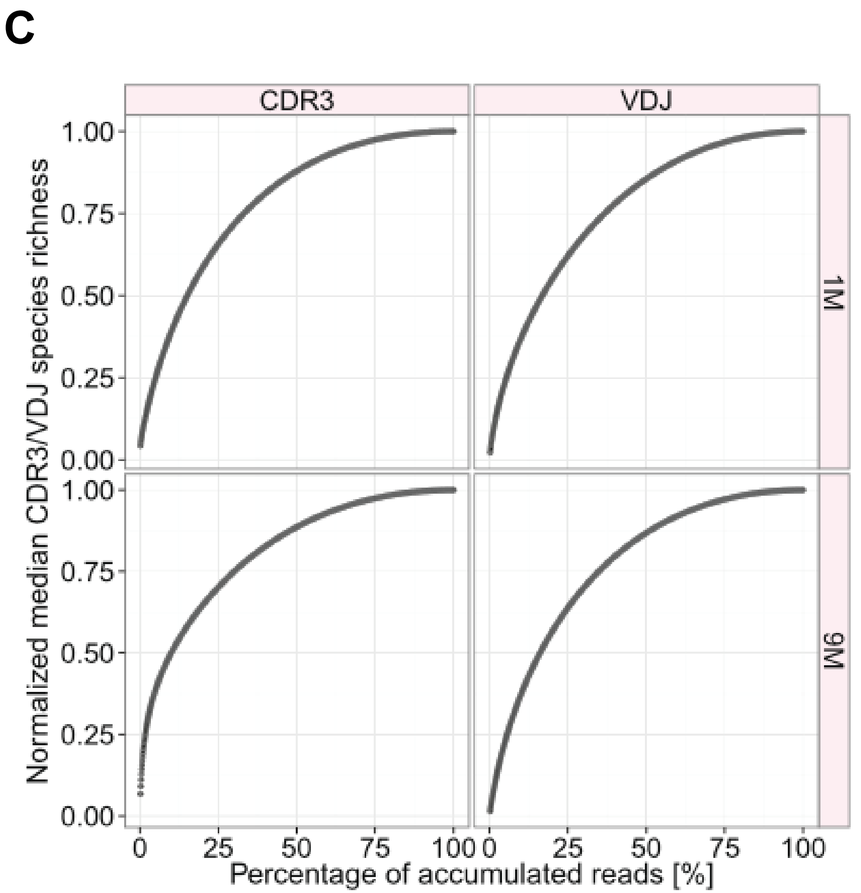


Greiff, Trends Immunol, 2015.



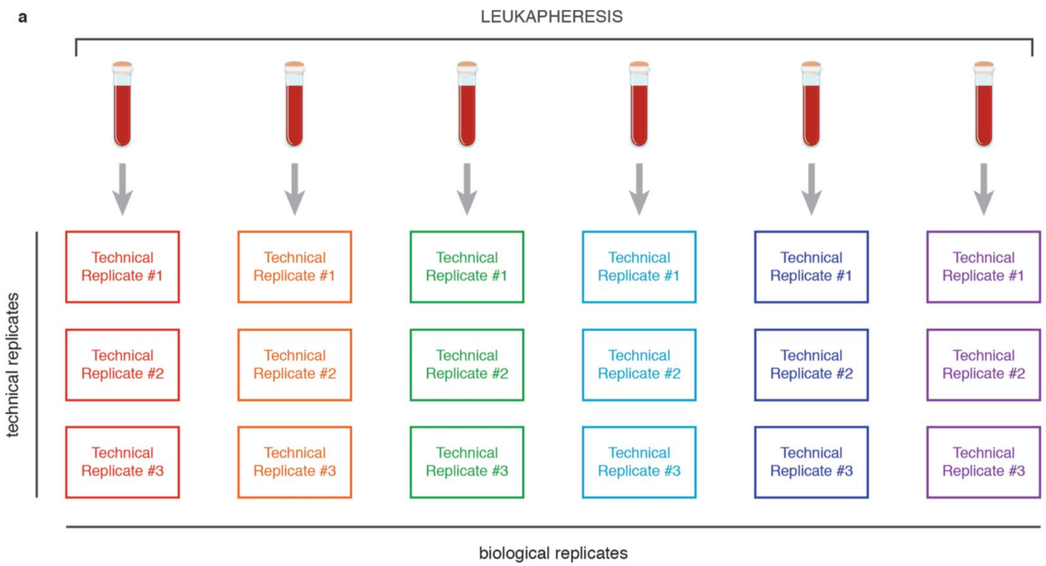
High-diversity scenario

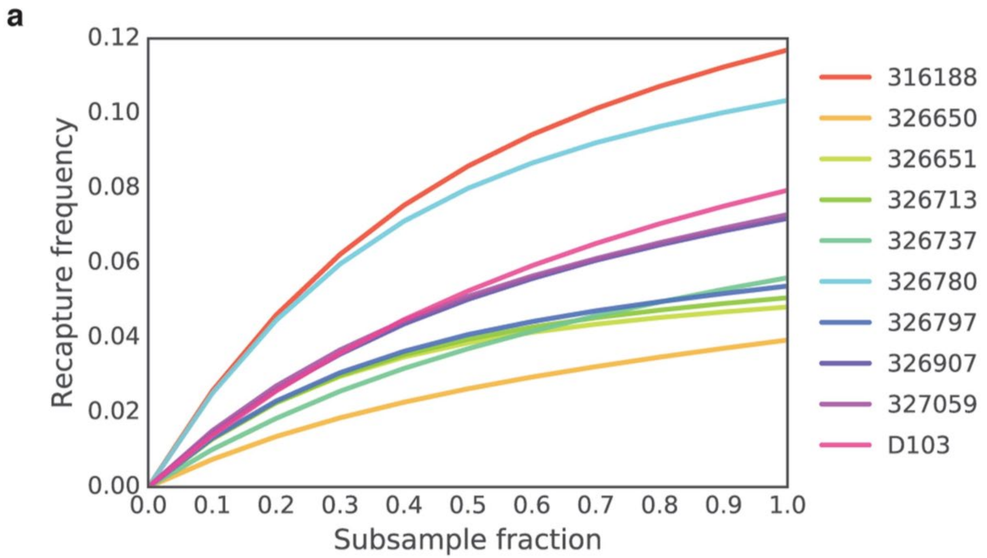
Super high diversity

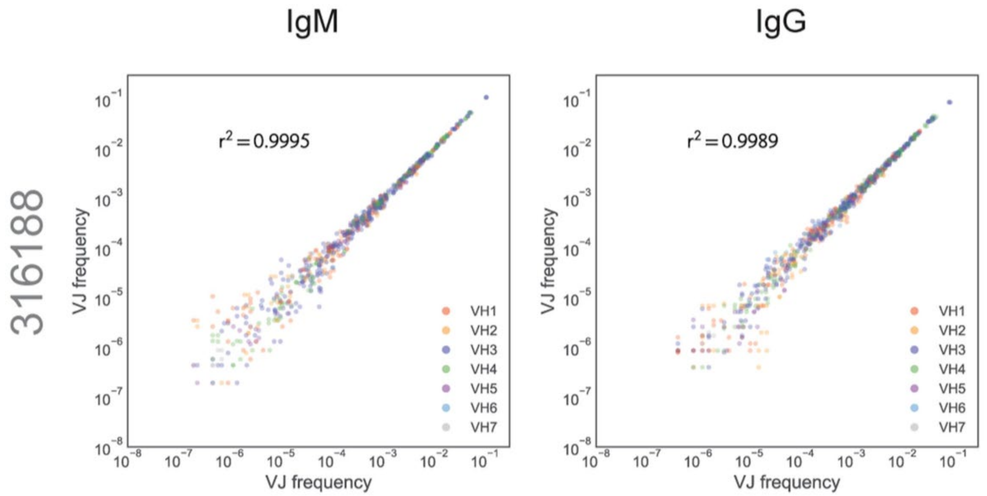


Greiff, BMC Immunol, 2014

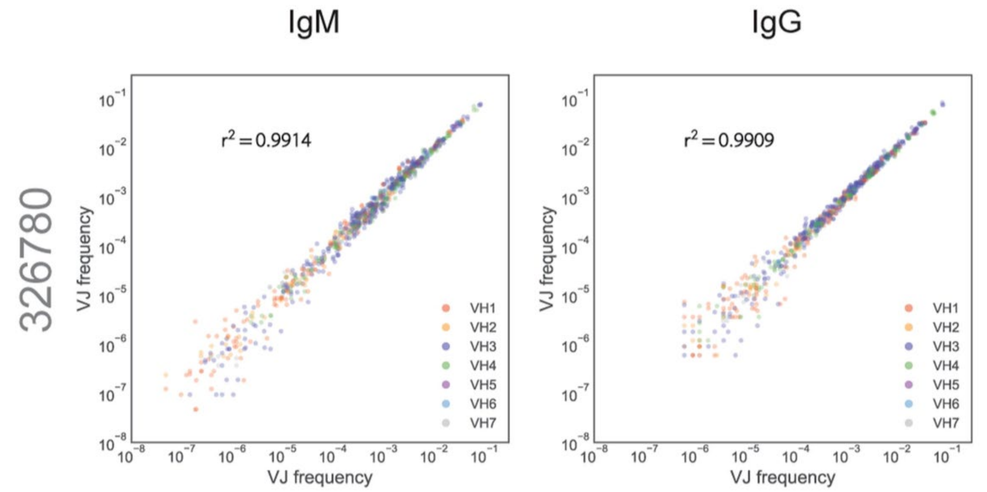
Testing sample coverage by species accumulation curves (human)







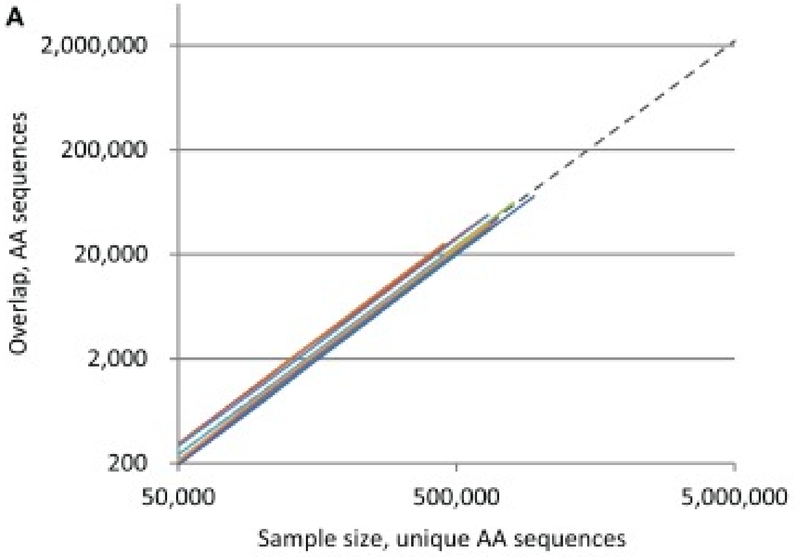
Technical replicate

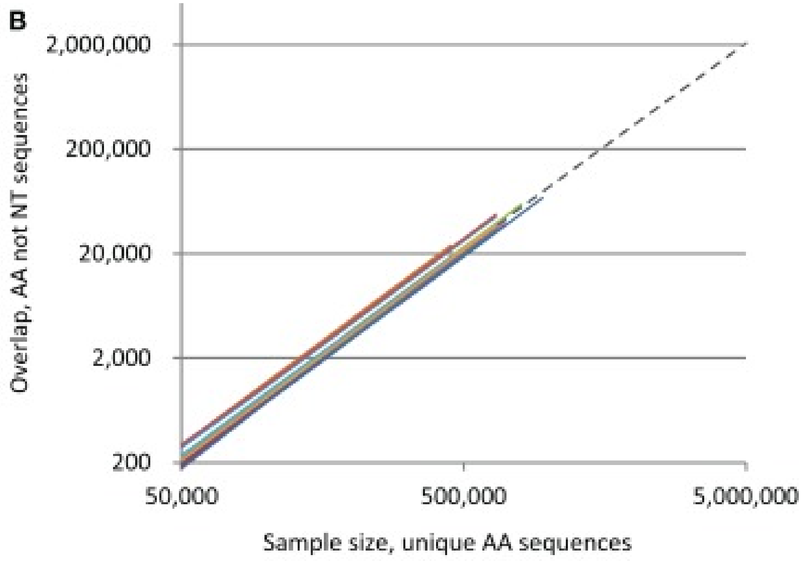


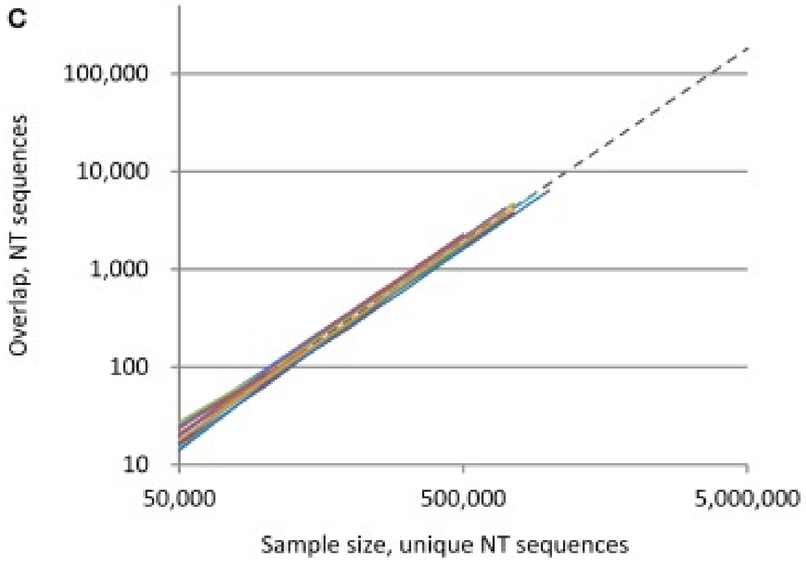
Biological replicate

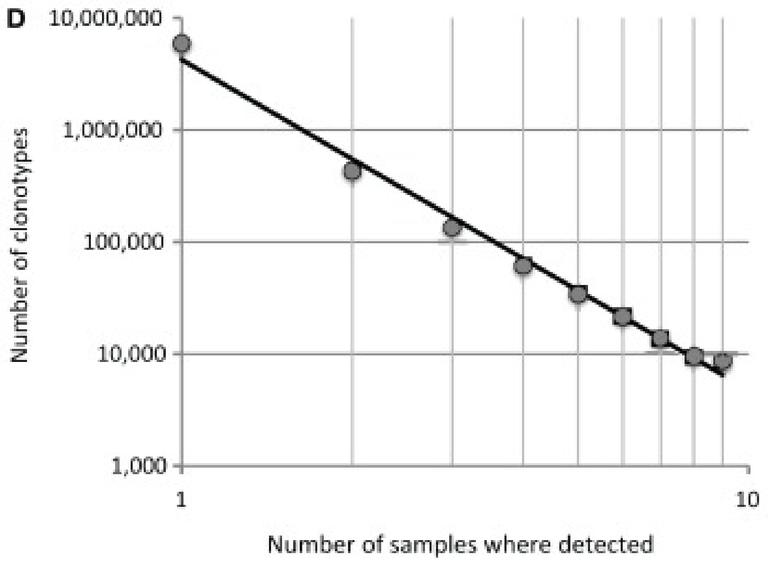
Briney, Nature, 2019

Higher coverage leads to higher discovery of public clones









Overlap of individual TCR beta CDR3 repertoires grows geometrically with the number of sequence pairs sampled. Plots indicate the number of shared sequences for 12 unrelated donor pairs in relation to sample size at the level of

(A) all amino acid sequences,

(B) amino acid sequence only, excluding matches with identical nucleotide sequences, and

(C) nucleotide sequences. Each of the 12 colored lines represents the observed overlap between randomly drawn samples of

unique CDR3 variants for a different pair of unrelated donors. To extrapolate the predicted level of overlap if the full individual TCR beta repertoires were to be sampled, we plotted fittings of averaged data with a power law ($Y = aX^b$) as dashed lines.

(D) We plotted the degree to which unique clonotypes were shared among our nine donors, and found that the frequency with which TCR beta clonotypes occur in human repertoires is distributed according to a power law.

Shugay, Front Imm, 2013

Ensuring exp. data reliability by replicate sequencing I

Definition of a clone

1. CDR3 with exact (100% identity) a.a. sequence
2. Full-length VDJ with exact (100%) a.a. sequence

A Bioinformatics workflow

Illumina MiSeq raw data 2x250 bp paired-end → Data pre-processing → IMGT-based read annotation → Unique CDR3s ranked by decreasing frequencies

B Statistical analysis

Determination of reliable detection cut-off → Determination of range of reliable detection + Corresponding total reads + Comparison of CDR3 ranking

Example of 85% reliable detection cut-off applied to two hypothetical datasets:

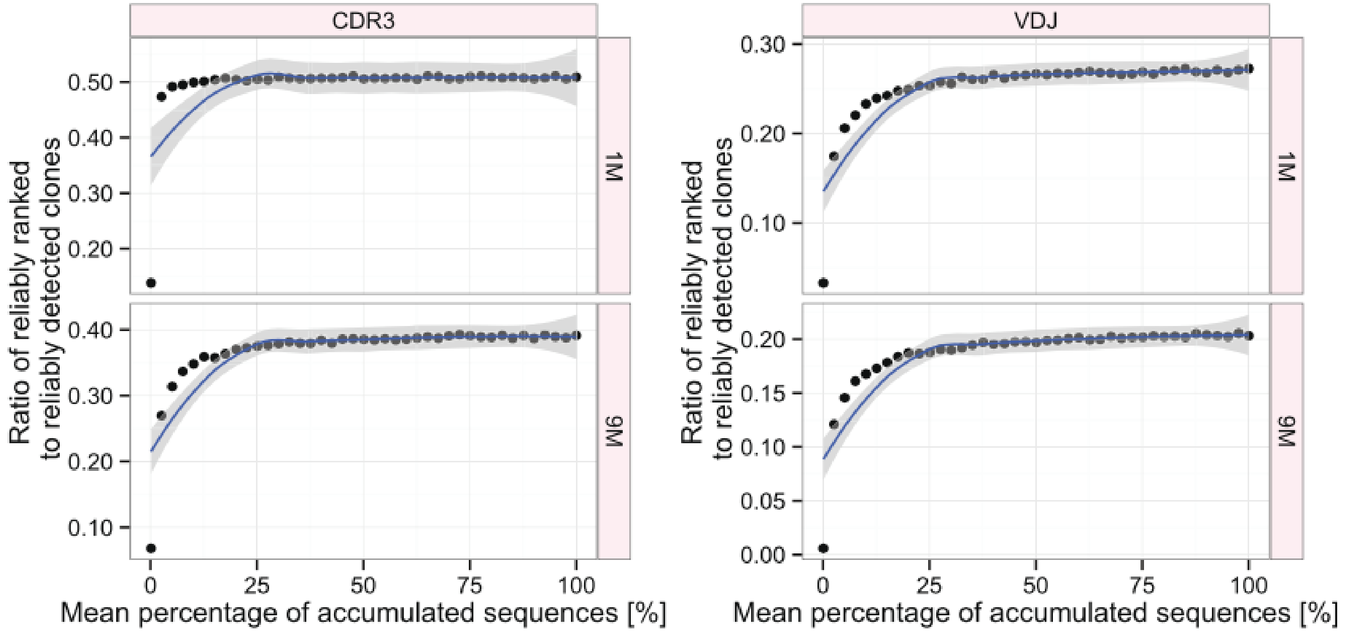
Dataset 1 vs. Dataset 2

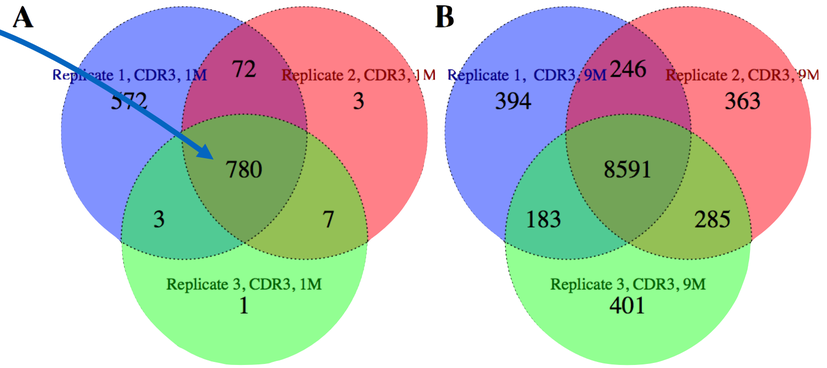
	Ratio	Dataset 1	Freq.	Dataset 2	Freq.
✓	100.0	TRGD	3.2	TRGD	2.9
✓	100.0	ARHAY	1.3	NYYGLA	1.9
✓	100.0	NYYGLA	0.9	ARHAY	0.8
✓	100.0	GFADS	0.7	YGYLN	0.7
✓	100.0	WELGR	0.6	GFADS	0.6
✓	100.0	RLSYYDL	0.6	WELGR	0.5
✓	85.7	TIGGF	0.5	LAWFA	0.4
✓	87.5	YGYLN	0.3	GWFAY	0.3
✗	77.8	RVFFD	0.2	RLSYYDL	0.2
✗	77.0	FPRMDY	0.1	YYGYF	0.1

Dataset 2 vs. Dataset 1

	Ratio	Dataset 2	Freq.	Dataset 1	Freq.
✓	100.0	TRGD	2.9	TRGD	3.2
✓	100.0	NYYGLA	1.9	ARHAY	1.3
✓	100.0	ARHAY	0.8	NYYGLA	0.9
✓	100.0	YGYLN	0.7	GFADS	0.7
✓	100.0	GFADS	0.6	WELGR	0.6
✓	100.0	WELGR	0.5	RLSYYDL	0.6
✓	85.7	LAWFA	0.4	TIGGF	0.5
✗	75.0	GWFAY	0.3	YGYLN	0.3
✗	77.8	RLSYYDL	0.2	RVFFD	0.2
✗	77.0	YYGYF	0.1	FPRMDY	0.1

Reliably detected clones:
clones that are present in all replicates

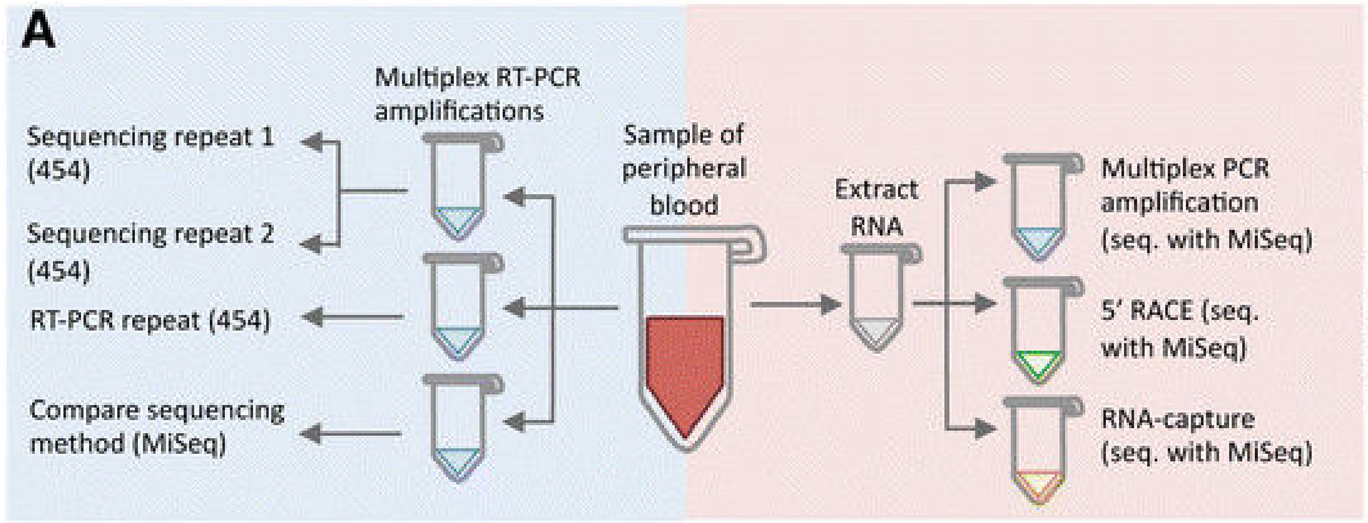


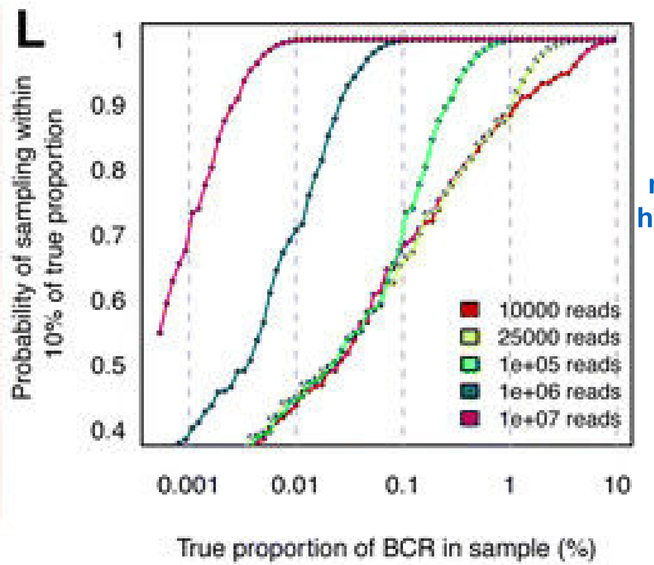


→Complete clonal coverage does not imply correct clonal ranking. To achieve, correct clonal ranking an even higher sampling depth is needed.

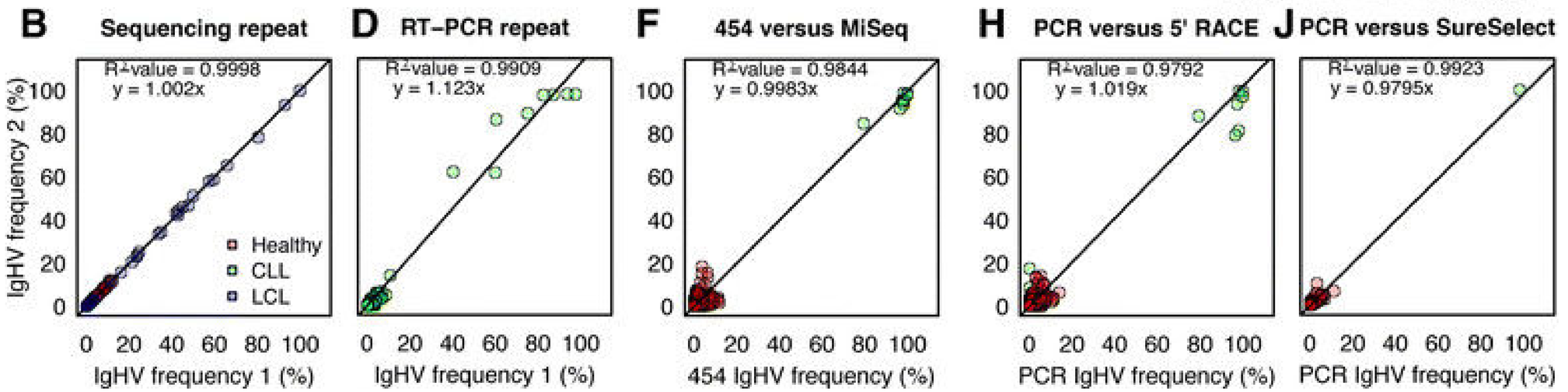
Greiff, BMC Immunol, 2014

Ensuring exp. data reliability by replicate sequencing II



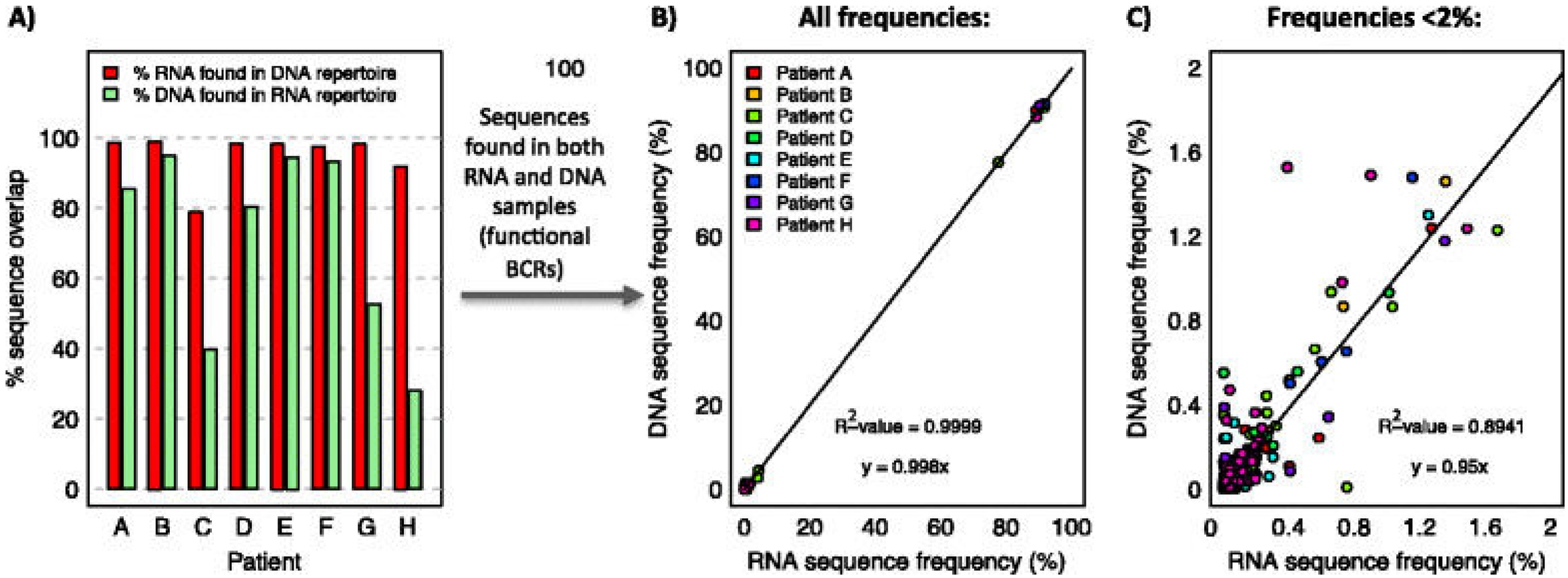


→ **Accurate clonal ranking requires very high sequencing depth**



Bashford-Rogers, BMC Immunol, 2014

Ensuring exp. data reliability: DNA vs RNA sequencing



RNA and DNA were extracted from each peripheral blood sample from 8 CLL patients, on which multiplex RT-PCR or PCR was performed respectively and sequenced by MiSeq (250 bp paired-end).

A) The percentage of DNA sequences found in each RNA sample. The correlation between the BCR frequency in RNA and functional DNA repertoires (DNA sequences that were found also in the RNA repertoire) for the 8 CLL patients in

B) all IgHV gene usage frequencies and

C) the low frequency IgHV gene usage frequencies only (<2%). **If unequal numbers of RNA molecules per cell significantly skewed the RNA BCR repertoires, then deviation from y = x correlation would be expected.**

Bashford-Rogers, BMC Immunol, 2014

Measuring the replicability, reliability and sensitivity of different TCR methods

Table 1 | Comparative performance of the nine TCRseq molecular methods

TR chain	Method	Replicability	Reliability	Sensitivity	Cost per sample ($)	Controls and standards	Format type	fastq data availability
TRA	RACE-1	7	4	4	~230	-	Lab protocol	Yes
	RACE-1_U	4	5	4	~230	UMI	Lab protocol	Yes
	RACE-2	5	4	5	230–280	-	Service or kit	Yes
	RACE-2_U	4	5	5	230–280	UMI	Service or kit	Yes
	RACE-3	3	2	3	~150	-	Kit	Yes
	RACE-4	5	6	4	~150	-	Lab protocol	Yes
	RACE-5	2	3	3	~300	-	Lab protocol	Yes
TRB	mPCR-1	3	3	3	~350–550[a]	Synthetic TCRs	Service or kit	No
	mPCR-2	6	7	7	~25	-	Lab protocol	Yes
	mPCR-3	5	5	3	~350–550[a]	-	Service or kit	Yes
	RACE-1	6	5	4	~230	-	Lab protocol	Yes
	RACE-1_U	4	6	5	~230	UMI	Lab protocol	Yes
	RACE-2	6	6	6	230–280	-	Service or kit	Yes
	RACE-2_U	6	6	7	230–280	UMI	Service or kit	Yes
	RACE-3	2	2	3	~150	-	Kit	Yes
	RACE-4	3	5	4	~150	-	Lab protocol	Yes

For each method, an average rank score for TRA (top) and TRB (bottom) sequencing was calculated for replicability, reliability and sensitivity (first three columns), and practical information was summarized (last four columns). Ranks were calculated as the average of the ranks for results from Figs. 1e, 2c, 3b and 4c for replicability; Figs. 1e, 2b, 4b and 5a,b for reliability; and Figs. 4c and 5b and Supplementary Figs. 2a and 5c for sensitivity. Rank values range from 2 (best) to 7 (worst). Details are provided as Supplementary Data 1. Cost per sample is expressed in US dollars as per current prices for a depth of 1 million TCR sequences per sample on a 25-million-reads sequencing format. The costs cover reagents for library preparation to sequencing. [a]mPCR1 and mPCR3 price ranges correspond to the cost for purchasing either kits (lowest price) or service up to sequencing and basic data analyses from the provider.

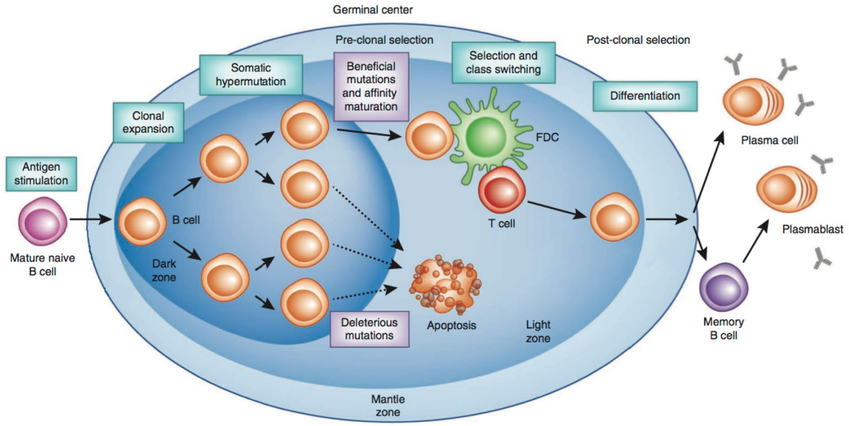
→ **Replicability:** the ability of each method to reproduce the same repertoire from different sub-samples from the same individual)

→ **Reliability:** the extent to which different methods record the same results when applied to the same sample

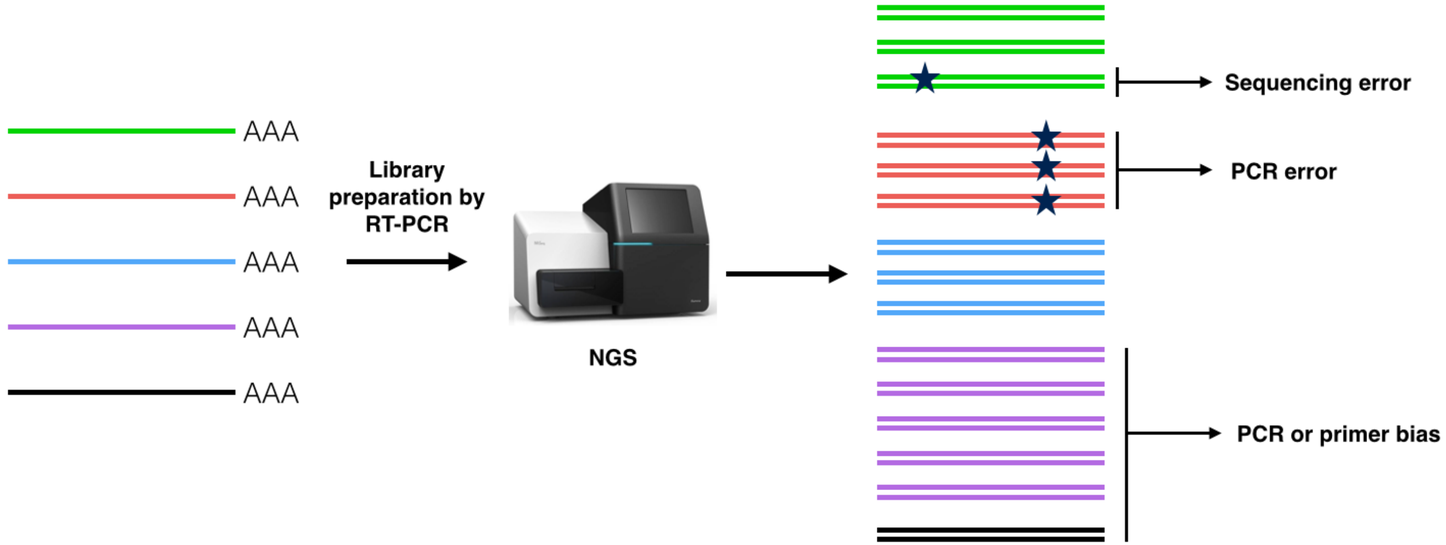
→ **Sensitivity:** capability of a given method to capture low abundant clonotypes

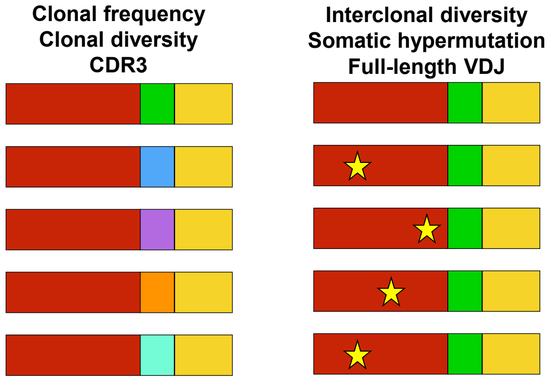
Barennes, NBT, 2020

AIRR-seq errors may compromise immunological interpretation



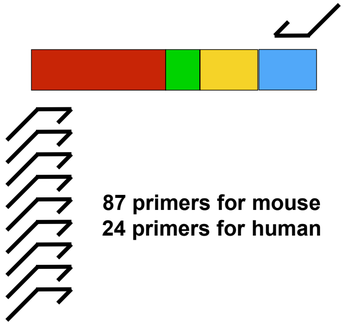
Georgiou, Nature Biotech, 2014.





Errors → Artificial Clonal diversity and somatic hypermutation

Bias → Artificial clonal selection and expansion



Reliable cloning of functional antibody variable domains from hybridomas and spleen cell repertoires employing a reengineered phage display system

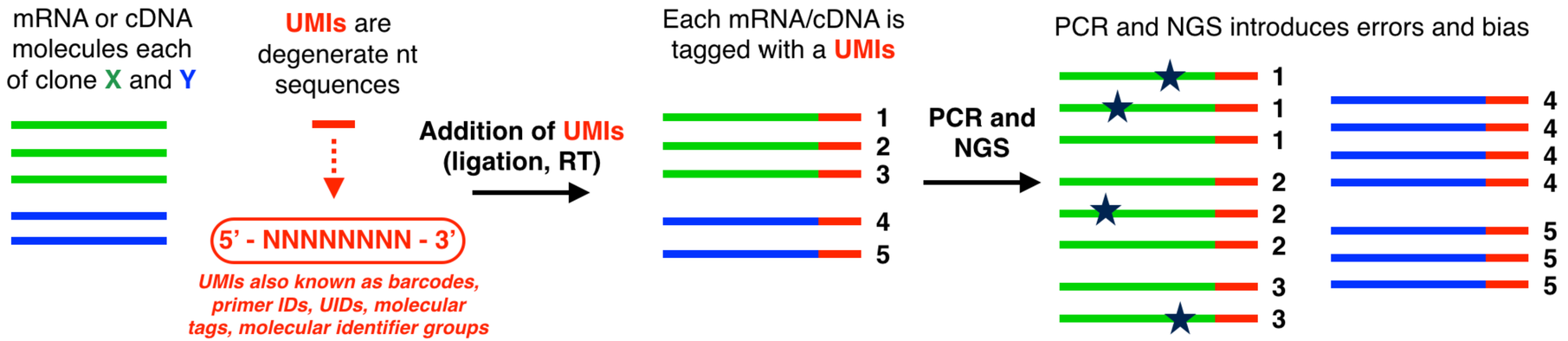
Anke Krebber, Susanne Bornhauser, Jörg Burmester, Annemarie Honegger, Jörg Willuda, Hans Rudolf Bosshard, Andreas Plückthun *

Biochemisches Institut der Universität Zürich, Winterthurerstr. 190, CH-8057 Zürich, Switzerland

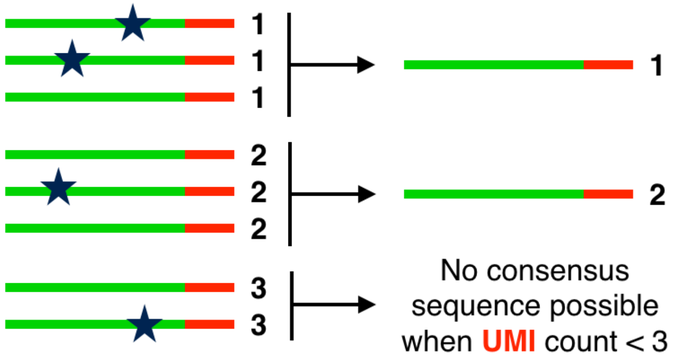
Journal of Immunological Methods 201 (1997) 35–55

Unique molecular identifiers (UMI) for error correction

Experimental library preparation using Unique Molecular Identifiers (UMIs)



Error correction by grouping UMIs and building a consensus sequence



Bias correction by counting UMIs

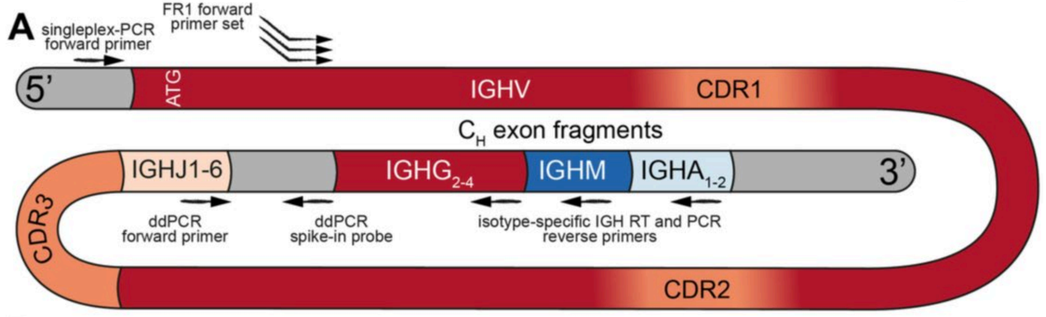
Clone	Original count	# Reads	# UMI	Clonal frequency based on # reads	Clonal frequency based on # UMI	True clonal frequency
X	3	8	3	0.533	0.6	0.6
Y	2	7	2	0.467	0.4	0.4

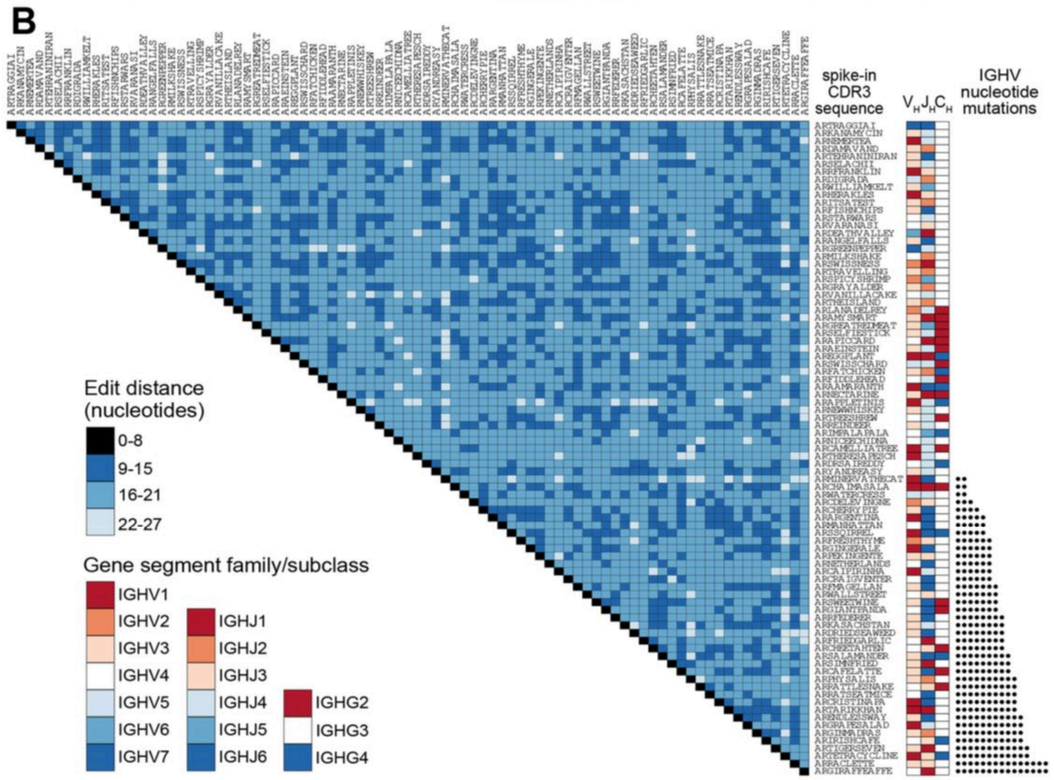
Fan, **PNAS**, 2011.
Kinde, **PNAS**, 2011.
Kivioja, **Nature Meth**, 2011.
Jabara, **PNAS**, 2011.
Shiroguchi, **PNAS**, 2012.
Lundberg, **Nature Meth**, 2013.

Khan, Science Advances, 2016

AIRR-seq error correction using UMI (mouse and human)

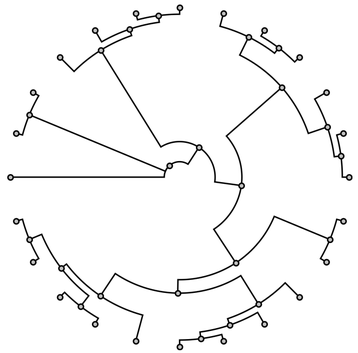
Validation of UMI error correction via spike-in design





100% clonal error correction across all 16 spike-ins

Clonal variants for spike-in clone: CRISTINAW

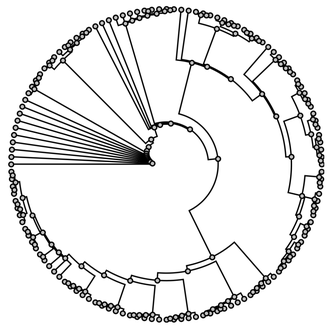


Uncorrected
24 clones

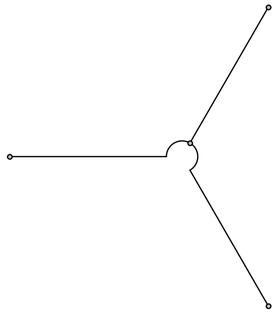
MAF-corrected
1 clone

98% intraclonal error correction across all 16 spike-ins

Intraclonal variants for spike-in clone: CRISTINAW



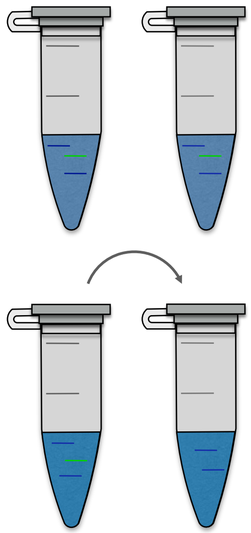
Uncorrected
178 intraclonal variants



MAF-corrected
3 intraclonal variants

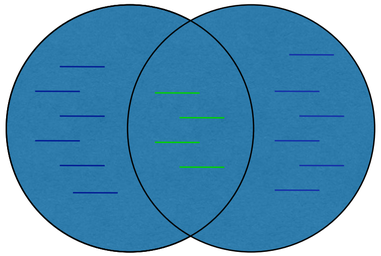
Khan, Science Advances, 2016
Friedensohn, Front Imm, 2018

Another benefit of UMIs: exclusion of sample contamination

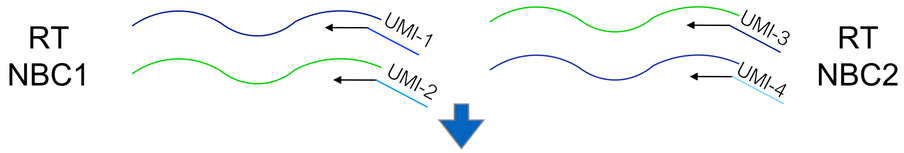


overlapping clones can be **biological**

overlapping clones can also result from **contamination**

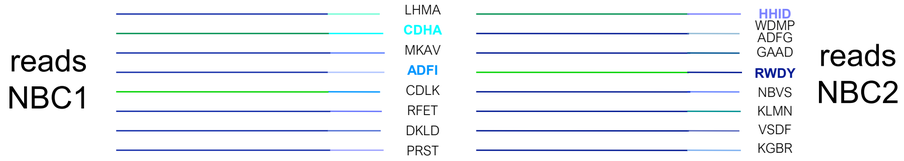


Unique tagging of each mRNA molecule in RT step with UMI



PCR library preparation and handling

Overlapping clones with different tags must be biological



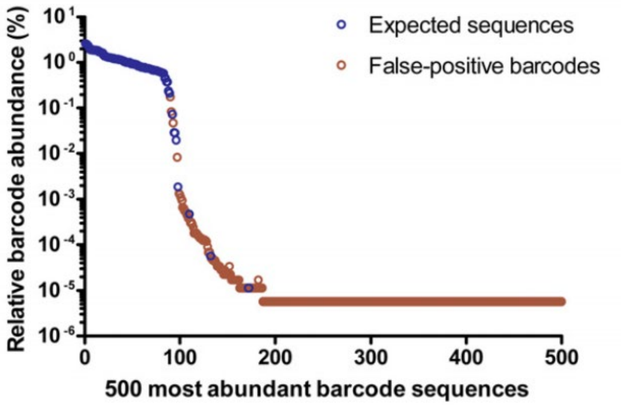
	CDR3 overlap (%)				UID overlap (%)				UID overlap of overlapping CDR3s (%)			
HBsAg7_nBC2	0.23	0.46	3.09		0.1	0.13	0.14		0	0	0	
HBsAg7_nBC1	0.62	0.25			0.14	0.11			0	0		
NP−HEL2_nBC2	7.15				0.12				0			
NP−HEL2_nBC1												
	NP−HEL2_nBC1	NP−HEL2_nBC2	HBsAg7_nBC1	HBsAg7_nBC2	NP−HEL2_nBC1	NP−HEL2_nBC2	HBsAg7_nBC1	HBsAg7_nBC2	NP−HEL2_nBC1	NP−HEL2_nBC2	HBsAg7_nBC1	HBsAg7_nBC2

No contamination

Greiff, Cell Reports, 2017.

I) Issues in UMI use: errors in UMIs

A library of 100 UMIs (barcodes) showed 7 false positives in top 100



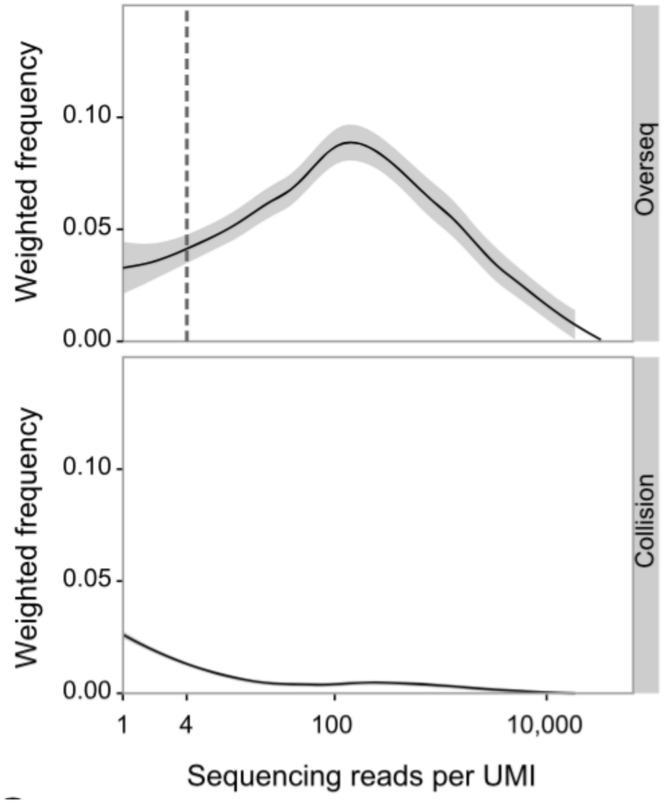
Deakin, Nucleic Acids Res, 2014.

Overestimation of diversity due to errors in UMIs (PID)

PID	No. of reads
AG.ATGGCCTG.	2178
AG.ATGGCCTA.	3
AG.ATGTCCTG.	5
AG.ATGGCCCG.	4
AG.GTGGCCTG.	7
AG.ATGCCCTG.	4
AG.ATGGC.TGA	15
AG.ATG.CCTGA	4
AGTATGGCCT..	3
CACT.G.CTATT	1865
CACC.G.CTATT	3
CACT.G.CTATC	14
CACT.G.CTACT	3
CACT.GTCTAT.	7
CACTAG.CTAT.	4

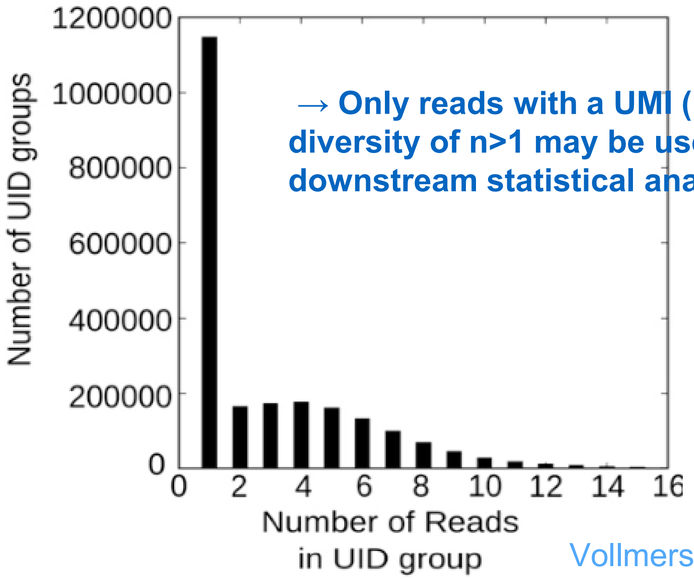
Brodin, PLoS ONE, 2015.

→ Setting a UMI threshold (n≈>3–4) to eliminate erroneous UMI diversity



Egorov, J Immunol, 2015.

II) Issues in UMI use: undersampling of UMIs



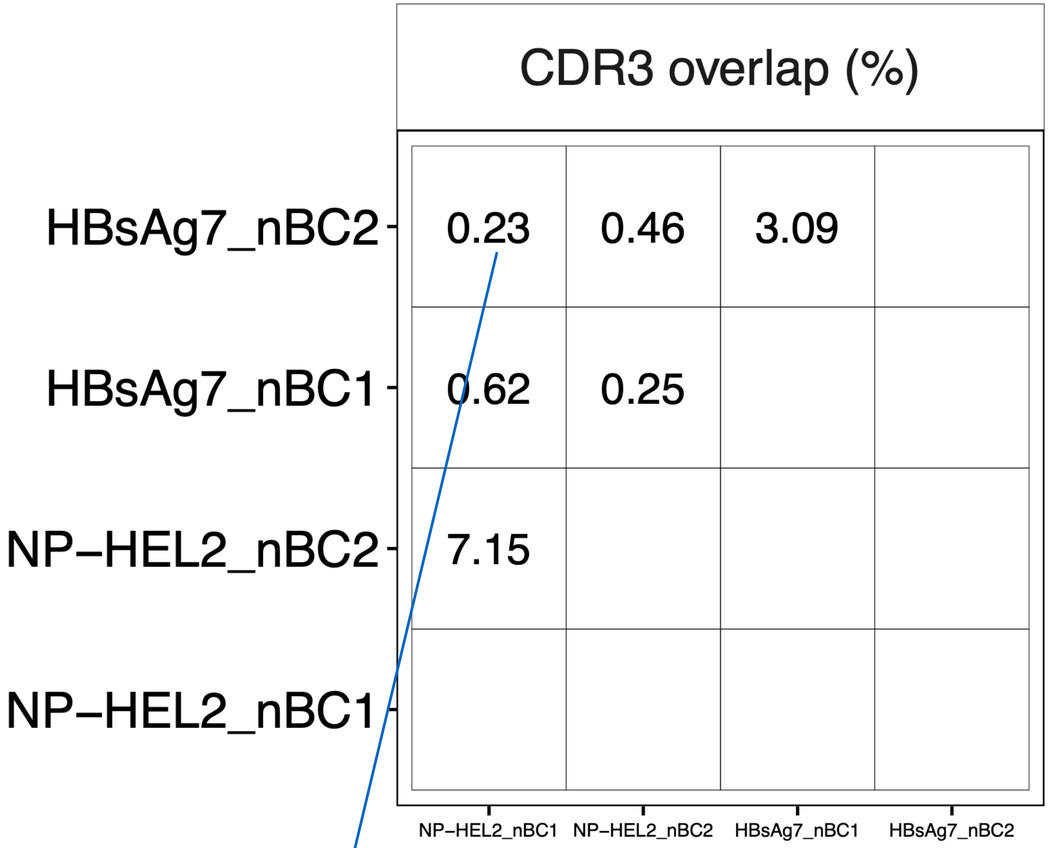
→ Only reads with a UMI (UID) diversity of n>1 may be used for downstream statistical analyses

Vollmers, PNAS, 2013

is not sufficient. For example, introducing a hard cutoff that discards all UMIs with fewer than five reads leads to a decrease in observed TCR diversity. UMI-based methods might be more accurate for assessing clonotype frequency, in line with their use to quantify and correct for PCR errors and bias[41]. Furthermore, a threshold of 2–4 reads per UMI efficiently protects against artifacts and cross-sample contamination[42], which become critical with tighter cluster density on modern Illumina machines. UMI-based methods might require several replicates or higher sequencing coverage to consistently and unambiguously identify rare TCR sequence clonotypes. Notably, both RACE-1 and RACE-2 methods performed better after UMI correction (see Table 1).

Barennes, NBT, 2020

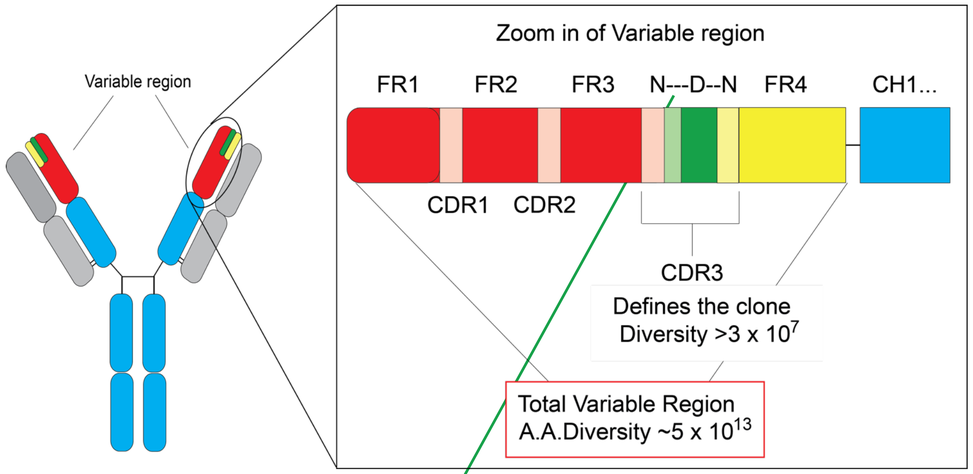
→ High-diversity repertoires (>500,000 unique clones) such as naïve B cells may not be sufficiently covered using UMI technology



Greiff, Cell Reports, 2017

→ Expected CDR3 overlap: 14% (Figure 5 in Greiff/Menzel et al. , Cell Reports, 2017).

High-throughput annotation of AIRR-seq data



Annotation of **D-region** is unreliable

Data	Platform	% of wrong V genes	% of wrong D genes	% of wrong J genes	% of wrong CDR3
Synthetic TRB					
	MiXCR	0.0	35.3	0.2	0.4
	IMGT	0.6	21.6	9.3	19.0
	Decombinator	3.8	N/A	2.3	N/A
	IgBlast	0.0	28.5	0.0	N/A
Synthetic IGH					
	MiXCR	0.0	27.8	0.2	1.6
	IMGT	1.3	54.4	11.6	14.1
	IgBlast	0.0	20.3	0.0	N/A

Bolotin, Nat Met, 2015

https://b-t.cr/t/list-of-v-d-j-annotation-software/18

Dedicated aligners

- BRILIA 24 (Lee et al. 2017 6)
- CloAnalyst 38 (Kepler 2013 9)
- Decombinator 17 (Thomas et al. 2013 3) : uses a finite state automaton
- iHMMune-align 16 (Gaeta et al. 2007 3) : Hidden Markov Model
- IgBLAST 14 (Ye et al 2013) : highly tuned BLAST
- IgSCUEAL 11 (Frost et al 2015 7) : phylogenetic placement
- IMSEQ 10 (Kuchenbecker et al 2015 1)
- Joinsolver 6 (Souto-Carneiro et al. 2004) : webserver only
- MiXCR 18 (Bolotin et al 2015 3)
- partis 10 , also Ighutil 8 (Ralph and Matsen 2016 5)
- reppertHMM 3 (Elhanati et al. 2015 2)
- SoDA (binary available in Automation) 8 (Volpe et al. 2006 8 ; see also Munshaw and Kepler 2010)
- VDJFasta 13 (Glanville et al. 2009 4)
- VDJsolver 7 (Ohm-Laursen et al. 2006 8)

Alignment wrappers and webservers

- Change-O 19 (Gupta et al. 2015 2)
- IMGT V-QUEST 5 (Lefranc et al 2008)
- ImmuneDB 12 (Rosenfeld et al. 2017 3) : implements alignment method described in (Zhang et al. 2015)
- SONAR 5 (Schramm et al. 2016)
- VBASE2 3 (Retter et al. 2004)

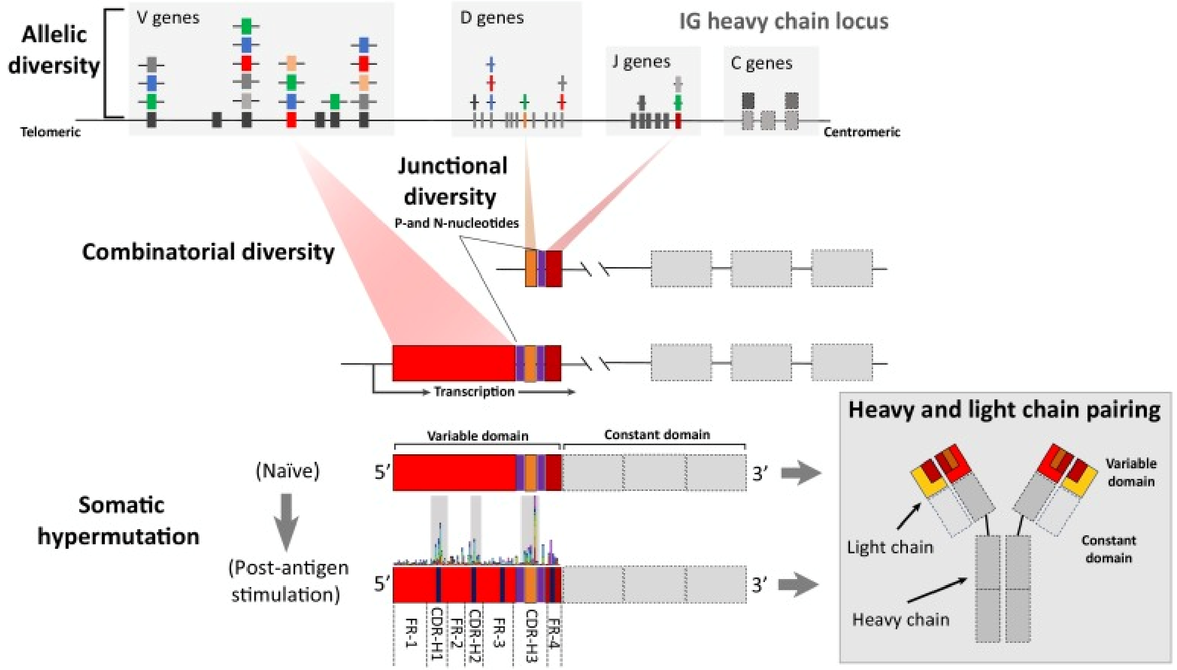
Without publications

- abstar 11 : Python; focus on scale-up
- MiGMAP 2 : wraps IgBLAST and includes extra features
- IgValve 5 : Ruby, for validation
- vdj 12 : Python; last update 2014

	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

Greiff, Trends Immunol, 2015

Genetic source of repertoire differences: germline gene loci



Watson, Trend Imm, 2017

Collins, Curr Op Sys Bio, 2020

Implication in disease

Open Access | Published: 16 February 2016

IGHV1-69 polymorphism modulates anti-influenza antibody repertoires, correlates with IGHV utilization shifts and varies by ethnicity

Yuval Avnir, Corey T. Watson, Jacob Glanville, Eric C. Peterson, Aimee S. Tallarico, Andrew S. Bennett, Kun Qin, Ying Fu, Chiung-Yu Huang, John H. Beigel, Felix Breden, Quan Zhu & Wayne A. Marasco

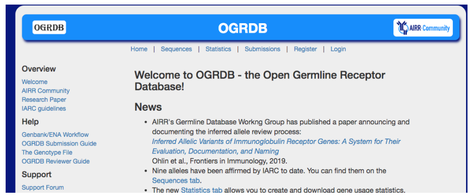
Scientific Reports **6**, Article number: 20842 (2016) | Cite this article

Brief Definitive Report | February 17 2014

Epitope-specific antibody response is controlled by immunoglobulin V_H polymorphisms

Bruno Raposo, Doreen Dobritzsch, Changrong Ge, Diana Ekman, Bingze Xu, Ingrid Lindh, Michael Förster, Hüseyin Uysal, Kutty Selva Nandakumar, Gunter Schneider, Rikard Holmdahl

Allele databases



Inferred Allele Review Committee (IARC)

Purpose:

The Inferred Allele Review Committee was formed after the third AIRR-Community meeting in Rockville, MD in December 2017. **The IARC is responsible for judging the validity of germline immunoglobulin and TCR genes, inferred from RepSeq data. It will advise IMGT and the IUIS/IMGT nomenclature committee of their findings. It will also work with IMGT to make inferred sequences, and evidence in support of their existence, available to the AIRR community and other researchers.** The work of the committee will initially focus on human

VDJbase: an adaptive immune receptor genotype and haplotype database

Aviv Omer, Or Shemesh, Ayelet Peres, Pazit Polak, Adrian J Shepherd, Corey T Watson, Scott D Boyd, Andrew M Collins, William Lees, Gur Yaari

Author Notes

Experimental and computation allele/haplotype detection

- Germline gene alleles might differ by ethnicity.
- Ideally, germline gene reference databases for antibody sequence annotation should be compiled for each individual (Corcoran, Nat Comm, 2015, Gadala-Maria, PNAS, 2015, Ralph, PLoSCompBio, 2019, Peres 2019 Bioinformatics, Gidoni 2019 Nat Comms, Omer 2020 Bioinformatics, Rodriguez, Frontiers in Imm 2020)

Summary: Error correction and annotation of AIRR-seq data

- Biologically conclusive AIRR-seq depends on deep coverage of immune repertoires. Coverage may be assessed via replicates and species accumulation curves
- AIRR-seq library preparation can introduce numerous errors: primer bias, PCR bias
- Error correction can be performed both experimentally (e.g., UMI, replicates) and computationally (e.g., clonotype clustering, exclusion of singleton reads)
- UMI-based error correction may not not applicable for highly diverse sample
- Numerous AIRR-seq sequence annotation tools exist. However, care should be taken when choosing the reference genome in order to avoid introducing artificial diversity or mutations
- Identification of germline gene alleles enables more accurate germline gene annotation and SHM quantification | potential link between germline genes and antibody-antigen binding/disease

Outline

Introduction to Adaptive immune receptor repertoire sequencing (AIRR-seq)

- Generation of immune repertoire diversity
- Workflow and applications of AIRR-seq

Error correction and Standardization of AIRR-seq data

- Experimental design and considerations
- Error and bias correction
- Standardization

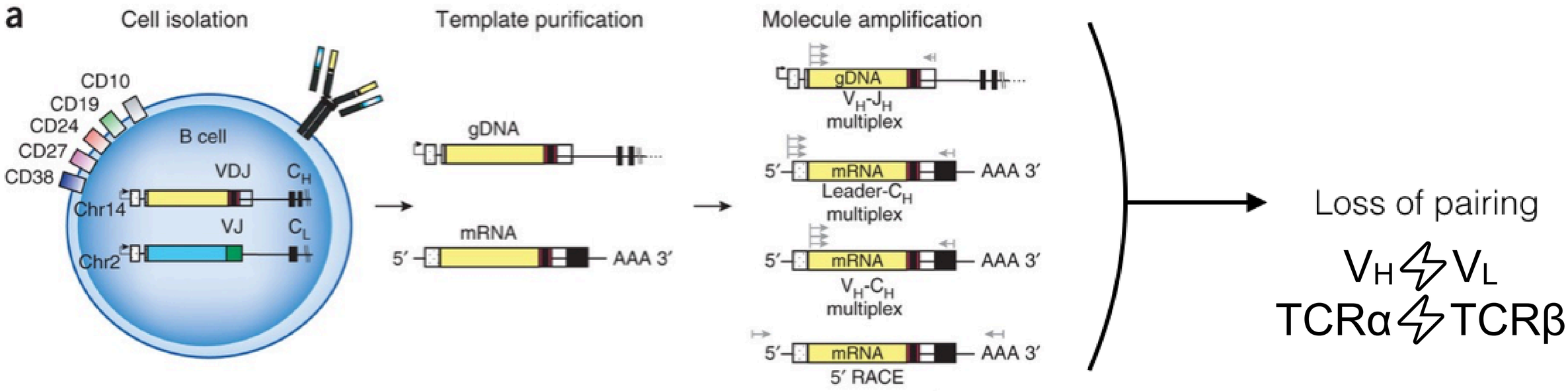
Single-cell AIRR-seq

- Pairing by targeted amplification
- Single-cell sequencing

Computational strategies for immune repertoire analysis

- Diversity and convergence analysis
- Network analysis
- Machine learning

Issues in bulk AIRR-sequencing



Input material:

- DNA
- RNA

Methods:

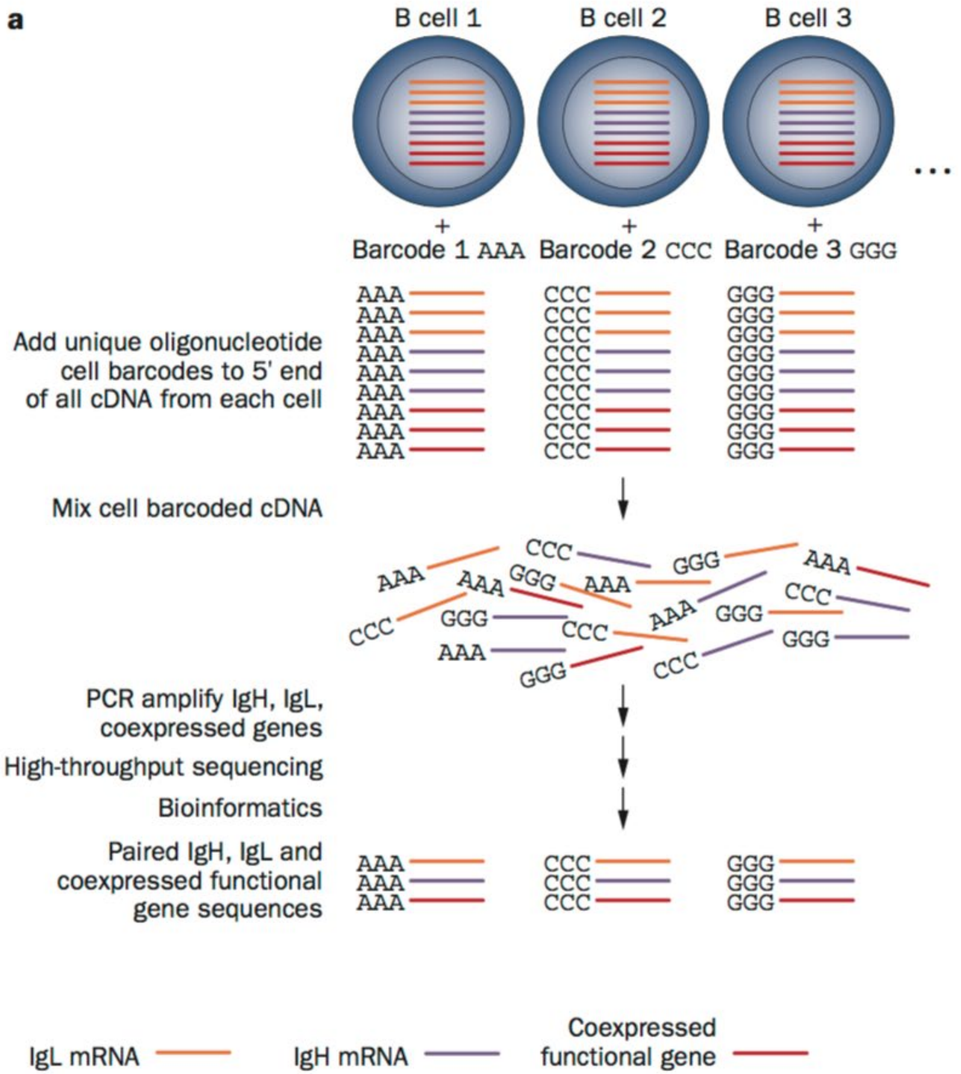
- Multiplex PCR
- RACE

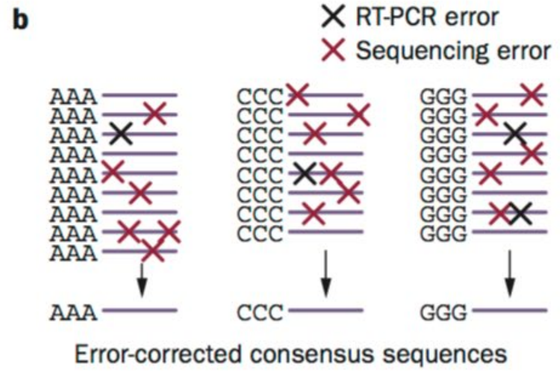
Georgiou, Nature Biotech, 2014.

Benefits of single-cell sequencing: gene expr. info and error corr.

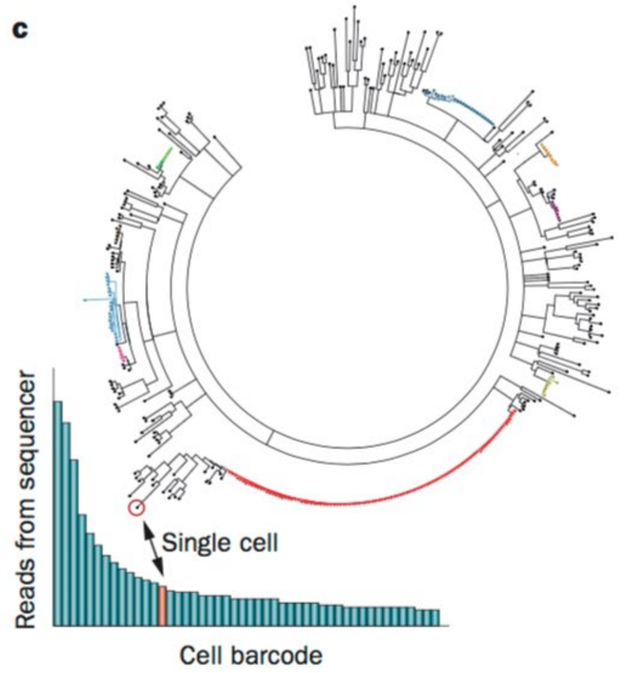
Benefit 1:
Simultaneously measure TCR, B cell Ig, and gene expression in the same sample

Benefit 2:
preserve IgH–IGL, TCRa–TCRb pairing





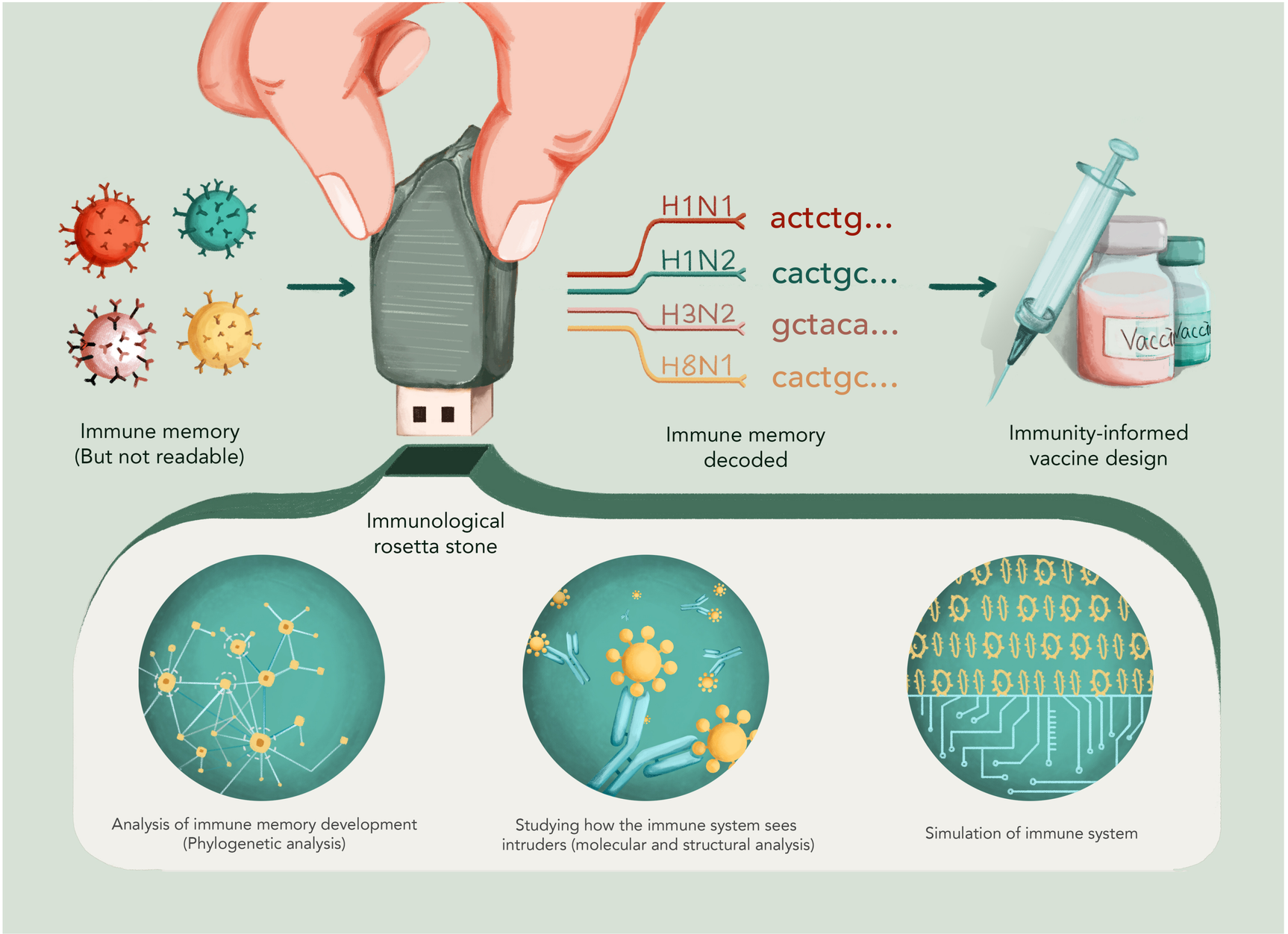
Benefit 3:
Reduce error by cell barcoding



Benefit 4:
Obtain correct clonal frequency ranking

Stubbington, Science, 2017 Robinson, Nat Rev Rheu, 2014

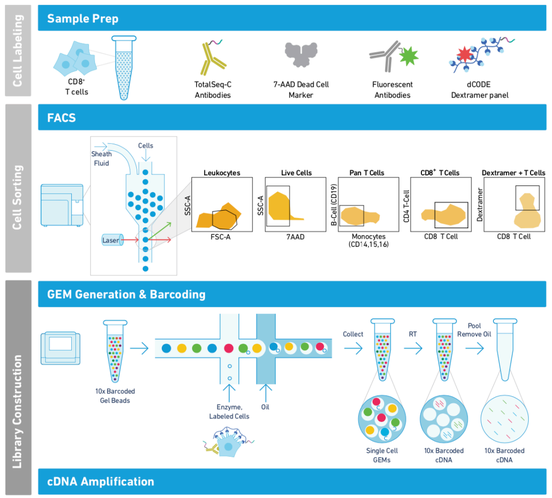
A Rosetta Stone for immunology is needed to map immune receptors to antigen recognition



https://www.vaksinebloggen.no/vaccine-research-in-need-of-a-rosetta-stone/

Novel technologies for linking immune receptor sequence to function

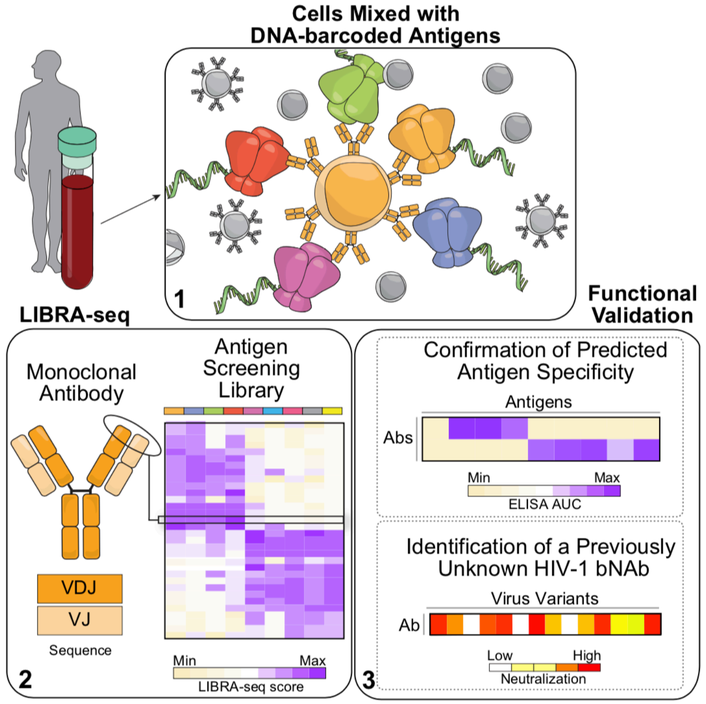
Assess the binding specificities of over 150,000 CD8⁺ T cells from 4 human donors across a highly multiplexed panel of 44 distinct, specific peptide–MHC (pMHC) multimers



High-Throughput Mapping of B Cell Receptor Sequences to Antigen Specificity

Cell

Ian Setliff,[1,2,16] Andrea R. Shiakolas,[1,3,16] Kelsey A. Pilewski,[1,3] Amyn A. Murji,[1,3] Rutendo E. Mapengo,[4] Katarzyna Janowska,[5] Simone Richardson,[4,11] Charissa Oosthuysen,[4,11] Nagarajan Raju,[1,3] Larance Ronsard,[7] Masaru Kanekiyo,[8] Juliana S. Qin,[1] Kevin J. Kramer,[1,3] Allison R. Greenplate,[1] Wyatt J. McDonnell,[3,9,17] Barney S. Graham,[8] Mark Connors,[10] Daniel Lingwood,[7] Priyamvada Acharya,[5,6] Lynn Morris,[4,11,12]



An integrated immune discovery solution

- Takes days compared to months
- Highly customizable

BEAM-T | Coming H1 2022
Massive Scale T Cell Receptor Discovery

BEAM-Ab | Coming H1 2022
Massive Scale Antibody Discovery

Letter | Published: 30 March 2020

High-throughput single-cell activity-based screening and sequencing of antibodies using droplet microfluidics

Annabelle Gérard, Adam Woolfe, [...] Colin Brenan ✉

Nature Biotechnology (2020) | Cite this article

Determinants governing T cell receptor α/β-chain pairing in repertoire formation of identical twins

Hidetaka Tanno[a,b,1], Timothy M. Gould[c,d,1], Jonathan R. McDaniel[a], Wenqiang Cao[c,d], Yuri Tanno[a], Russell E. Durrett[a], Daechan Park[e], Steven J. Cate[f], William H. Hildebrand[f], Cornelia L. Dekker[g], Lu Tian[h], Cornelia M. Weyand[c,d], George Georgiou[a,b,2,3], and Jörg J. Goronzy[c,d,2,3]

[a]Department of Chemical Engineering, University of Texas at Austin, Austin, TX 78712; [b]Institute for Cellular and Molecular Biology, University of Texas at Austin, Austin, TX 78712; [c]Division of Immunology and Rheumatology, Department of Medicine, Stanford University School of Medicine, Stanford, CA 94305; [d]Department of Medicine, Palo Alto Veterans Administration Healthcare System, Palo Alto, CA 94304; [e]Department of Life Sciences, Ajou University, Suwon 16499, South Korea; [f]Department of Microbiology and Immunology, University of Oklahoma Health Sciences Center, Oklahoma City, OK 73104; [g]Department of Pediatrics (Infectious Diseases), Stanford University School of Medicine, Stanford, CA 94305; and [h]Department of Biomedical Data Science, Stanford University School of Medicine, Stanford, CA 94305

Article | Published: 16 March 2020

Massively parallel interrogation and mining of natively paired human TCRαβ repertoires

Matthew J. Spindler, Ayla L. Nelson, Ellen K. Wagner, Natasha Oppermans, John S. Bridgeman, James M. Heather, Adam S. Adler, Michael A. Asensio, Robert C. Edgar, Yoong Wearn Lim, Everett H. Meyer, Robert E. Hawkins, Mark Cobbold & David S. Johnson ✉

Nature Biotechnology **38**, 609–619(2020) | Cite this article

Open software for analysing AIRR single-cell data

Immcantation Tutorials » 10x Genomics V(D)J Sequence Analysis Tutorial

Edit on Bitbucket

10x Genomics V(D)J Sequence Analysis Tutorial

Overview

This tutorial is a basic walkthrough for defining B cell clonal families and building B cell lineage trees using 10x Genomics BCR sequencing data. **It is intended for users without prior experience with Immcantation.** If you are familiar with Immcantation, then this page may be more useful.

Knowledge of basic command line usage is assumed. Please check out the individual documentation sites for the functions detailed in this tutorial before using them on your own data. For simplicity, this tutorial will use the Immcantation Docker image which contains all necessary software. It is also possible to install the packages being used separately (see pRESTO, Change-O, and Alakazam).

SOFTWARE TOOL ARTICLE

REVISED scRepertoire: An R-based toolkit for single-cell immune receptor analysis [version 2; peer review: 2 approved]

Nicholas Borcherding [1-4], Nicholas L. Bormann[5], Gloria Kraus[6]

Author details

`immunarch` — Fast and Seamless Exploration of Single-cell and Bulk T-cell/Antibody Immune Repertoires in R

Scirpy: a Scanpy extension for analyzing single-cell T-cell receptor-sequencing data

Gregor Sturm, Tamas Szabo, Georgios Fotakis, Marlene Haider, Dietmar Rieder, Zlatko Trajanoski, Francesca Finotello

Bioinformatics, Volume 36, Issue 18, 15 September 2020, Pages 4817–4818,
https://doi.org/10.1093/bioinformatics/btaa611
Published: 02 July 2020 **Article history** ▾

Why `immunarch`?

- **Work with any type of data:** single-cell, bulk, data tables, databases — you name it.
- **Community at the heart:** ask questions, share knowledge and thrive in the community of almost 30,000 researchers and medical scientists worldwide. **Pfizer, Novartis, Regeneron, Stanford, UCSF** and **MIT** trust us.
- **One plot — one line:** write a whole PhD thesis in 8 lines of code or reproduce almost any publication in 5-10 lines of `immunarch` code.
- **Be on the bleeding edge of science:** we regularly update `immunarch` with the latest methods. Let us know what you need!
- **Automatic format detection and parsing** for all popular immunosequencing formats: from **MiXCR** and **ImmunoSEQ** to **10XGenomics** and **ArcherDX**.

Caveats single-cell analysis

- Reduced throughput as compared to bulk seq

- Benchmarking of technology still in its infancy (keep in mind that even bulk-sequencing is still not fully standardized and mostly incomparable across sequencing protocols and technologies)

- Data analysis pipelines (downstream of data processing) are mostly developed for bulk-sequencing and cannot be readily transferred to single-cell seq (how to treat paired information in data analysis remains unclear)

Summary: Single-cell AIRR-seq

- Bulk and single-cell (b/sc)AIRR-seq allow asking different research questions
- bAIRR-seq remains the state-of-the-art in case deep coverage of immune repertoire diversity is the main research focus
- scAIRR-seq preserves pairing information and is therefore superior if exact clonal/pairing information is needed such as for: phylogenetics, and antibody/TCR engineering
- scAIRRseq allows stricter error correction and coupling of transcriptome and repertoire analysis
- Recently several scAIRR-seq approaches and (commercial) platforms have emerged. However, given their relative niche presence, they have not been sufficiently compared and validated by a wider community (e.g., spike-in controls)

Outline

Introduction to Adaptive immune receptor repertoire sequencing (AIRR-seq)

- Generation of immune repertoire diversity
- Workflow and applications of AIRR-seq

Error correction and Standardization of AIRR-seq data

- Experimental design and considerations
- Error and bias correction
- Standardization

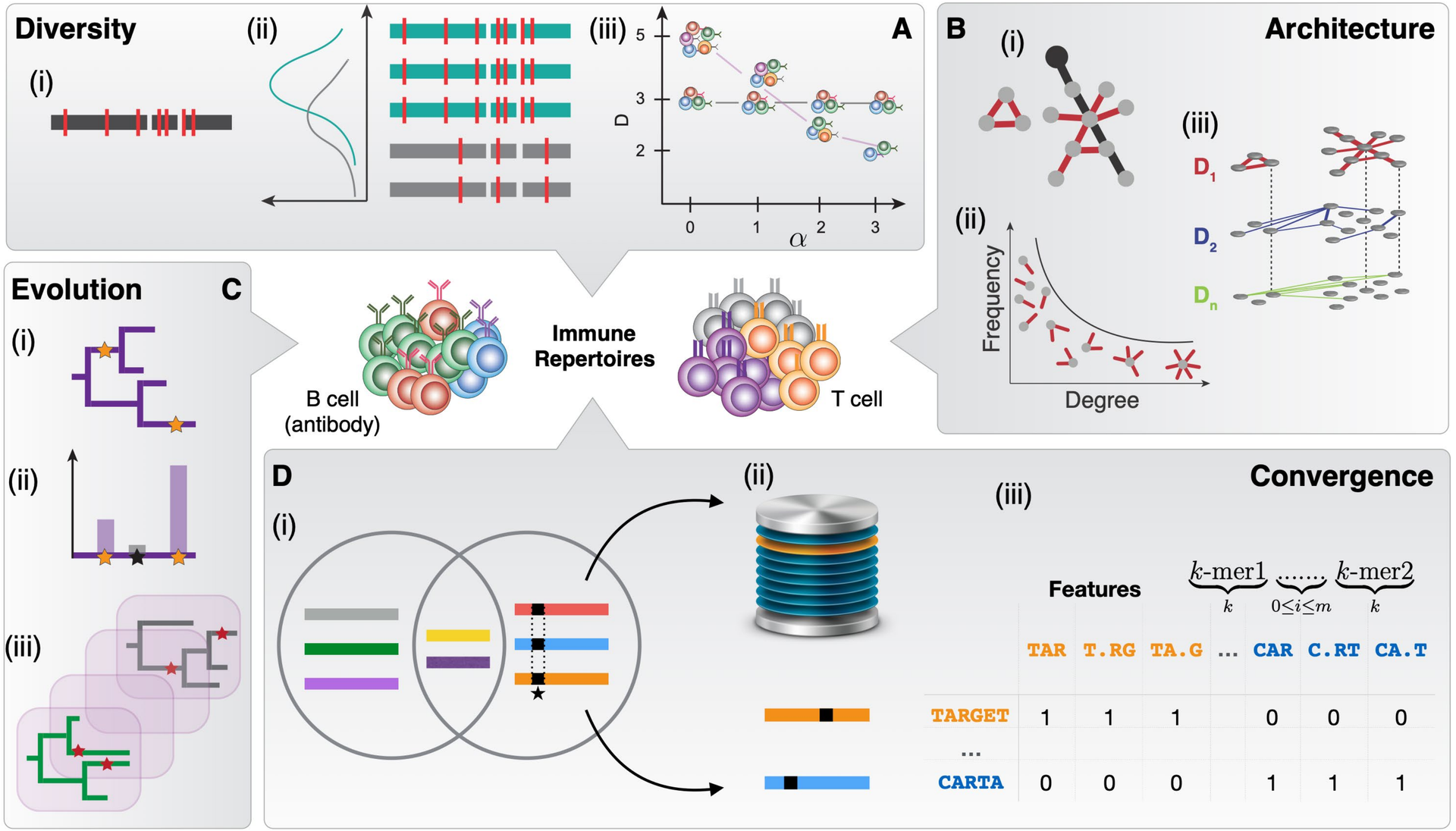
Single-cell AIRR-seq

- Pairing by targeted amplification
- Single-cell sequencing

Computational strategies for immune repertoire analysis

- Diversity and convergence analysis
- Network analysis
- Machine learning

Computational strategies for dissecting the high-dimensional complexity of adaptive immune repertoires



Structural diversity of B-cell receptor repertoires along the B-cell differentiation axis in humans and mice

Aleksandr Kovaltsuk, Matthew I. J. Raybould, Wing Ki Wong, Claire Marks, Sebastian Kelm, James Snowden, Johannes Trück, Charlotte M. Deane

Miho, Front Imm, 2018

→Given the similarity of antibody and T cell receptor genomic structure, **most computational analyses can be applied interchangeably**

→ For an **in depth overview of current computational strategies and future directions for immune repertoire analysis**, please refer to

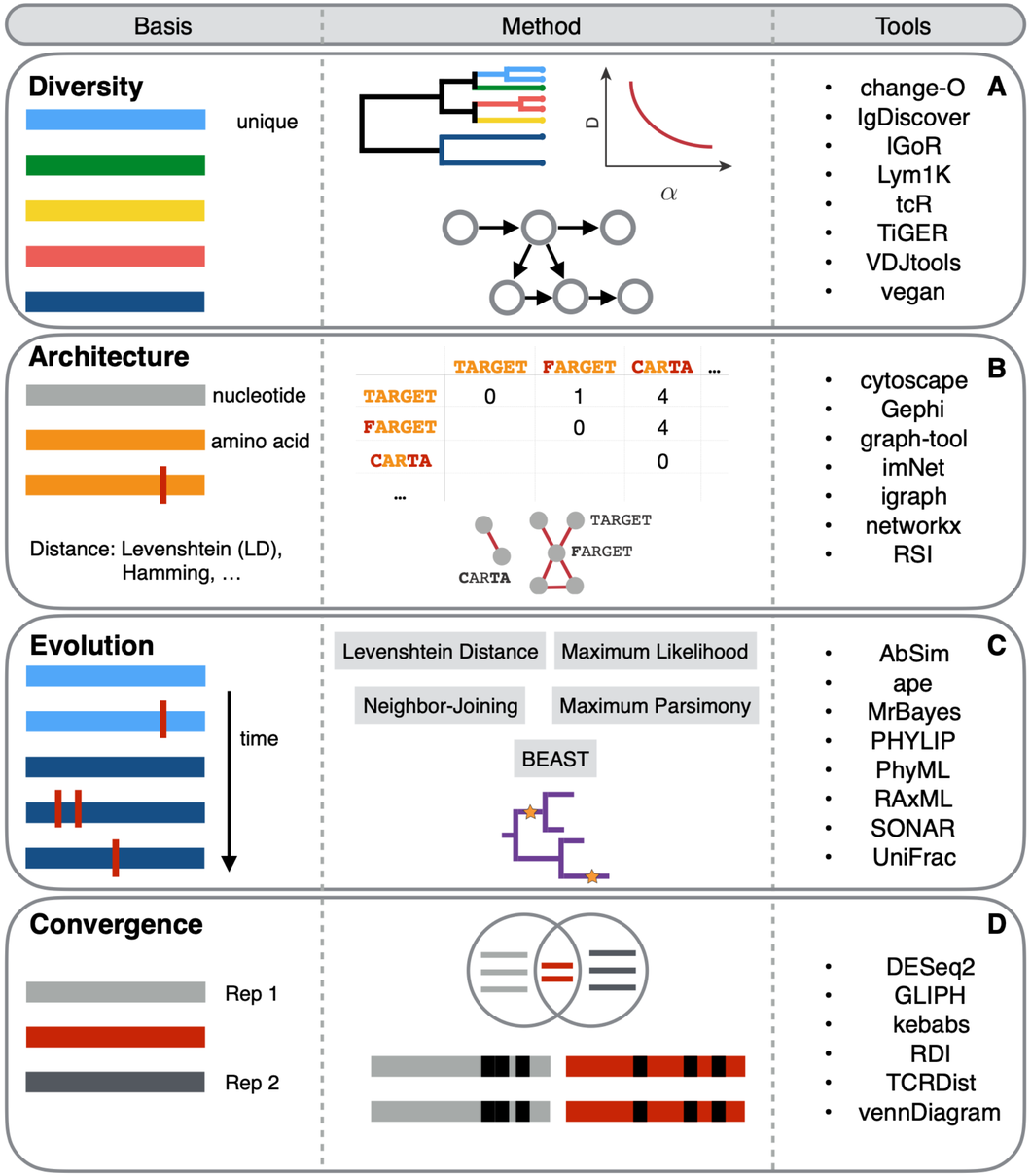
- Rosati, BMCBiotech, 2017
- Miho, Front Imm, 2018
- López-Santibáñez-Jácome, PeerJ, 2018
- Brown, MSDE, 2019
- Bradley, AnnRevImm, 2019
- Lees, Curr Op in SysBio2020

Exemplary list of comp. tools for immune repertoire analysis

Most tools are written in **python** and **R** (with C/Java being used to improve performance of certain subroutines)

→ For an **in depth overview of current computational strategies and future directions for immune repertoire analysis**, please refer to

- Rosati, BMCBiotech, 2017
- Miho, Front Imm, 2018
- López-Santibáñez-Jácome, PeerJ, 2018
- Brown, MSDE, 2019
- Bradley, AnnRevImm, 2019
- Lees, Curr Op in SysBio2020



→ Diversity tools can be subdivided into 3 groups: (i) inference of germline gene diversity, (ii) inference of VDJ recombination statistics, (iii) quantification of clonal diversity

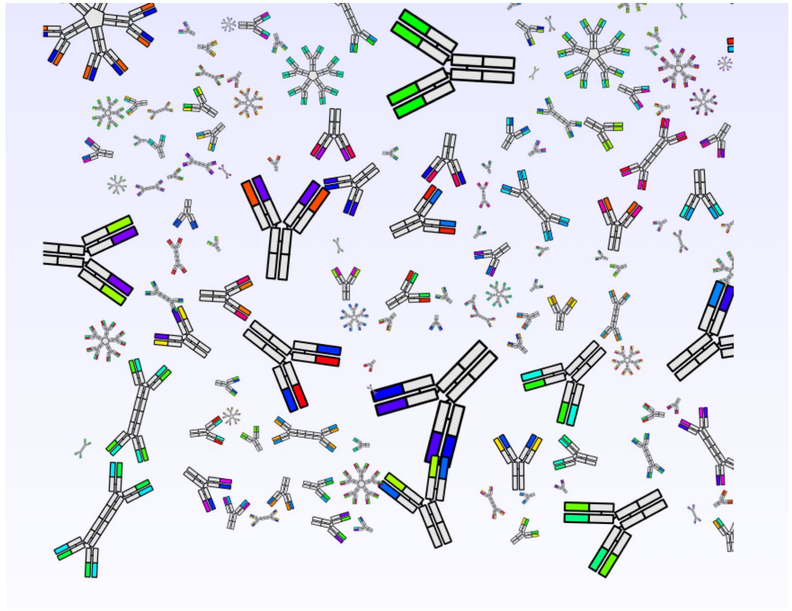
→ While igraph and networkx are predominantly used for *quantification* of network measures, cytoscape and gephi's main purpose is to *visualize* networks

→ Phylogenetic methods are being used exclusively for antibody data since SHM is absent from T cells.

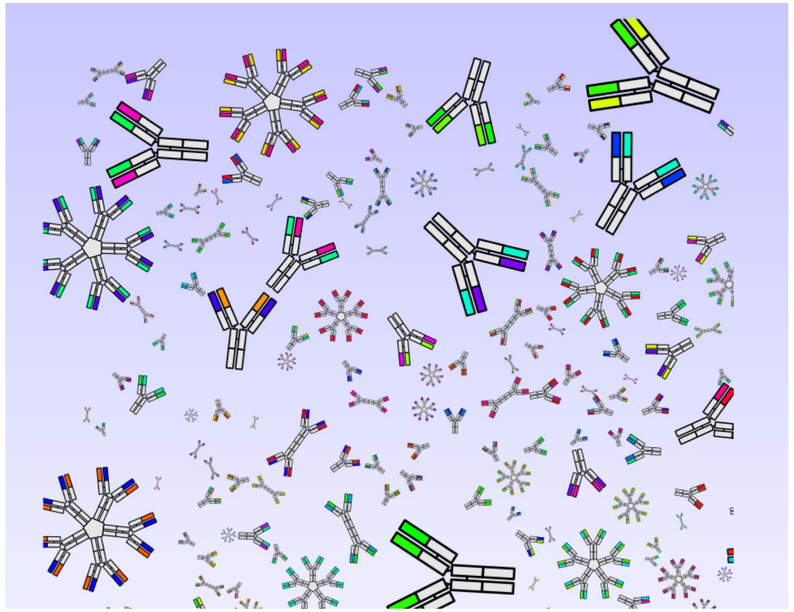
→ Repertoire convergence (overlap) can be quantified (i) using overlap, (ii) distance and (iii) machine learning methods

Miho, Front Imm, 2018

Quantifying and comparing the **diversity** of immune repertoires



Repertoire 1



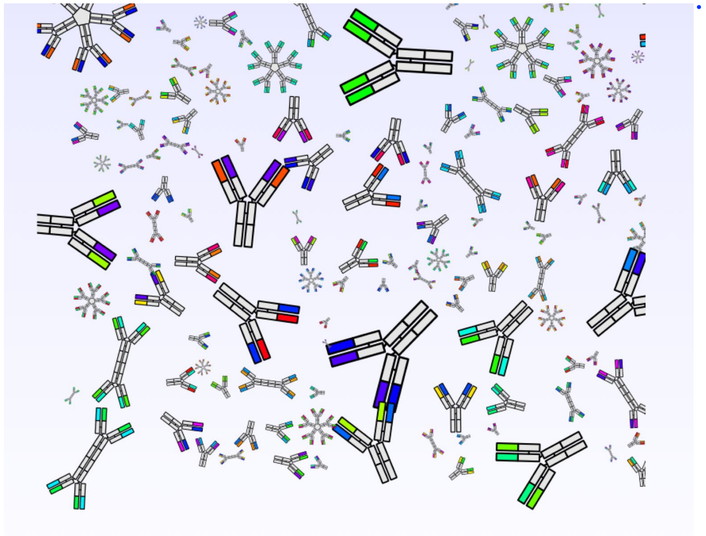
Repertoire 2

How are immune receptors distributed within a repertoire?
-uniform or uneven?

How does one compare those distributions across repertoires?

Quantifying repertoire diversity using diversity indices I

Repertoire 1

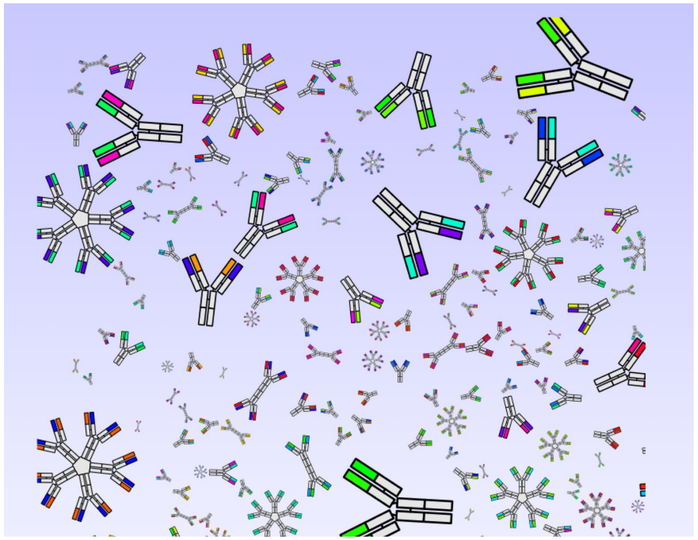


Antibody sequence	Antibody frequency [%]
CARTRGDYW	3.0
CARARHAYDYW	1.3
CARNYYGLADYW	0.9
CARGFADSDYW	0.7
...	...

Antibody (clonal) frequency distribution:

$$\vec{f} = (3.0, 1.3, 0.9, 0.7, \ldots)^T$$

Repertoire 2



Antibody sequence	Antibody frequency [%]
CARGHJADYW	10
CARYARHADY	4.3
CARGLANYYDY	2.7
CARDSGFADY	0.6
...	...

Antibody frequency distribution:

$$\vec{f} = (10, 4.3, 2.7, 0.6)^T$$

Repertoires **are not comparable** based on frequency distributions because Ab sequences do not overlap

Diversity indices solve this problem by mapping frequency distributions to a common coordinate system

Quantifying repertoire diversity using diversity indices II

Rényi entropy (new coordinate system)

$$H_\alpha = \frac{1}{1-\alpha}\log\left(\sum_i f_i^\alpha\right)$$

The higher alpha, **the higher is the weight of abundant sequences**

Antibody frequency distribution:

$$\vec{f} = (3.0, 1.3, 0.9, 0.7, \ldots)^T$$

$\alpha = 0$ → log (Species richness) := number of unique immune receptors

$\alpha = 1$ → $-\sum_i f_i \log f_i$ Shannon entropy

$\alpha = 2$ → $1 - \sum_i f_i^2$ log(Simpson's index)

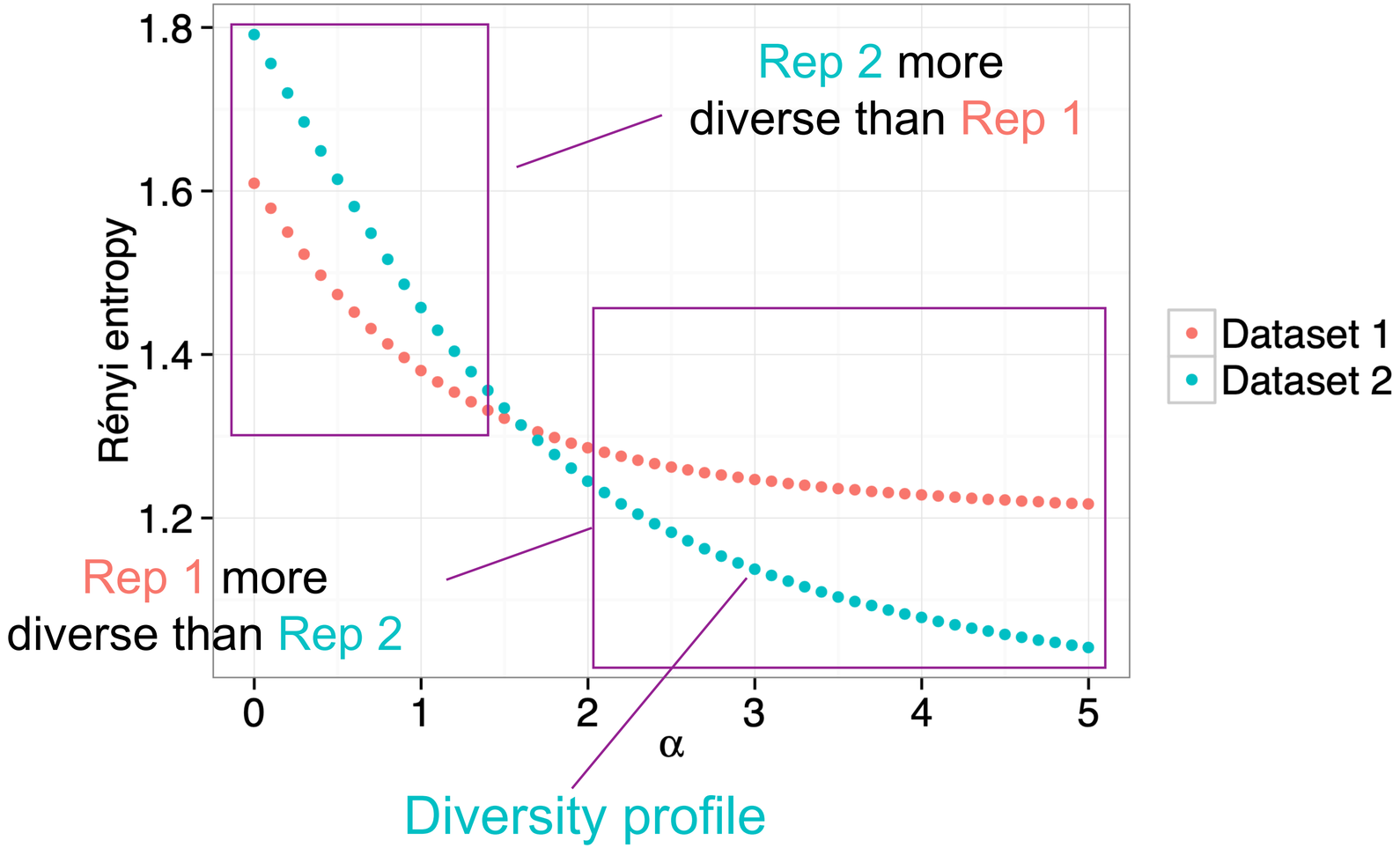
$\alpha \to \infty$ → $\max(\vec{f})$ log(Berger-Parker index)

Challenges in repertoire diversity analysis |

Repertoire 1 (33,29,28,5,5)%

Repertoire 2 (42,30,10,8,5,5)%

Frequency distributions



Challenge 1:

Rényi entropy is difficult to interpret biologically

Challenge 2:

Single diversity indices are insufficient to capture the sequence frequency space (qualitatively different results for different indices)

Challenge 1: Biological interpretation of diversity

$$H_\alpha = \frac{1}{1-\alpha}\log\left(\sum_i f_i^\alpha\right)$$

$$\exp\left(H_\alpha\right)$$

$${}^{\alpha}D = \left(\sum_{i=1}^{\mathrm{N}} f^\alpha\right)^{(1/1-\alpha)}$$

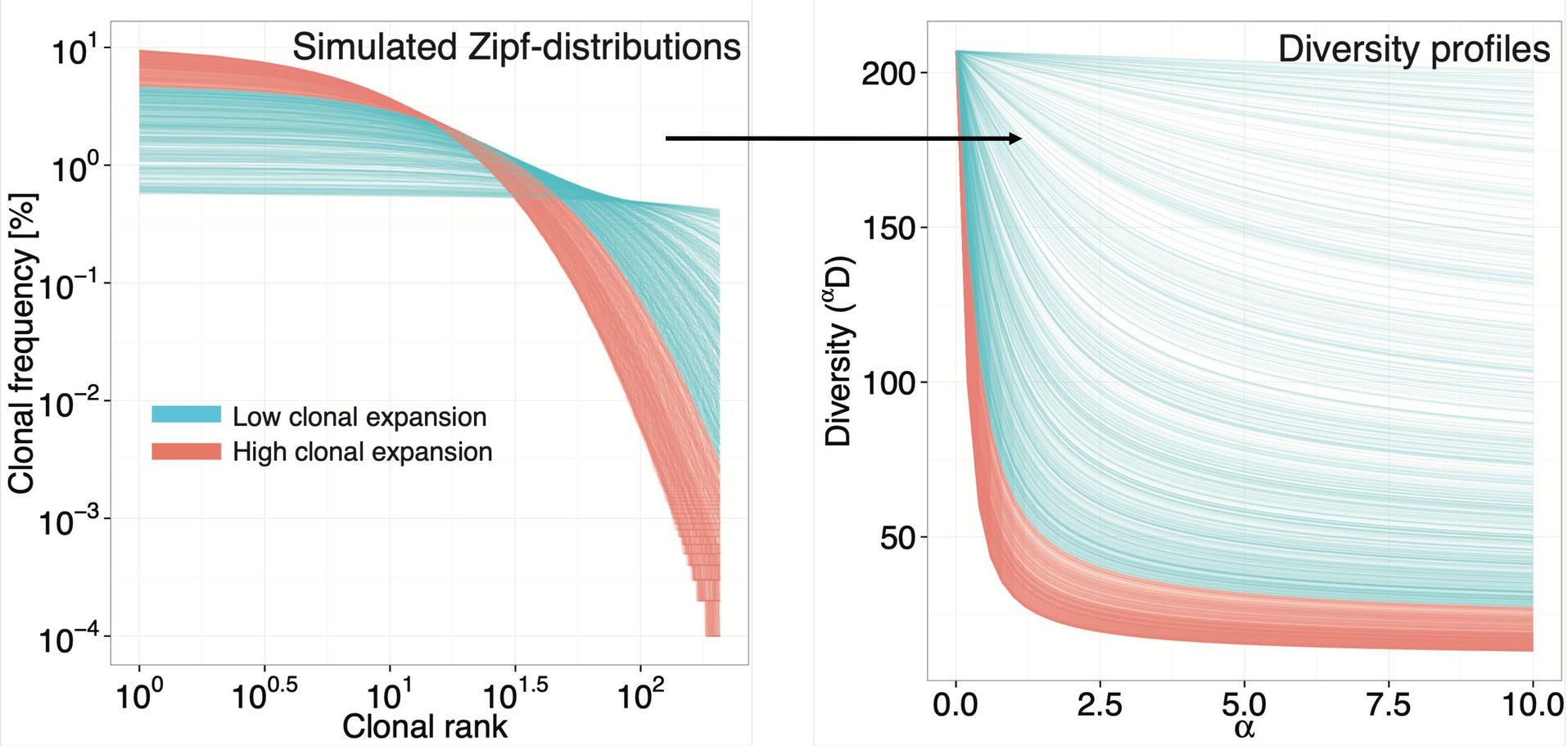
Example: Repertoire X is composed of 5 antibody sequences with a given frequency distribution (75,15,5,4,1)

$\alpha = {}^{1}D$(Repertoire X) = 2.28

Interpretation: The diversity of repertoire X **is equivalent** to a repertoire composed of ≈2 clones with equal frequency (50,50)

Hill-diversity (also termed: True diversity, effective number of species)

Challenge 2: capturing the entire frequency space using diversity profiles I



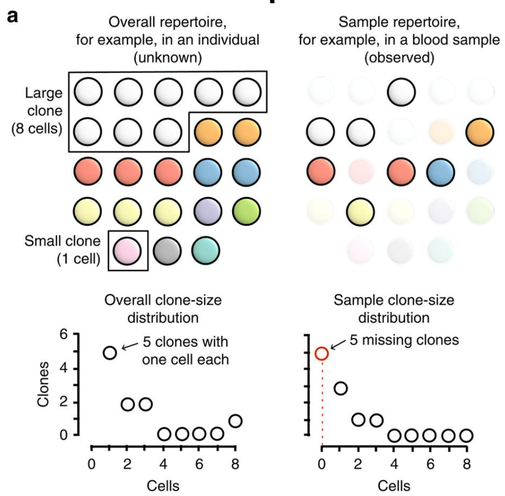
Zipf's law
(power law distribution)

$$g(\pi) := \begin{cases} \mathrm{C} \times \pi^{-\text{Zipf-}\alpha-1}, & 0 \leq \pi \leq \text{Zipf-B} \\ 0, & \text{otherwise.} \end{cases}$$

Greiff, Genome Medicine, 2015

Diversity estimations

Individual repertoire estimation from sample



Diversity measure	Overall	Sample	Ratio
Species richness	10.0	5.0	2.0x
Exp(entropy)	7.4	4.5	1.7x
Inverse Simpson index	5.6	4.2	1.3x
Inverse Berger-Parker index	2.9	2.7	1.1x

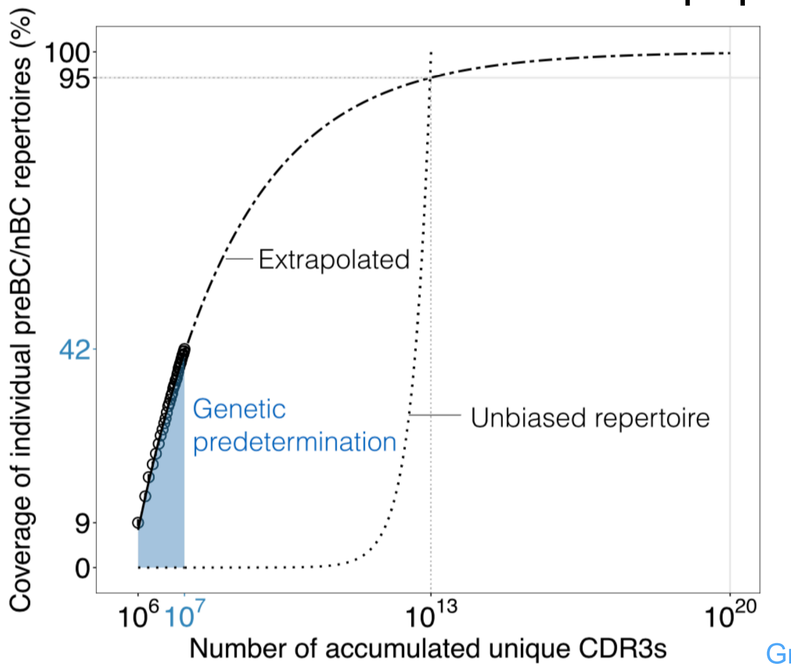
Arnaoult, Nat Comm, 2016

(Dis)Advantages of diversity estimators

estimator	advantages	disadvantages
parametric (e.g. Poisson abundance models, Power laws)	can estimate clonotype frequency distribution	requires *a priori* assumptions on analytical form of clonotype frequency distribution lack of validation: goodness-of-fit to observed data does not confirm model accuracy
non-parametric abundance-based estimators (e.g. Chao1, ACE, capture–recapture)	no *a priori* assumptions required on analytical form of clonotype frequency distribution	cannot estimate clonotype frequency distribution biased by sample size inaccurate in highly diverse immunological populations
non-parametric incidence-based estimators (e.g. Chao2, ICE)	does not require absolute count data	lack of validation in immunological populations biased by sample size
DivE	accurate in multiple validations, across all immunological populations tested unbiased by sample size	time consuming: multiple models must be fitted

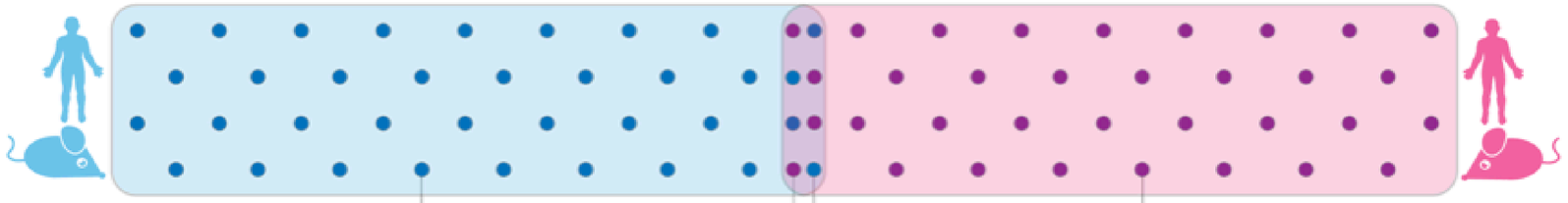
Laydon, Proc T. Soc B, 2015

Species accumulation curve to estimate population diversity



Greiff, Cell Reports, 2017

Quantifying sequence convergence based on entire sequences



Shared immune receptor sequences (public clones) across individuals

Castro, Dev & Comp Immunol, 2017

Shared subsequences (substrings) across individuals

Glanville, PNAS, 2011 (Figure 4)
Shugay, PLoS Comp Biol, 2015

Sequence-based (no frequency)

$$\text{overlap} = \frac{A \cap B}{\min(|A|, |B|)} \times 100$$

More weight to smaller repertoire

$$\text{overlap} = \frac{A \cap B}{\max(|A|, |B|)} \times 100$$

More weight to larger repertoire

$$\text{overlap} = \frac{A \cap B}{\text{mean}(|A|, |B|)} \times 100$$

Both repertoires (A, B) are weighted equally

Morisita-Horn (MH) index
Sequence/frequency-based

$$MH = \frac{2\sum_{i=1}^{S} a_i b_i}{\left(\frac{\sum_{i=1}^{S} a_i^2}{|A|^2} + \frac{\sum_{i=1}^{S} b_i^2}{B^2}\right)|A||B|}$$

Overlapping sequences are weighted by their frequency

S= #unique sequences

$$PG = \frac{\sum_{i=1}^{S} a_i^{\alpha} b_i^{\beta}}{\sum_{i=1}^{S} a_i^{2\alpha} + \sum_{i=1}^{S} b_i^{2\beta}}$$

Generalization of MH-index

$$\pi_{i,t} = \frac{\sum_{j=1}^{C_{i,t}-1} \sum_{k=j+1}^{C_{i,t}} d_{i,t,j} \cdot d_{i,t,k} \cdot G(s_{i,t,j}, s_{i,t,k})}{\binom{U_{i,t}}{2}}$$

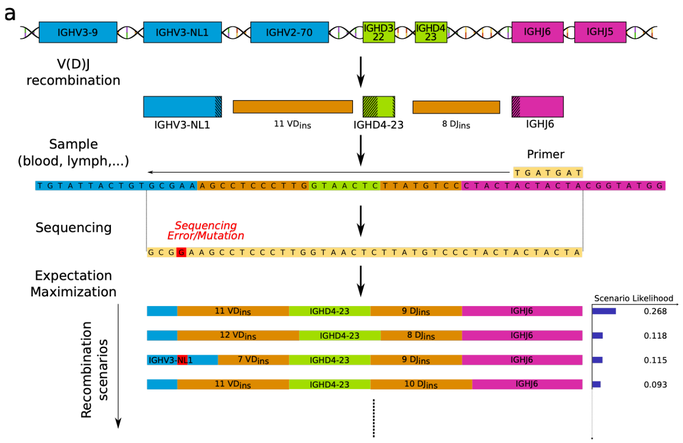
Repertoire similarity weighted by sequence frequency

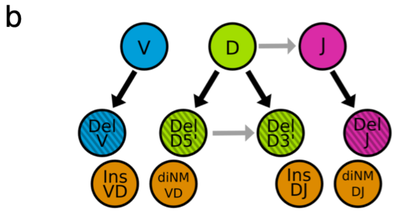
Strauli, Genome Medicine, 2016
Rempala, J Math Biol, 2013
Ahora, bioRxiv, 2018 Venturi, JIM, 2008

Repertoire dissimilarity index
Bolen, BMC Bioinformatics, 2017

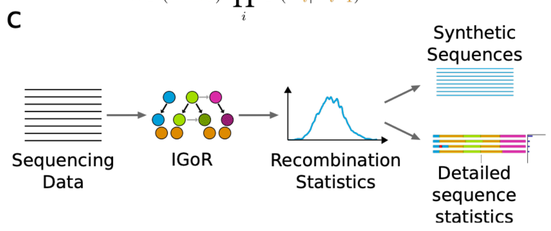
Inferring the recombination statistics of immune repertoires

Inference and Generation of Repertoires =IGoR





$$P(\text{scenario}) = P(V)P(D,J) \times P(\text{del}V|V)P(\text{Del}D5', \text{Del}D3'|D)P(\text{del}J|J) \times P(\text{InsVD}) \prod_{i}^{\text{InsVD}} P(n_i|n_{i-1}) \times P(\text{InsDJ}) \prod_{i}^{\text{InsDJ}} P(m_i|m_{i-1})$$



a IGoR lists putative recombination scenarios consistent with the observed sequence, and weighs them according to their likelihood.

b The likelihood of each scenario is computed using a Bayesian network of dependencies between the recombination features (V, D, J segment choices, insertions and deletions), as illustrated here for the human TRB locus.

c IGoR's pipeline includes three modes. In the **learning mode**, IGoR learns recombination statistics from data sequences. In the **analysis mode**, IGoR outputs detailed recombination scenario statistics for each sequence. In the **generation mode**, IGoR produces synthetic sequences with specified recombination statistics.

→ Distinguish between convergent recombination and convergent selection

→ Distinguish between public clones due to recombination and those due to antigen-driven selection

Marcou, Nat Comm 2018

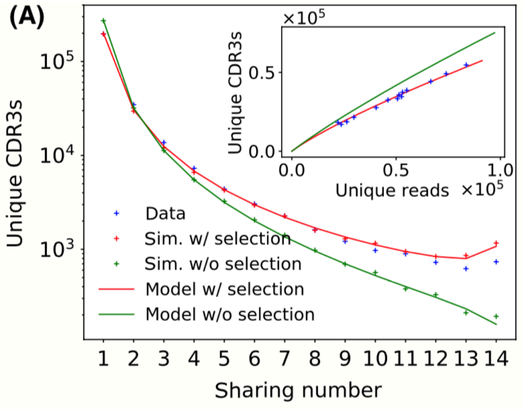
bitbucket.org/qmarcou/igor

Further literature:
Mora, PNAS, 2010
Murugan, PNAS,, 2012
Elhanati, Phil Trans R Soc B, 2015
Elhanati, Bioinformatics, 2016
Sethna, PNAS, 2017

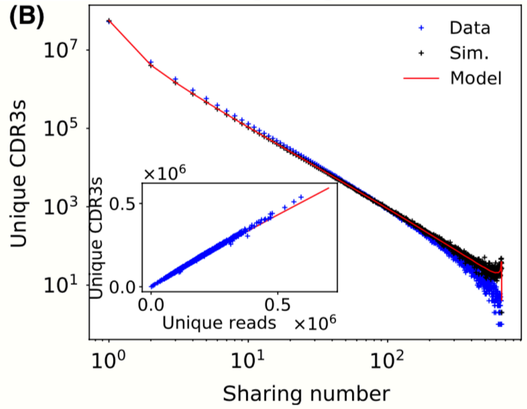
Predicting TCR public clone occurrence by modeling VDJ recombination

simple model of thymic selection. Whether a sequence is shared by many individuals is predicted to depend on the number of queried individuals and the sampling depth, as well as on the sequence itself, in agreement with the data. We introduce the *degree of publicness* conditional on the queried cohort size and the size of the sampled repertoires. Based on these observations, we propose a public/private sequence classifier,

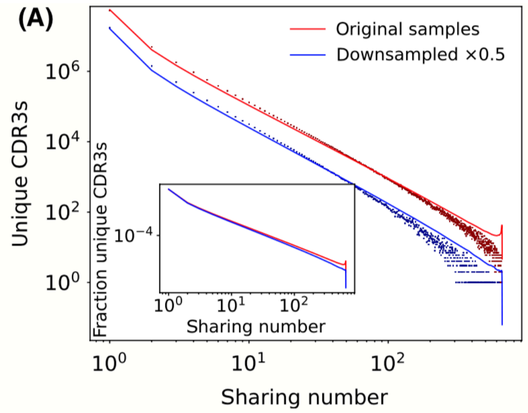
mouse

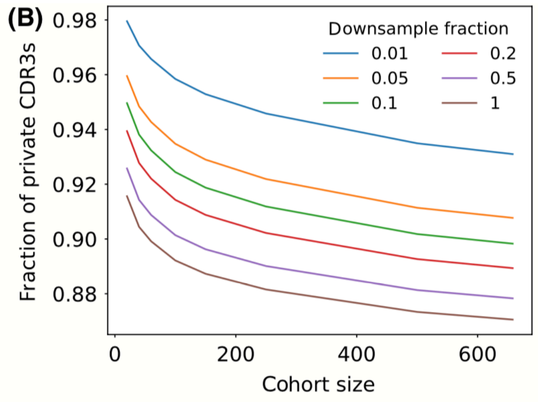


human

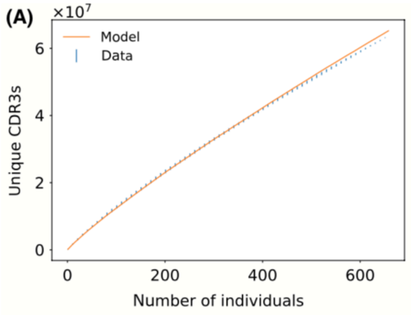


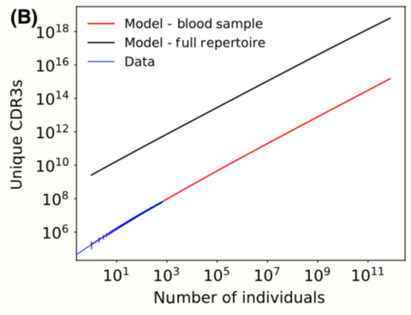
undersampling

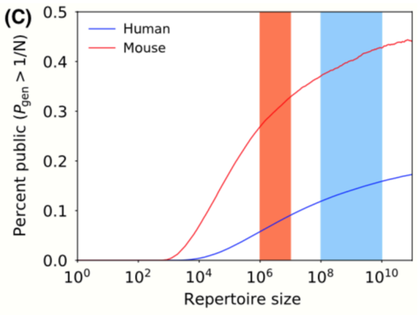




Number of public sequences per repertoire



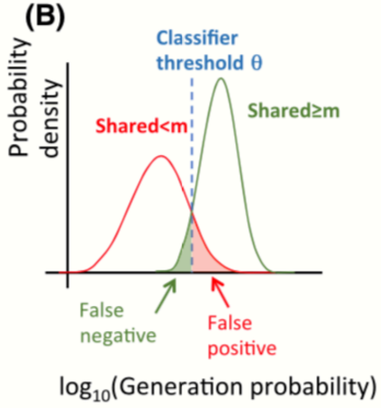


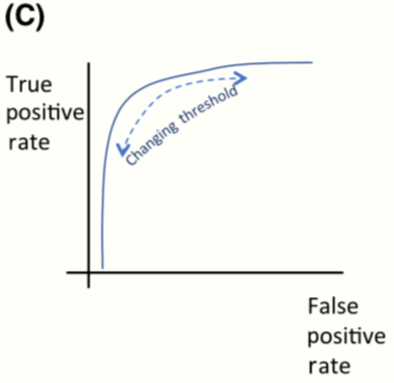


PUBLIC classifier

(A)

CDR3	Pgen	Sharing #
CASSELDTQYF	2.43E-08	23
CASSPPGELFF	4.36E-13	1
CASSLGTGARQPQHF	7.54E-10	4
CASSLGQGGSPLHF	2.53E-17	1

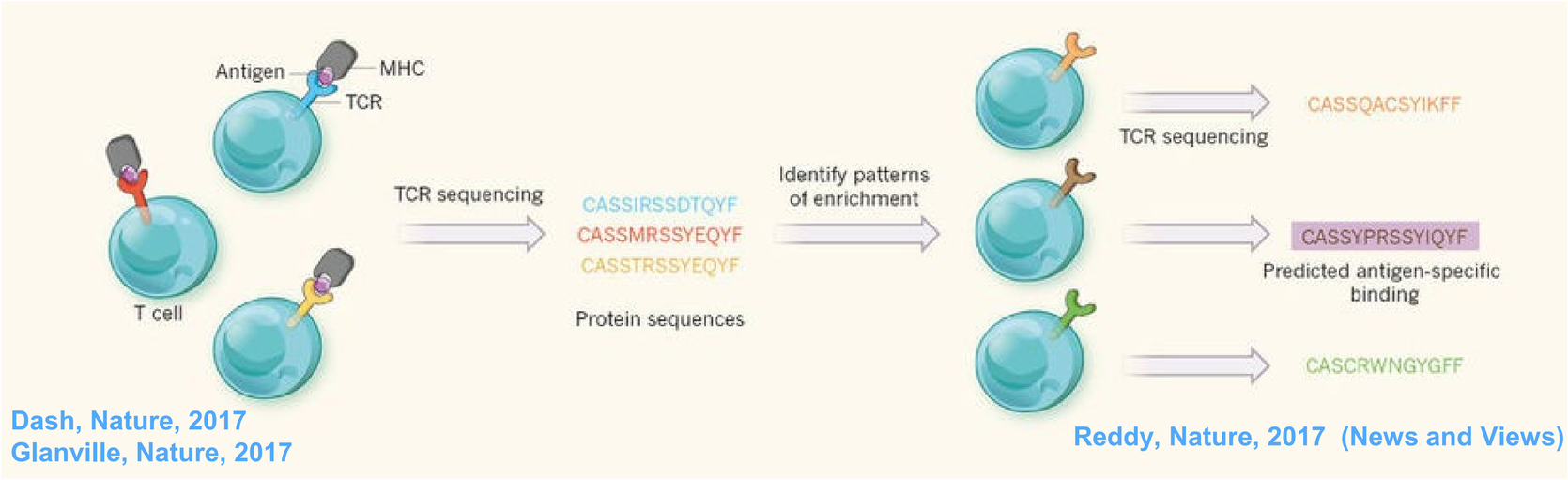




Open questions:
- influence of individual-specific models?
- influence of technology on models?

Elhanati, ImmuneRev, 2018

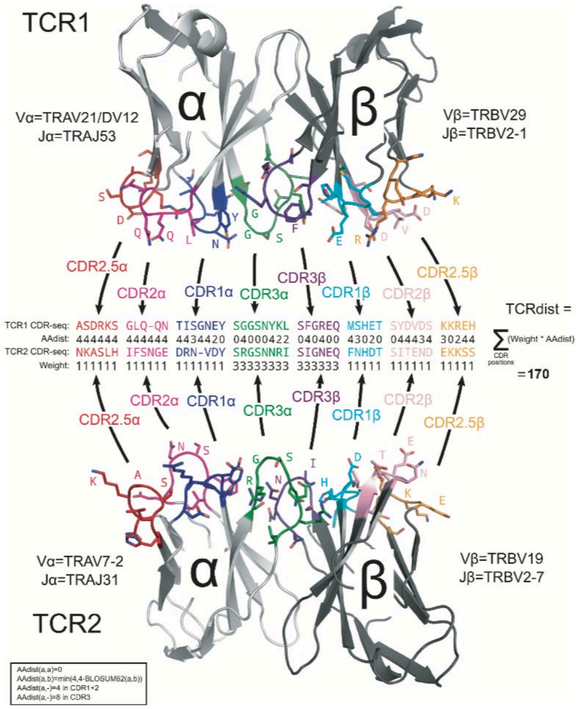
Prediction of TCR binding from the sequence by exploiting convergence



Dash, Nature, 2017
Glanville, Nature, 2017

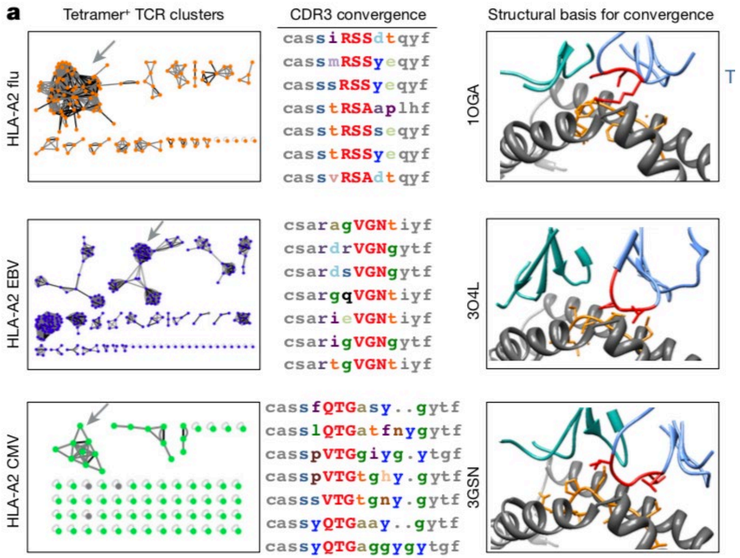
Reddy, Nature, 2017 (News and Views)

Epitope-specific TCR repertoires of CD8⁺ T cells from mice and humans, representing over 4,600 in-frame single-cell-derived TCRαβ sequence pairs from 110 subjects



Dash, Nature, 2017

Distance-based classifier (TCRdist) that assigns previously unobserved TCRs to characterize repertoires with robust sensitivity and specificity.



HC–tetramer-sorted antigen-specific TCR repertoires of EBV, influenza, CMV as well as public sources (*n* = 2,068).

GLIPH (grouping of lymphocyte interactions by paratope hotspots) to cluster TCRs with a high probability of sharing specificity owing to both conserved motifs and global similarity of complementarity-determining region 3 (CDR3) sequences.

Glanville, Nature, 2017
Huang, NBT, 2020

Summary: Measuring immune repertoire diversity

- Diversity is one of the hallmark features of adaptive immune repertoires. Therefore, its measurement lays the foundation for the majority of repertoire statistics

- Diversity can be quantified using methods borrowed from mathematical ecology

- Diversity profiles are superior to single diversity indices when comparing clonal frequency distribution across samples

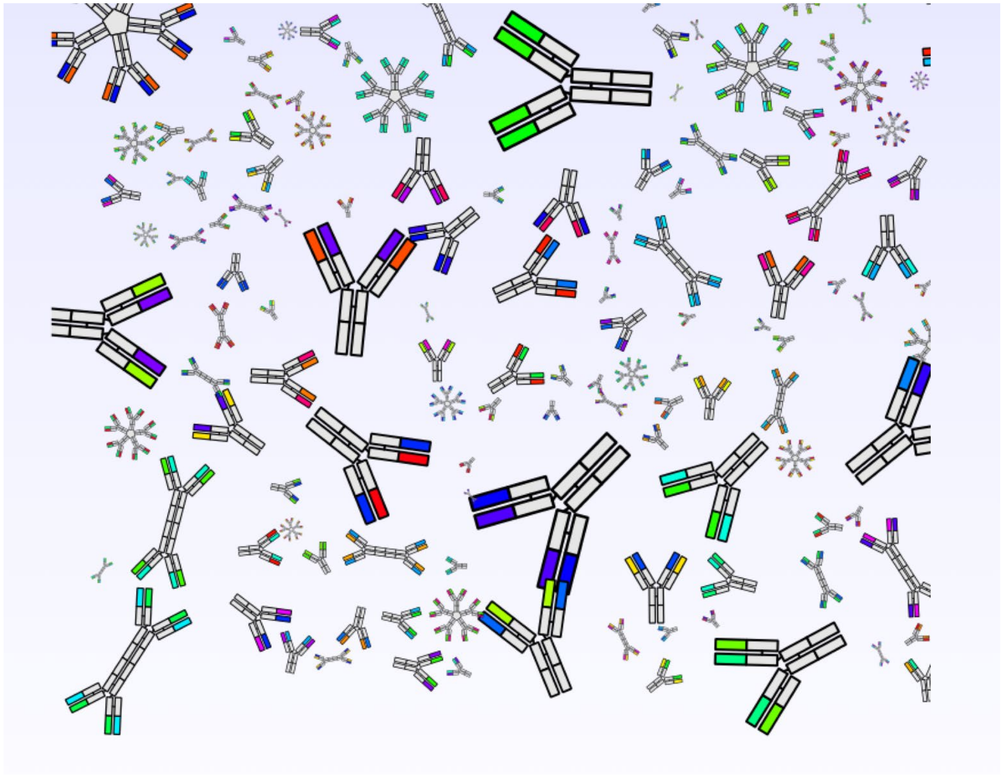
- Diversity holds immune information (the extent of which remains unclear)

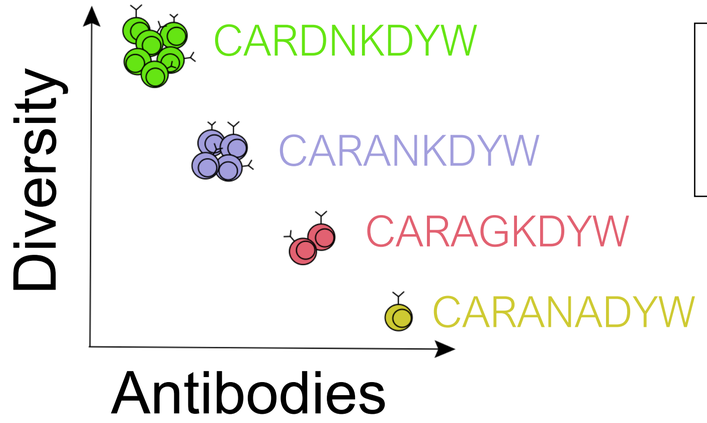
- VDJ recombination statistics can be inferred using Bayesian statistics

- Repertoire convergence (overlap) may be quantified from several perspectives and may be leveraged for the prediction of antigen specificity

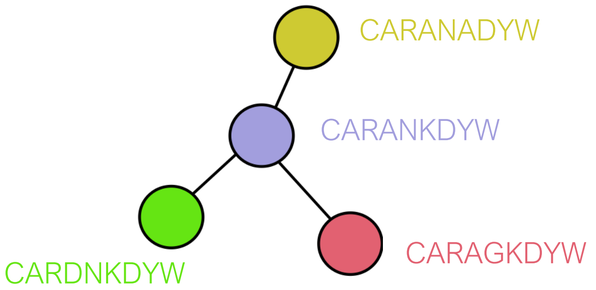
Networks for the analysis of antibody repertoire architecture (sequence similarity among sequences within a repertoire)

Immune repertoire





Diversity analysis provides **no information on sequence similarity**



Network analysis **resolves the similarity relation (architecture)** of antigen receptor sequences

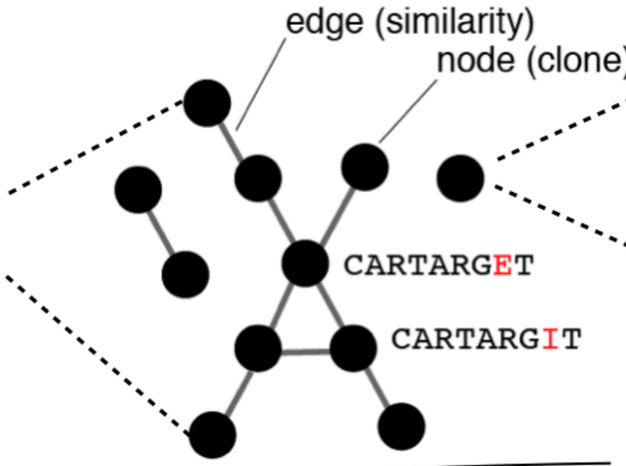
Together, **diversity and network analysis resolve the frequency and similarity** information of immune repertoires

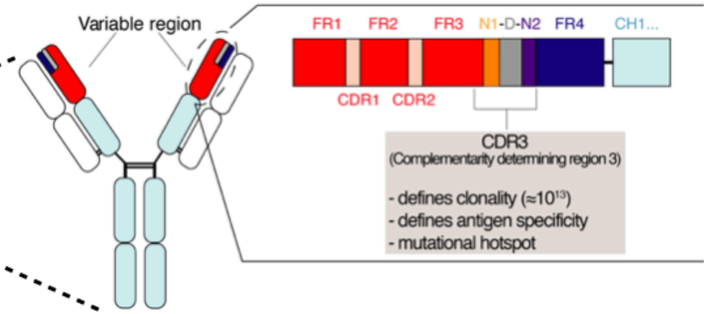
$$ {}^{q}D_s = \left(\sum_i p_i S_i{}^{q-1}\right)^{1/(1-q)} $$

Arora, bioRxiv, 2019

Building networks from immune repertoire sequence data

Network of entire antibody repertoire = Antibody repertoire architecture



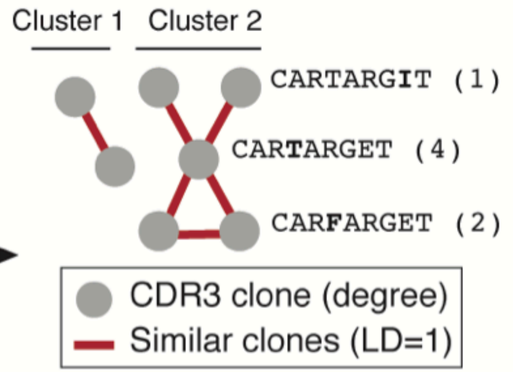


A **node** can represent e.g., on the **entire evariable region (FR1 – FR4)** or just the **CDR3**

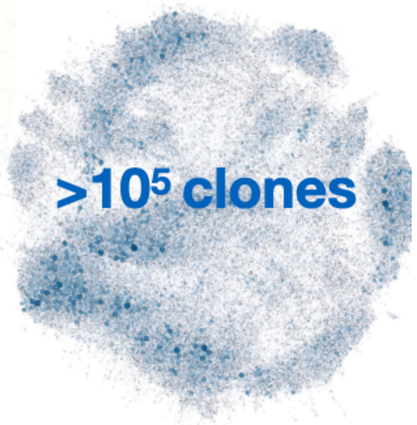
CDR3 repertoire	CARTARGET	CARFARGET	CARFARGIT	...	$CDR3_{10^2 \le n < 1}$
CARTARGET	0	1	2	...	LD_{1n}
CARFARGET		0	1	...	LD_{2n}
CARTARGIT			0	...	LD_{3n}
⋮	⋮	⋮	⋮	⋱	⋮
$CDR3_{10^2 \le n < 10^6}$	LD_{n1}	LD_{n2}	LD_{n3}	...	LD_{nn}

Levenshtein distance (LD) matrix

Antibody repertoire network construction →



Large-scale network visualization of the repertoire ⇢

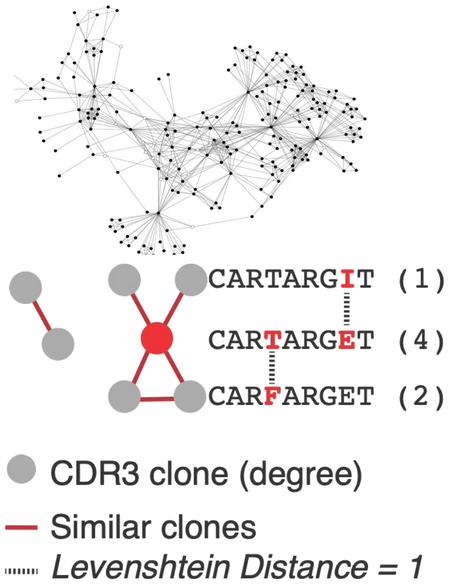


Computation of Levenshtein distance (LD) is **computation and memory expensive (all-by-all comparison) → high-performance computing needed**

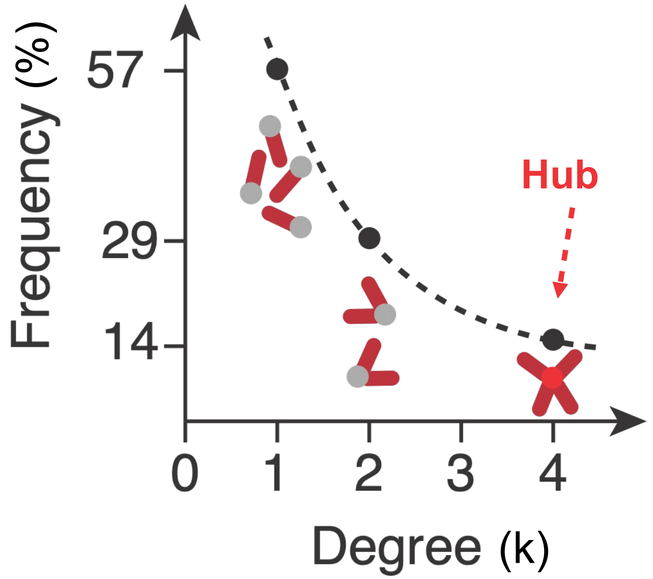
Similarity may also be defined **by LD =2, 3, etc.. (similarity layers)**

Miho, Nat Comms, 2019

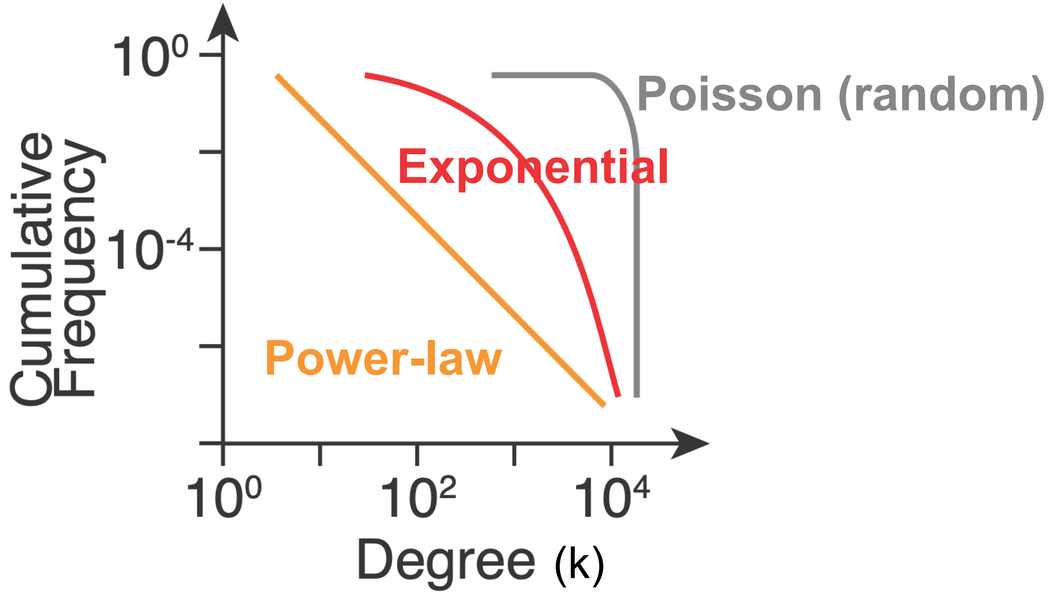
Quantitative analysis of immune repertoire networks



CDR3 degree (nr. of links) distribution

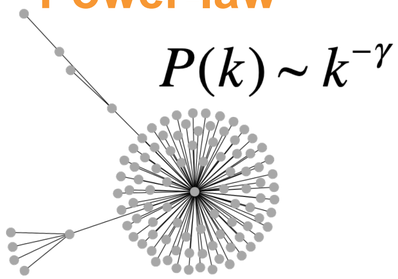


Cumulative (log-log)



The degree distribution quantifies the structure of the network

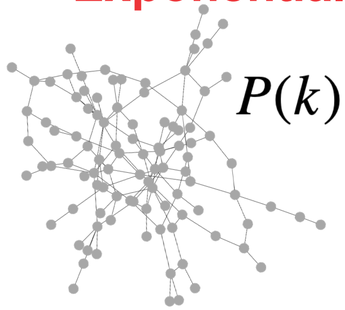
Power-law



$$P(k) \sim k^{-\gamma}$$

≈Antigen-experienced repertoire

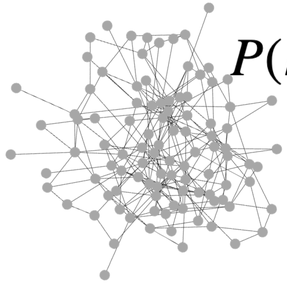
Exponential



$$P(k) \sim e^{\frac{-k}{c}}$$

≈Naïve repertoire

Poisson (random)

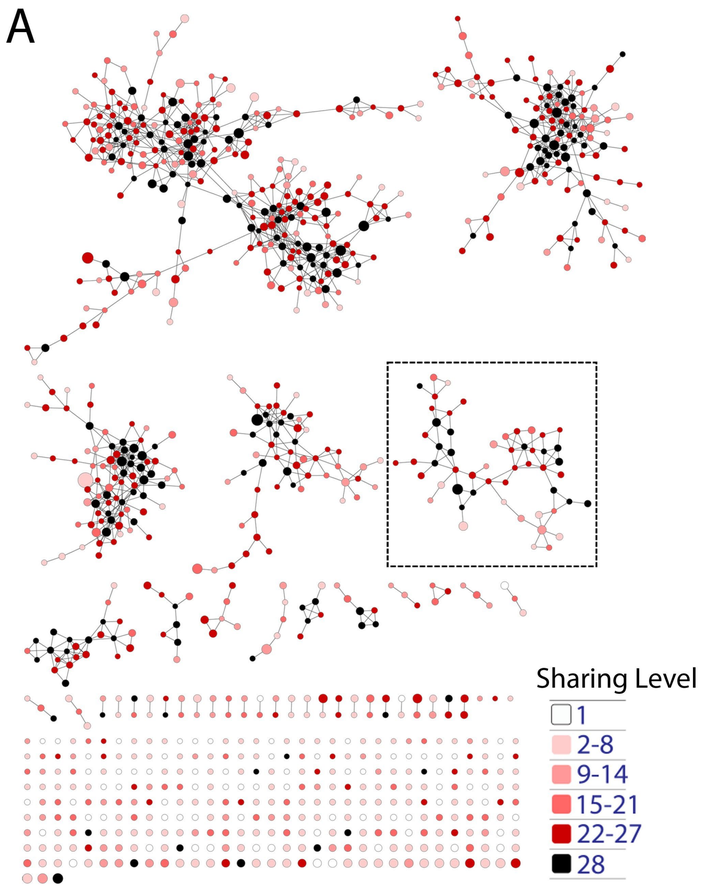


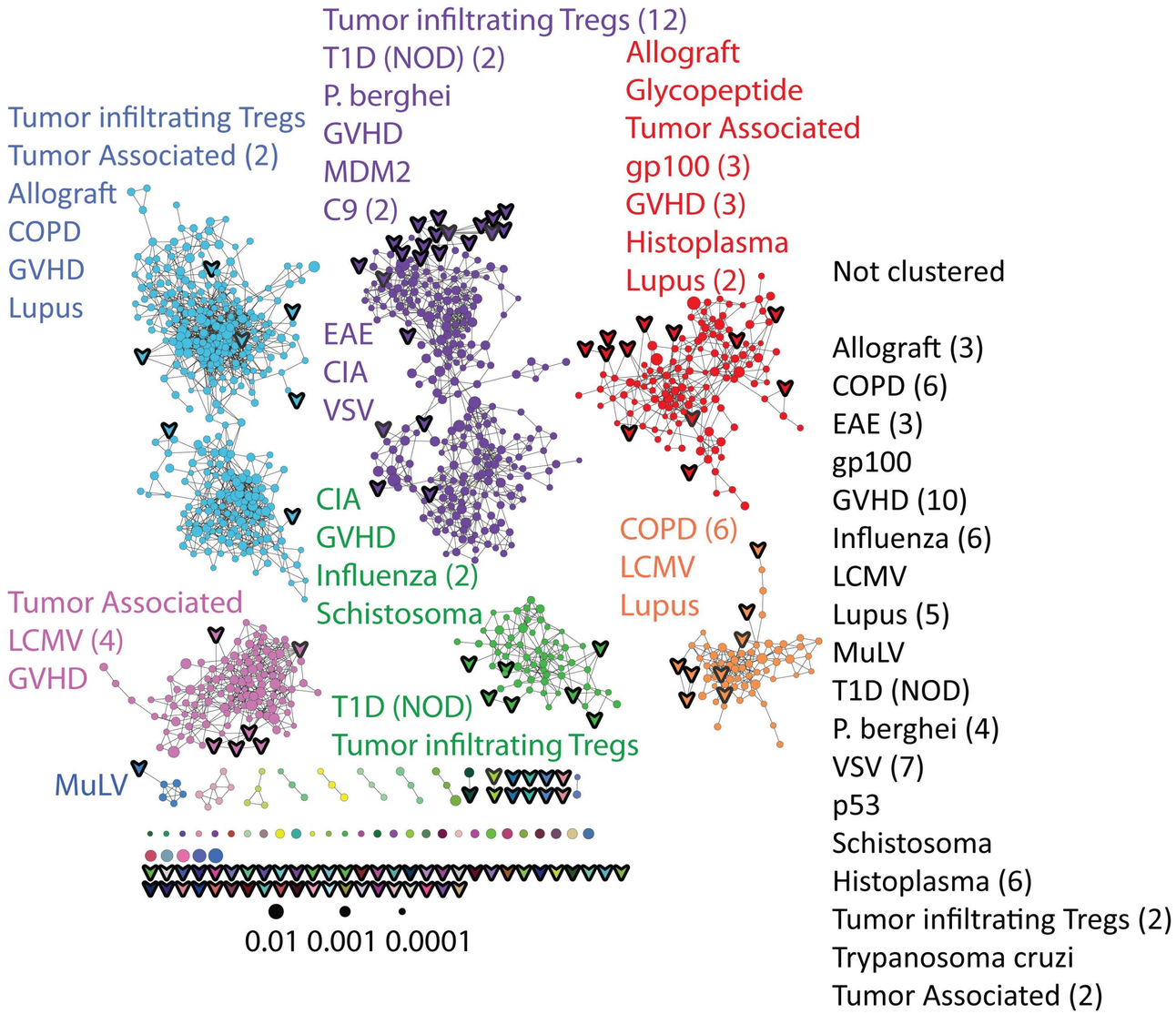
$$P(k) \sim e^{-<k>} \frac{<k>^k}{k!}$$

Figure modified from Enkelejda Miho

Insights into AIRR biology afforded by network analsis I

T cell receptor repertoires of mice and humans are clustered in similarity networks around conserved public CDR3 sequences





Further literature:
Pogorelyy et al., 2017, PNAS, 2018
Madi et al., 2017, Gen Res, 2017
Chang et al., Sci Rep, 2016
Linder et al., Nat Immunol, 2015
Hoehn et al., Philos Trans R Soc B, 2015
Bashford-Rogers et al., Genome Res, 2013

Madi et al., Gen Res, 2017

Insights into AIRR biology afforded by network analsis II

→ **Network structure is redundant across similarity layers**

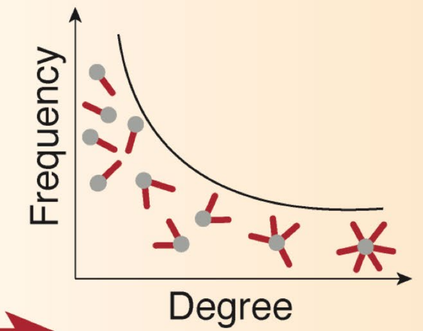
→ **Antibody networks are robust to random removal of clonal** (also T cell networks, Madi, elife, 2017)

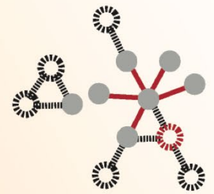
Levenstein Distance (LD) matrix calculation using high-performance computing platform

CDR3 repertoire	CARTARGET	CARFARGET	CARFARGIT	...	$CDR3_{10^2 \le n < 10^6}$
CARTARGET	0	1	2	...	LD_{1n}
CARFARGET		0	1	...	LD_{2n}
CARTARGIT			0	...	LD_{3n}
⋮	⋮	⋮	⋮	⋱	⋮
$CDR3_{10^2 \le n < 10^6}$	LD_{n1}	LD_{n2}	LD_{n3}	...	LD_{nn}

Antibody repertoire network construction

Deconvolution



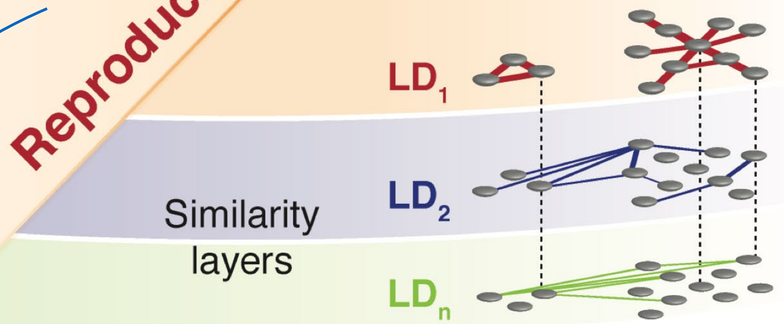


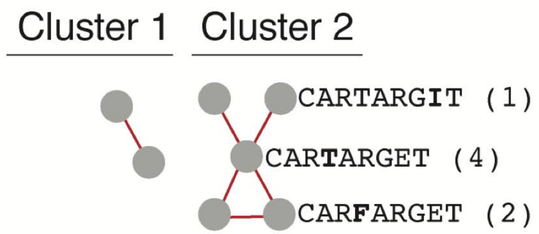
Public
Random

Robustness

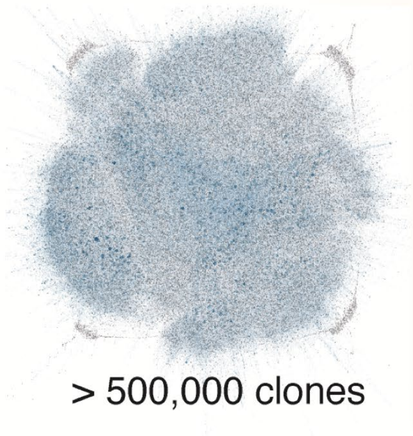
Redundancy

Reproducibility





Large-scale network visualization of the repertoire



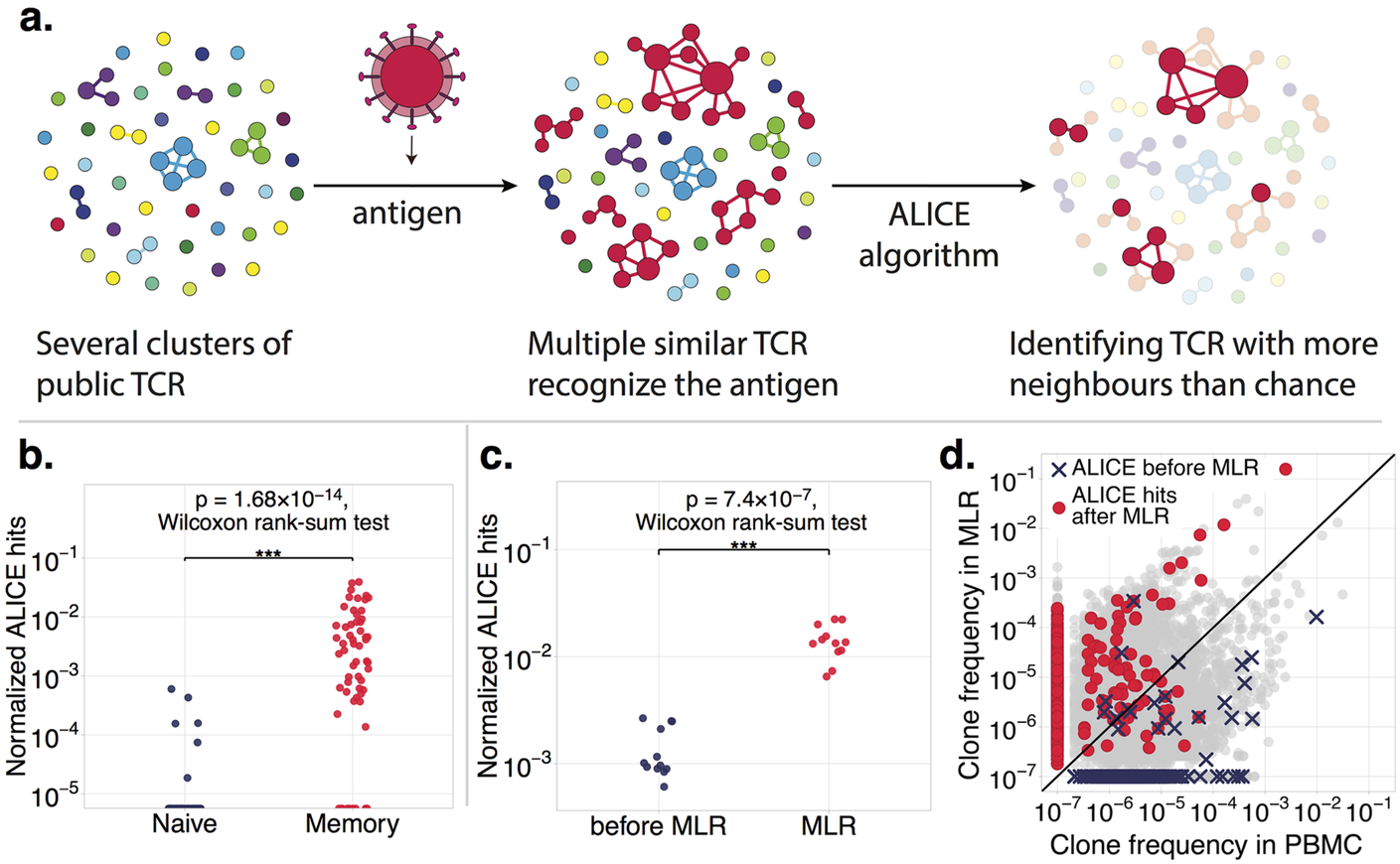
→ **Network measures are reproducible across mice**

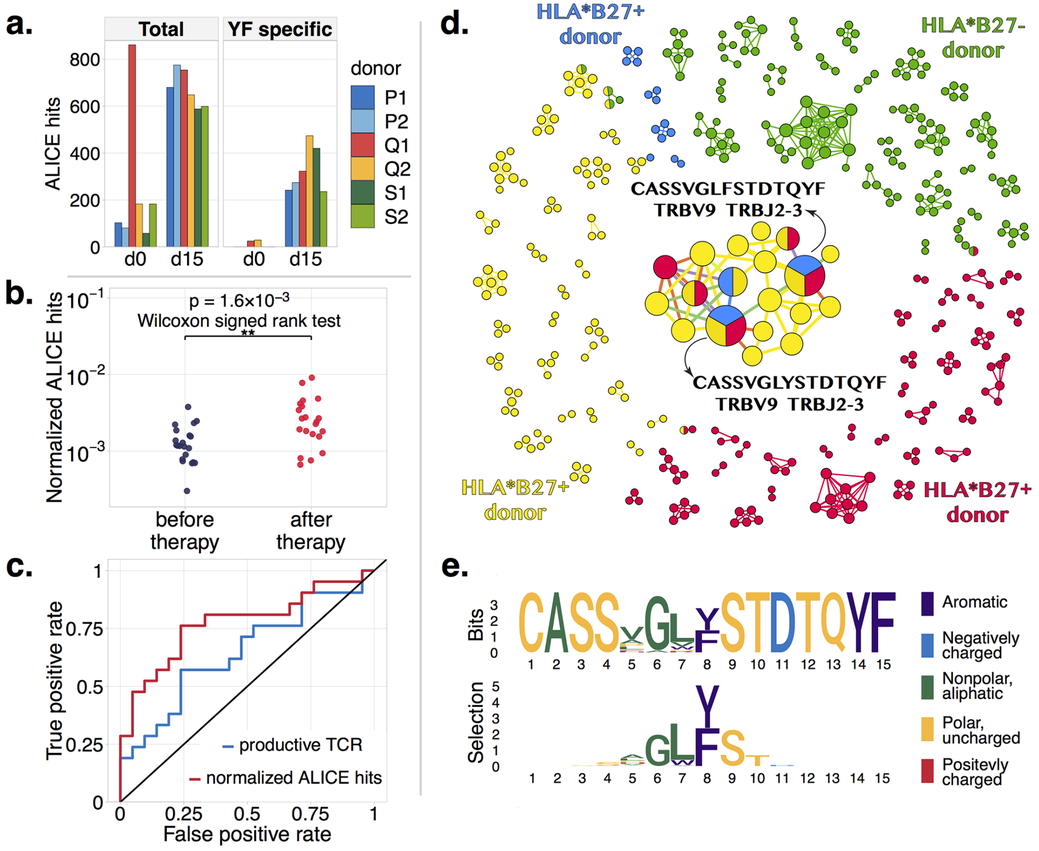
Miho, Nat Comm, 2019

Insights into AIRR biology afforded by network analysis III

Detecting T cell receptors involved in immune responses from single repertoire snapshots

Mikhail V. Pogorelyy, Anastasia A. Minervina, Mikhail Shugay, Dmitriy M. Chudakov, Yuri B. Lebedev, Thierry Mora, Aleksandra M. Walczak





Pogorelyy, PLOS Biology, 2019

Summary: Measuring immune repertoire architecture

- Network architecture determines antigen recognition breadth

- Quantitative and not visual analysis of antibody networks allows insight into the construction principles of antibody repertoires

- Construction of large-scale networks (>10^5 clonal sequences) requires high-performance computing

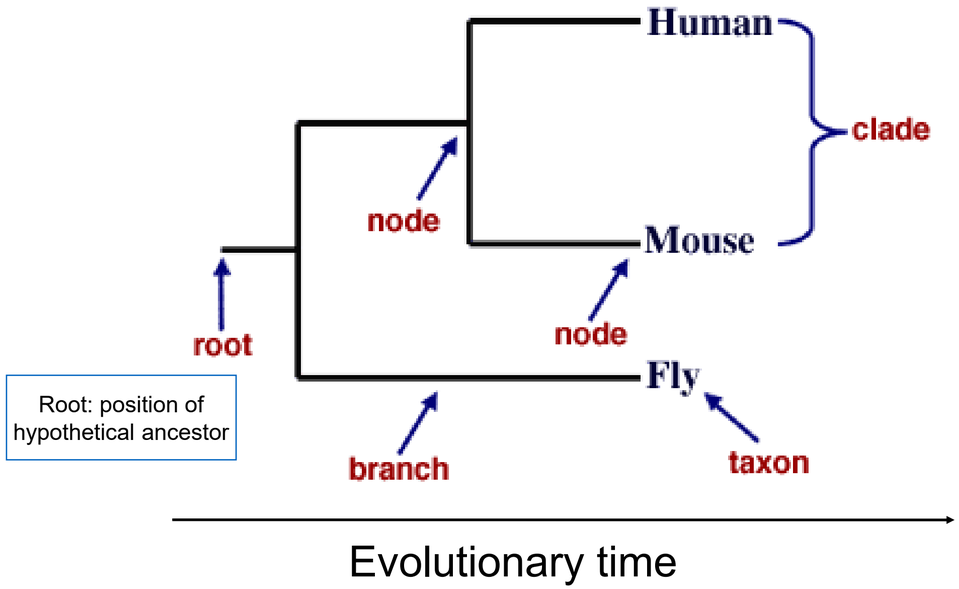
- Public clones play a special (but yet undetermined) structural role in antibody and T cell networks

Phylogenetics: retracing antibody evolution

Application of phylogenetics in antibody repertoire research

- Goal: infer evolutionary relationship between antibody sequences and visualize diversification of B-cell lineages in response to antigen
- Detect selection on B cell lineages
- Detect and quantify dynamics of affinity maturation
- Reconstruct evolution of broadly neutralizing antibodies

Hoehn, MBE, 2016

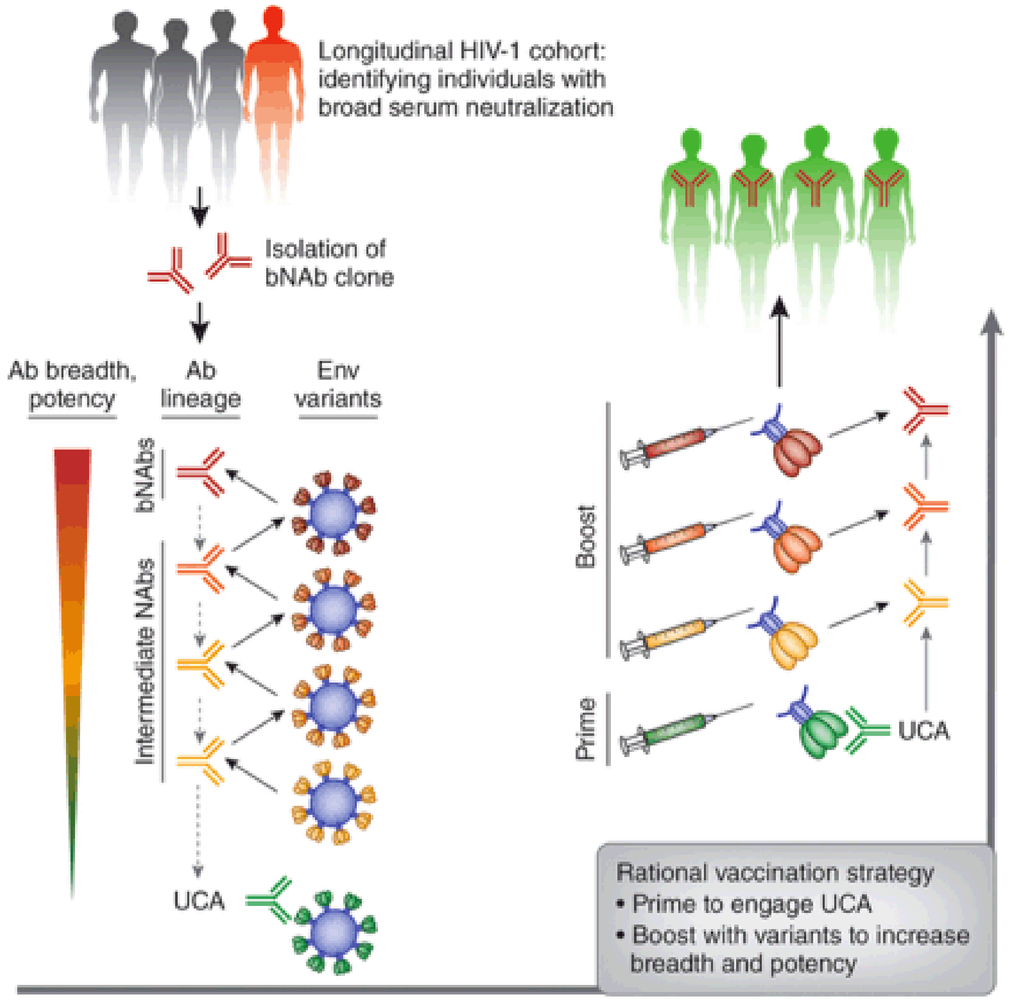


Root: position of hypothetical ancestor

Clade: common ancestor + descendants representing a single "branch" on a tree

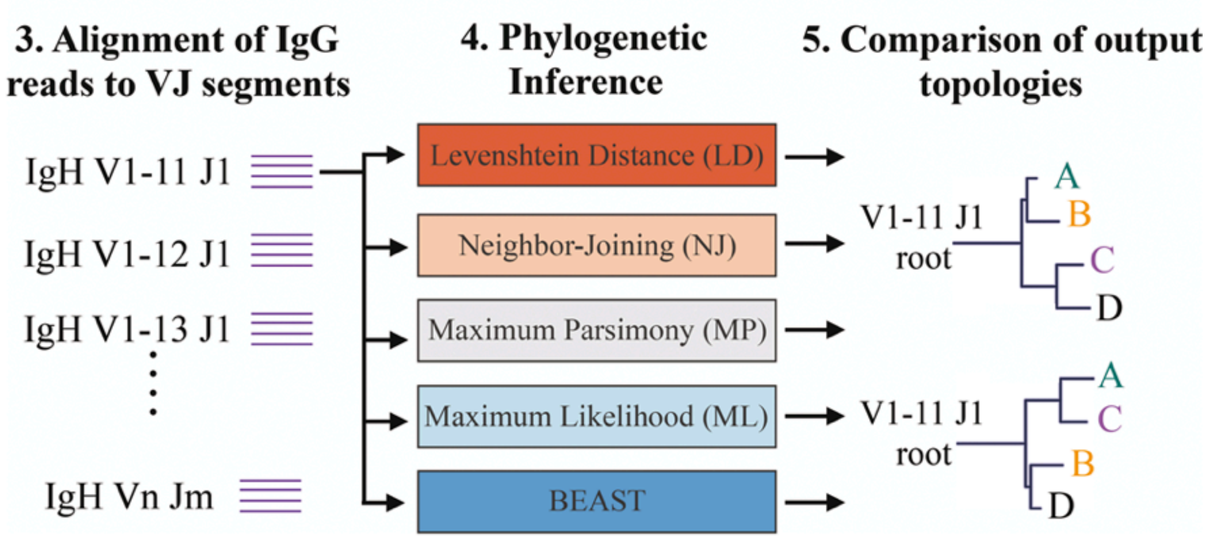
Lineage: separate V-D-J recombination events (can be computationally preselected for by restricting data to sequences sharing V, J, and CDR3 length)

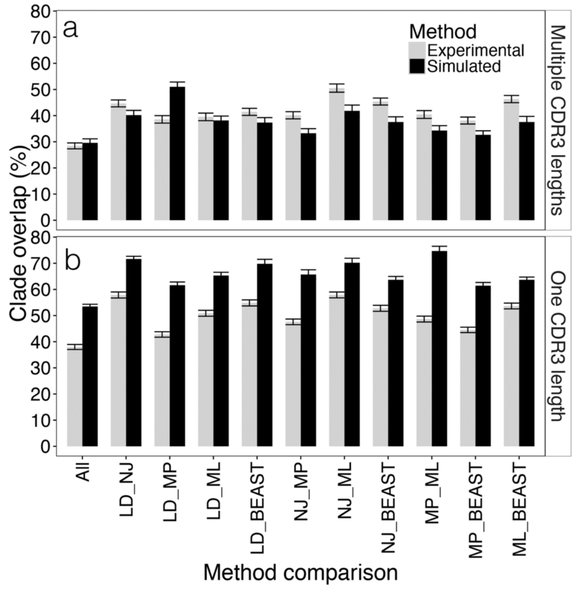
Figure 1: Deciphering bNAb development in an HIV-1–infected subject to guide vaccine strategies.

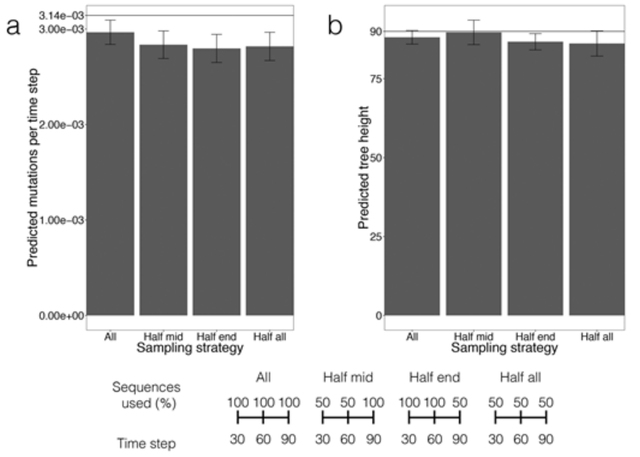


Gruell & Klein, Nat Med, 2014

Most common methods used for phylogenetic inference







LD, NJ

distance-based methods that rely upon an initial all-by-all distance matrix calculation (fast computation)

MP

non-parametrically selecting the shortest possible tree that explains the data (fast computation)

ML, BEAST

infer lineage evolution using probabilistic methods, which can incorporate biologically relevant parameters such as transition/transversion rate and nucleotide frequencies (slow computation)

Yermanos , Bioinformatics, 2017

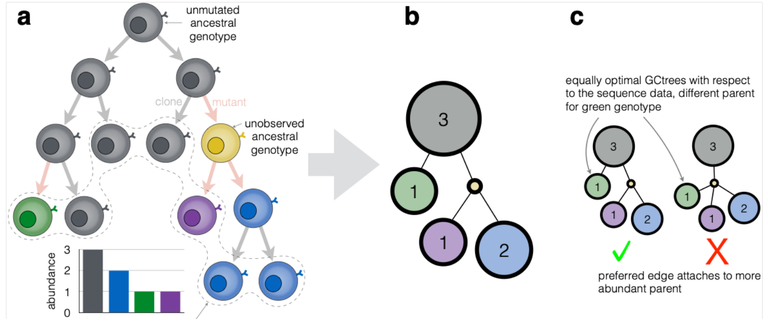
Overview article: Yang & Rannala, Nat Rev Gen, 2012

→Phylogenetic approaches may lead to different tree topologies

→ BEAST can be used to infer the duration of evolution and SHM rate

Yermanos , Bioinformatics, 2017

Using genotype abundance to improve phylogenetic inference



c Intuitively, the abundance information indicates that the tree on the left is preferable because the more abundant parent is more likely to have generated mutant descendants.

DeWitt, Mol Bio and Evo, 2018

Recent advances in AIRR phylogenetic analysis: tree significance and incorporation of antibody affinity

RESEARCH ARTICLE

Using B cell receptor lineage structures to predict affinity

PLOS Comp Biol 2020

Duncan K. Ralph*, Frederick A. Matsen IV

Fred Hutchinson Cancer Research Center, Seattle, Washington, USA

* dralph@fredhutch.org

- A method that uses evolutionary information from the family of related sequences that share a naive ancestor to predict the affinity of each resulting antibody for its antigen. When combined with information on the identity of the antigen, this method should provide a source of effective new antibodies.
- A method for a related task: given an antibody of interest and its inferred ancestral lineage, which branches in the tree are likely to harbor key affinity-increasing mutations

New Results

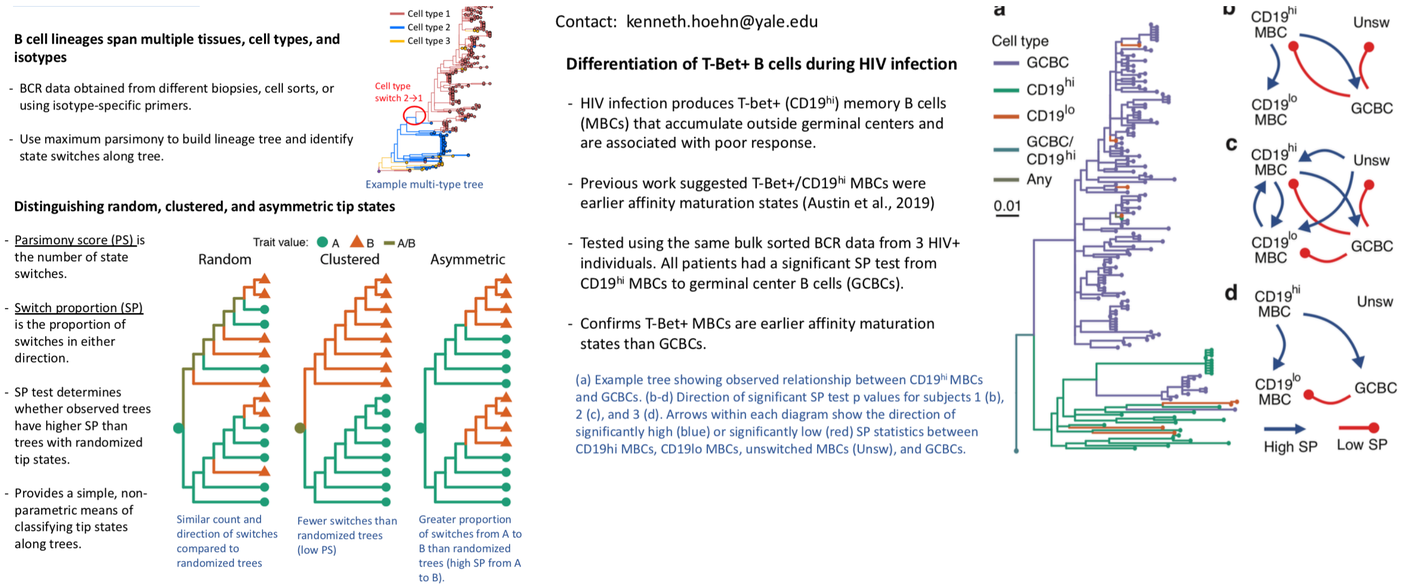
Comment on this paper

Phylogenetic analysis of migration, differentiation, and class switching in B cells

Kenneth B. Hoehn, Oliver G. Pybus, Steven H. Kleinstein

doi: https://doi.org/10.1101/2020.05.30.124446

- Statistical method for characterizing migration, differentiation, and isotype switching along B cell phylogenetic trees.



https://bitbucket.org/kleinstein/dowser

Summary: Retracing antibody evolution (phylogenetics)

- Antibody evolution is a hallmark feature of the antigen-driven adaptive immune response: its faithful reconstruction may lead to profound insight into the mechanisms of selection that govern the formation of antigen-specific repertoires

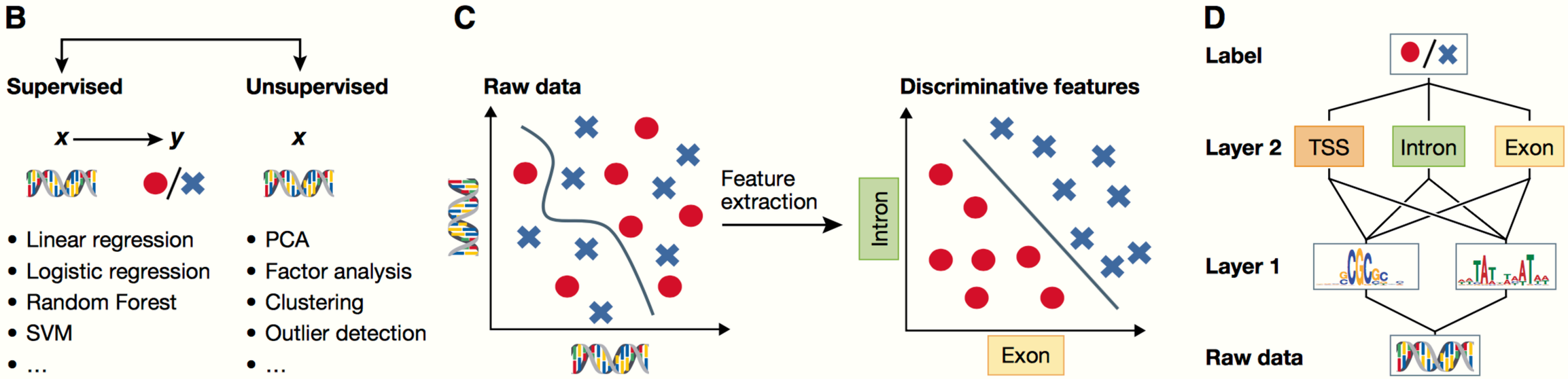
- Many methods exist for phylogenetic inference: they do not only differ in assumption and speed but may also differ in the resulting lineage trees

- Recently, progress has been made in coupling antibody abundance with antibody sequence information in order to more accurately reflect antibody evolution

- Mutability maps may help in increasing the accuracy of phylogenetic models

- Comparison of tree topologies remains a crucial challenge

Machine learning: a general overview



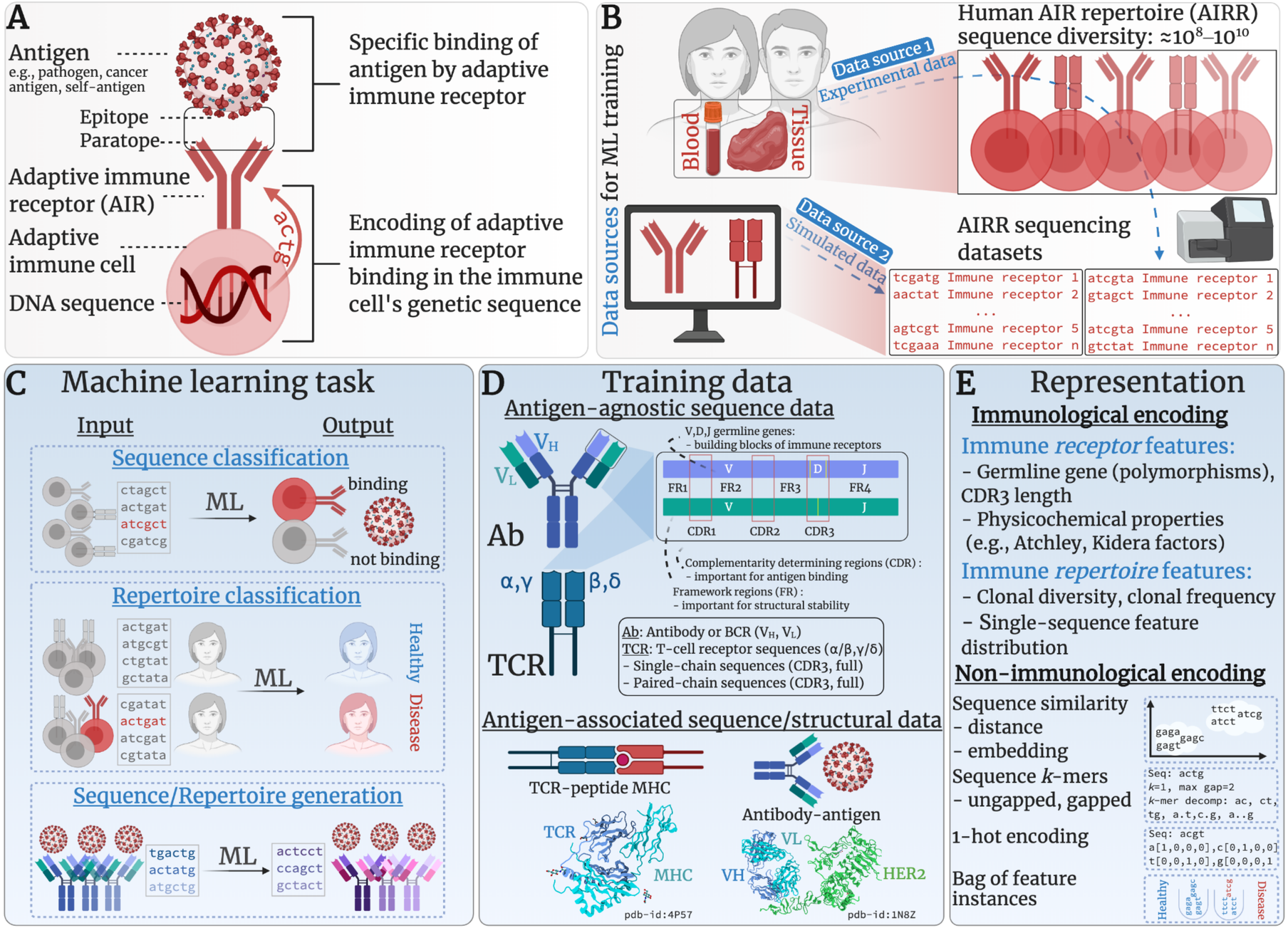
Supervised machine learning methods relate input features x to **an output class label** y, whereas **unsupervised methods** learn factors about x **without assigned class labels**.

Input data are often **high-dimensional,** which is challenging for many classical machine learning algorithms. **Alternatively, higher-level features extracted using a deep model may be able to better discriminate between classes**.

Deep networks use a hierarchical structure to learn increasingly **abstract feature representations** from the raw data.

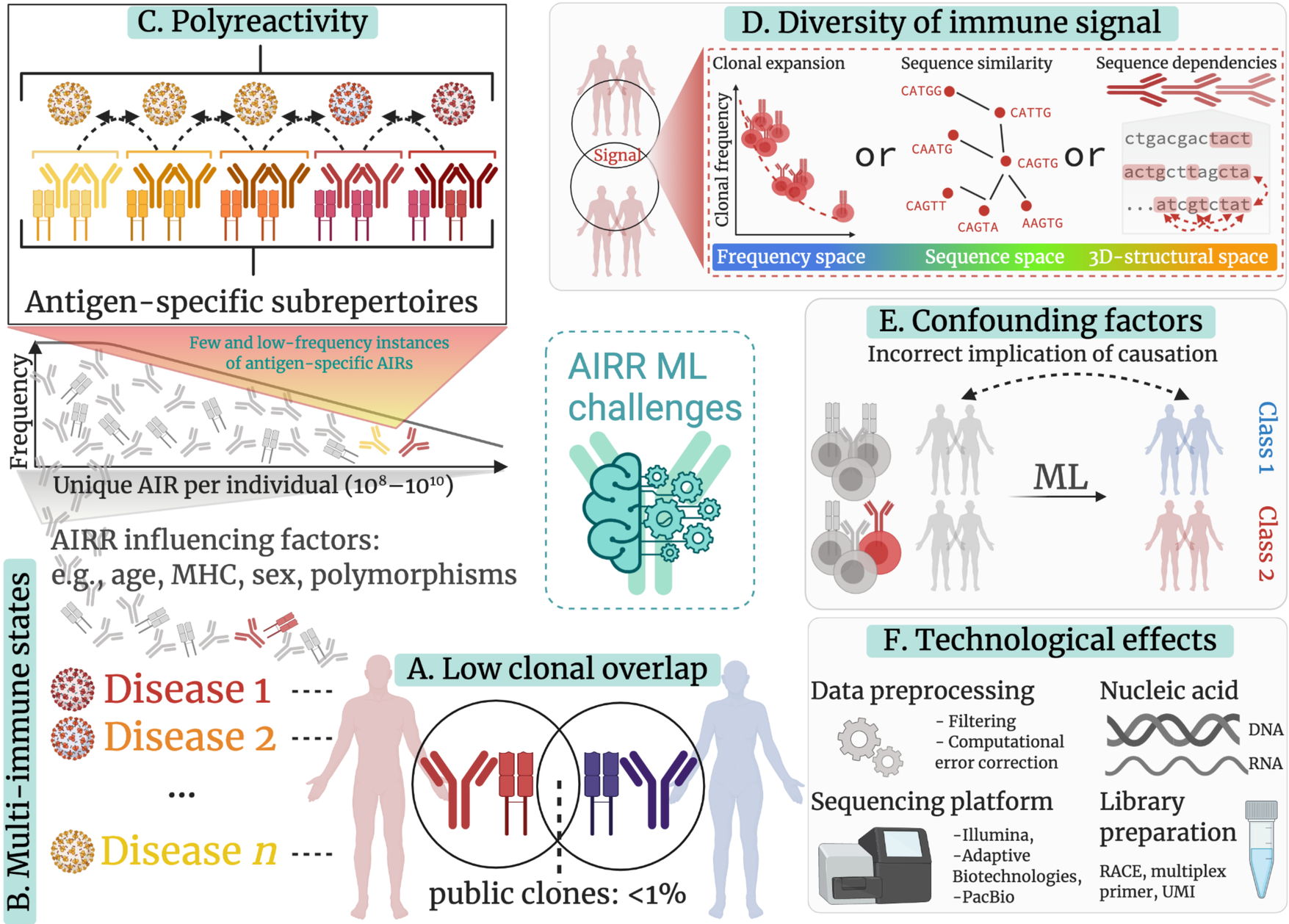
Angermueller, Mol Sys Biol, 2016

Machine learning enables the deciphering and prediction of immune receptors



Greiff, Cowell, Yaari
Curr Op Sys Bio, 2020

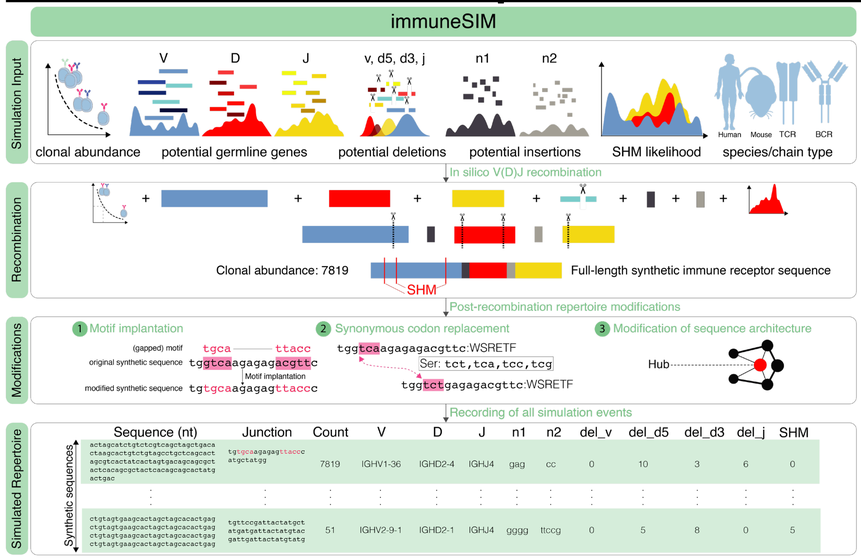
Specific biological and machine learning challenges for immune receptor research

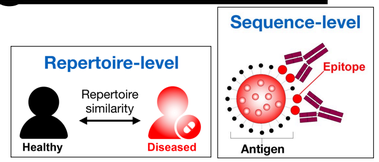


Greiff, Cowell, Yaari
Curr Op Sys Bio, 2020

Machine learning approaches applied to adaptive immune receptor data

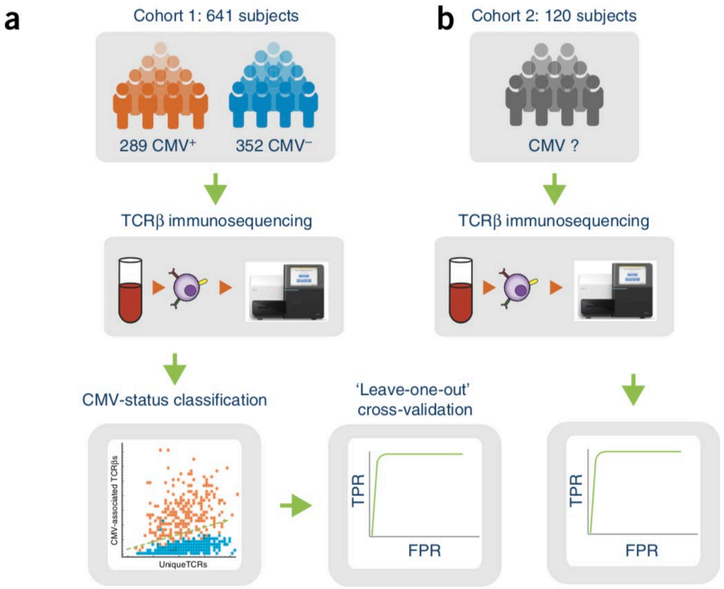
Ground truth sequence data generation

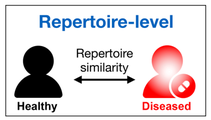




Marcou et al., Nat Comms, 2018
Olson et al., Front Imm, 2019
Weber, et al., Bioinformatics, 2020
…

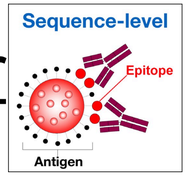
Prediction of immune state

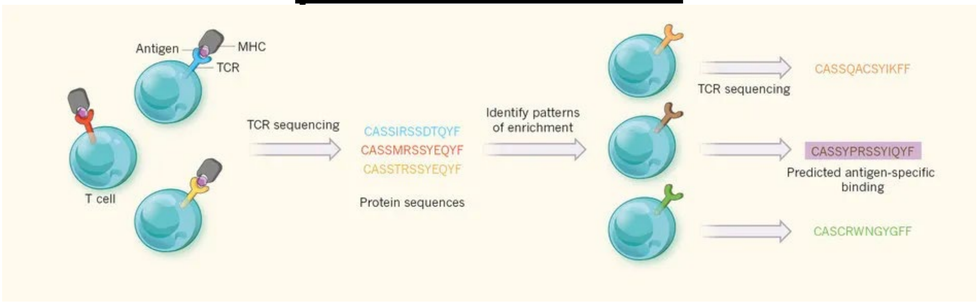




Emerson et al., Nat Gen, 2017
Ostmeyer et al., Cancer Res, 2019
Widrich et al., NeurIPS, 2020
Beshnova et al., Sci Trans Med, 2020
Sidhom et al., Nat Comms, 2021
Shemesh et al., Front Imm, 2021
…

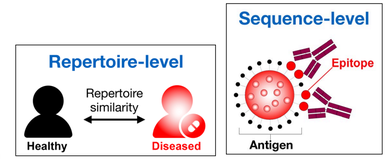
Prediction of antigen binding or public clones

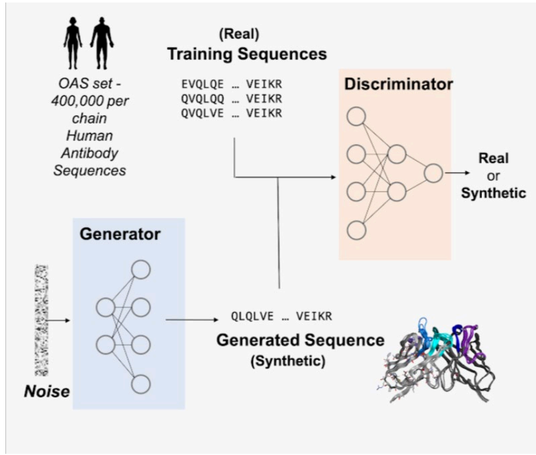




Dash et al., Nature, 2017
Glanville et al., Nature, 2017
Greiff et al., JI, 2017
Elhanati et al., ImmRev, 2018
Fischer et al., Mol Sys Bio, 2020
Moris et al., Brief in Bioinf, 2020
Huang et al. ; NBT, 2020
Akbar et al., Cell Reports, 2021
Sidhom et al., Nat Comms, 2021
…

Generative modeling





Davidsen et al., elife, 2019
Amimeur et al., bioRxiv, 2020
Friedensohn et al., bioRxiv, 2020
…

Multi-class accuracy assessment

Macro: biases your metric toward the least populated classes.

$$\text{Precision}_M = \frac{\sum_{i=1}^{n} \frac{TP_i}{(TP_i+FP_i)}}{n} \qquad \text{Recall}_M = \frac{\sum_{i=1}^{n} \frac{TP_i}{(TP_i+FN_i)}}{n} \qquad \text{F–score}_M = \frac{2 \times \text{Pre}_M \times \text{Rec}_M}{\text{Pre}_M + \text{Rec}_M}$$

Micro: bias your metric towards the most populated classes.

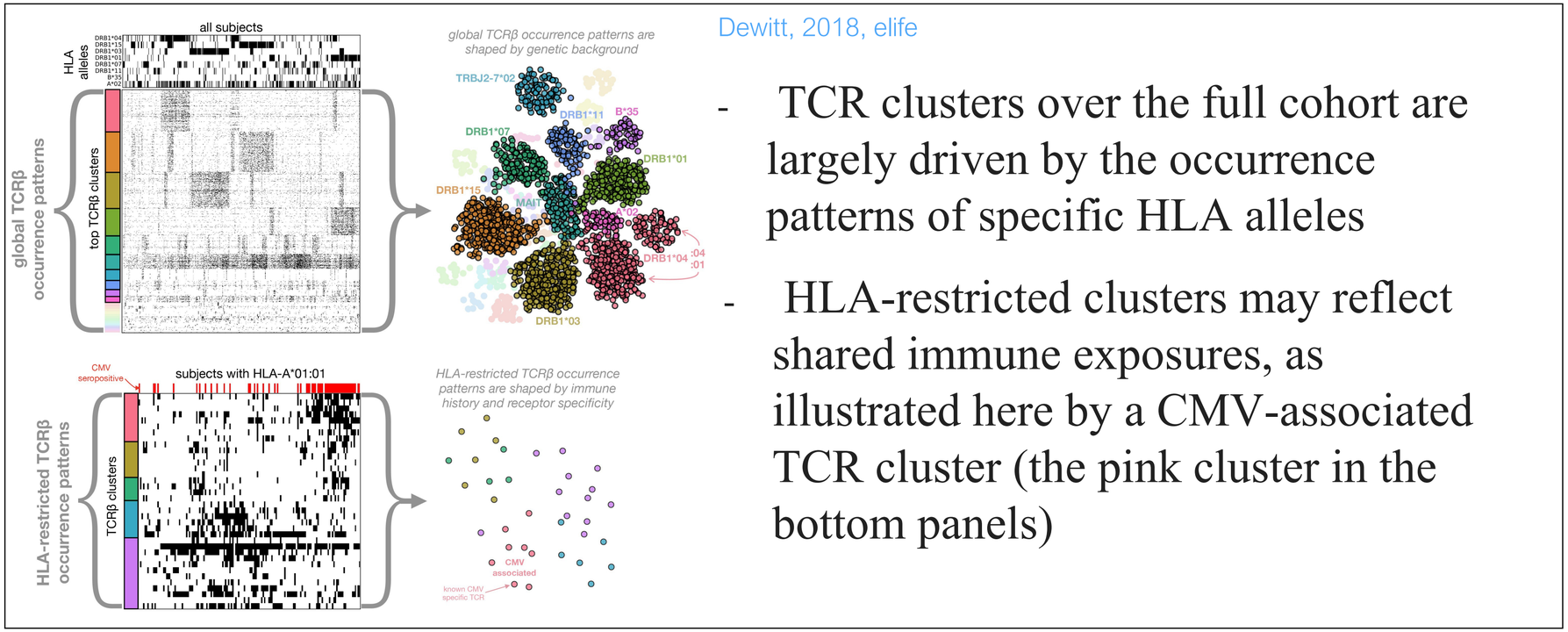
$$\text{Precision}_\mu = \frac{\sum_{i=1}^{n} TP_i}{\sum_{i=1}^{n} (TP_i + FP_i)} \qquad \text{Recall}_\mu = \frac{\sum_{i=1}^{n} TP_i}{\sum_{i=1}^{n} (TP_i + FN_i)} \qquad \text{F–score}_\mu = \frac{2 \times \text{Pre}_\mu \times \text{Rec}_\mu}{\text{Pre}_\mu + \text{Rec}_\mu}$$

If macro << micro:
smaller classes are poorly classified, whereas larger ones are likely correctly classified.

If macro >> micro:
gross misclassification in the most populated classes, whereas smaller classes are likely correctly classified.

Sokolova, Information Processing & Management, 2009

Effects of MHC, age, and sex on AIRR



Dewitt, 2018, elife

- TCR clusters over the full cohort are largely driven by the occurrence patterns of specific HLA alleles
- HLA-restricted clusters may reflect shared immune exposures, as illustrated here by a CMV-associated TCR cluster (the pink cluster in the bottom panels)

Sex bias in MHC I-associated shaping of the adaptive immune system

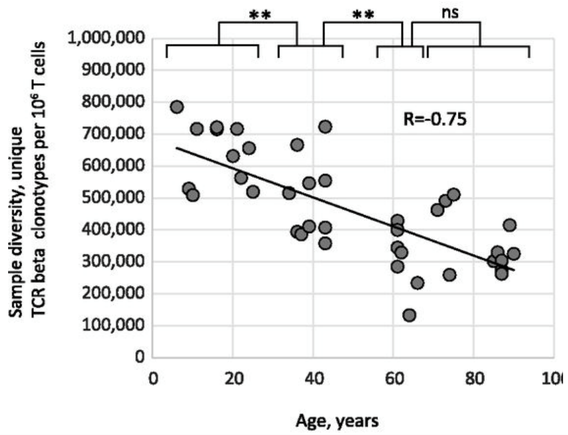
Tilman Schneider-Hohendorf[a], Dennis Görlich[b], Paula Savola[c], Tiina Kelkka[c], Satu Mustjoki[c], Catharina C. Gross[a], Geoffrey C. Owens[d], Luisa Klotz[a], Klaus Dornmair[e], Heinz Wiendl[a], and Nicholas Schwab[a,1]

"Next-generation TCR variable beta chain (TCRBV) immunosequencing of 824 individuals was evaluated in a multiparametric analysis including HLA-A -B/MHC class I background, TCRBV usage, sex, age, ethnicity, and TCRBV selection/expansion dynamics. We found that HLA-associated shaping of TCRBV usage differed between the sexes."

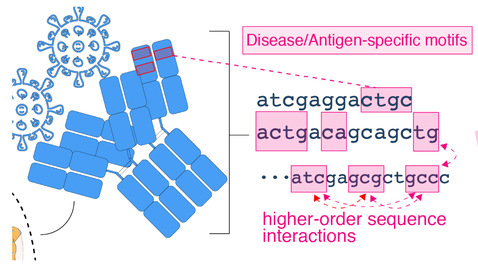
Age-Related Decrease in TCR Repertoire Diversity Measured with Deep and Normalized Sequence Profiling

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, Yuriy B. Lebedev and Dmitriy M. Chudakov

J Immunol March 15, 2014, 192 (6) 2689-2698; DOI: https://doi.org/10.4049/jimmunol.1302064



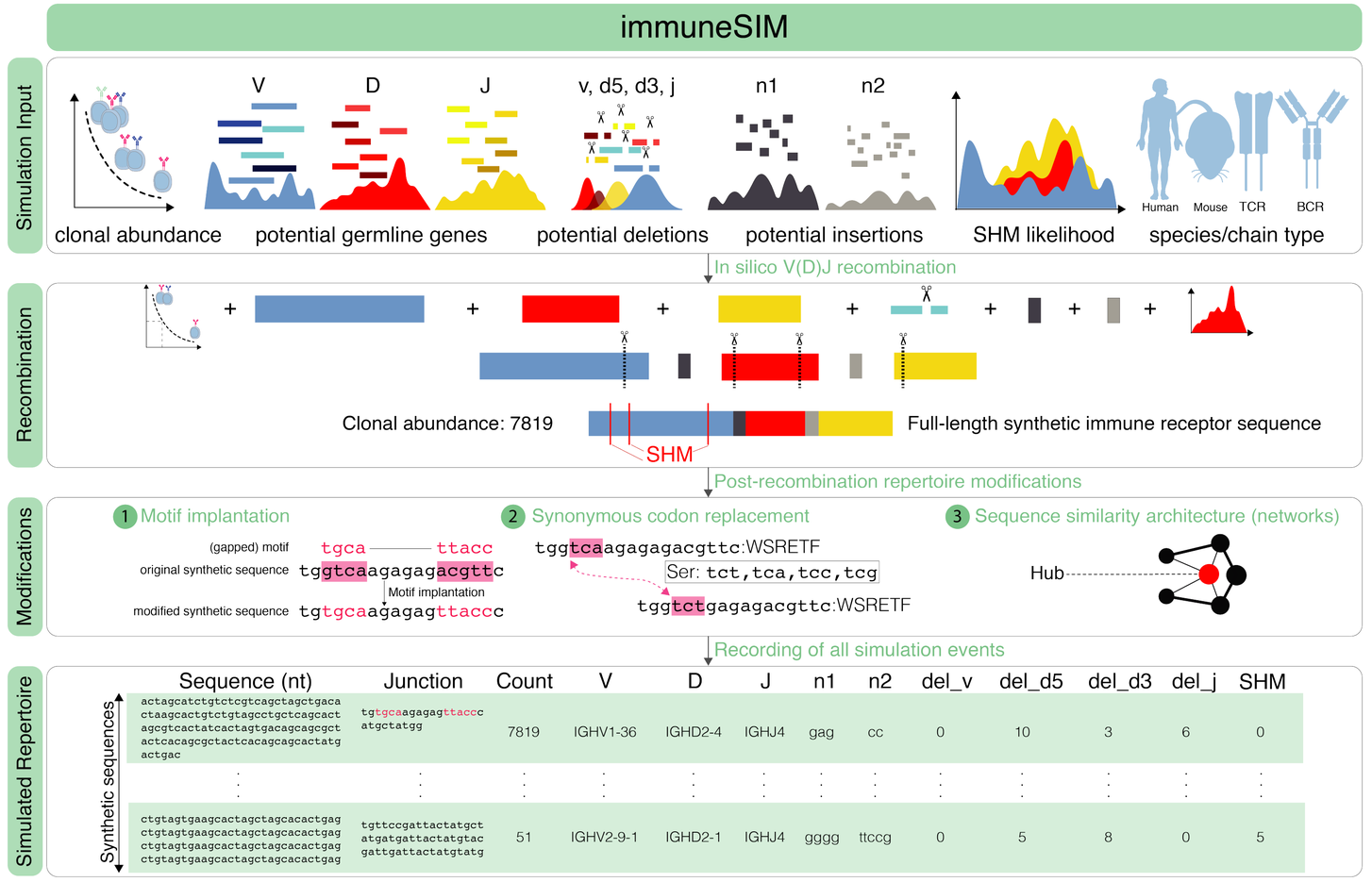
Generating immune repertoires with native-like immunosignature complexity for benchmarking machine learning approaches



Cédric Weber

Weber, Bioinformatics, 2020

Wout van Helvoirt



ImmunoProbs

BUILD PASSING | PYPI V0.2.0 | PYTHON 2.7 | LICENSE GPLV3

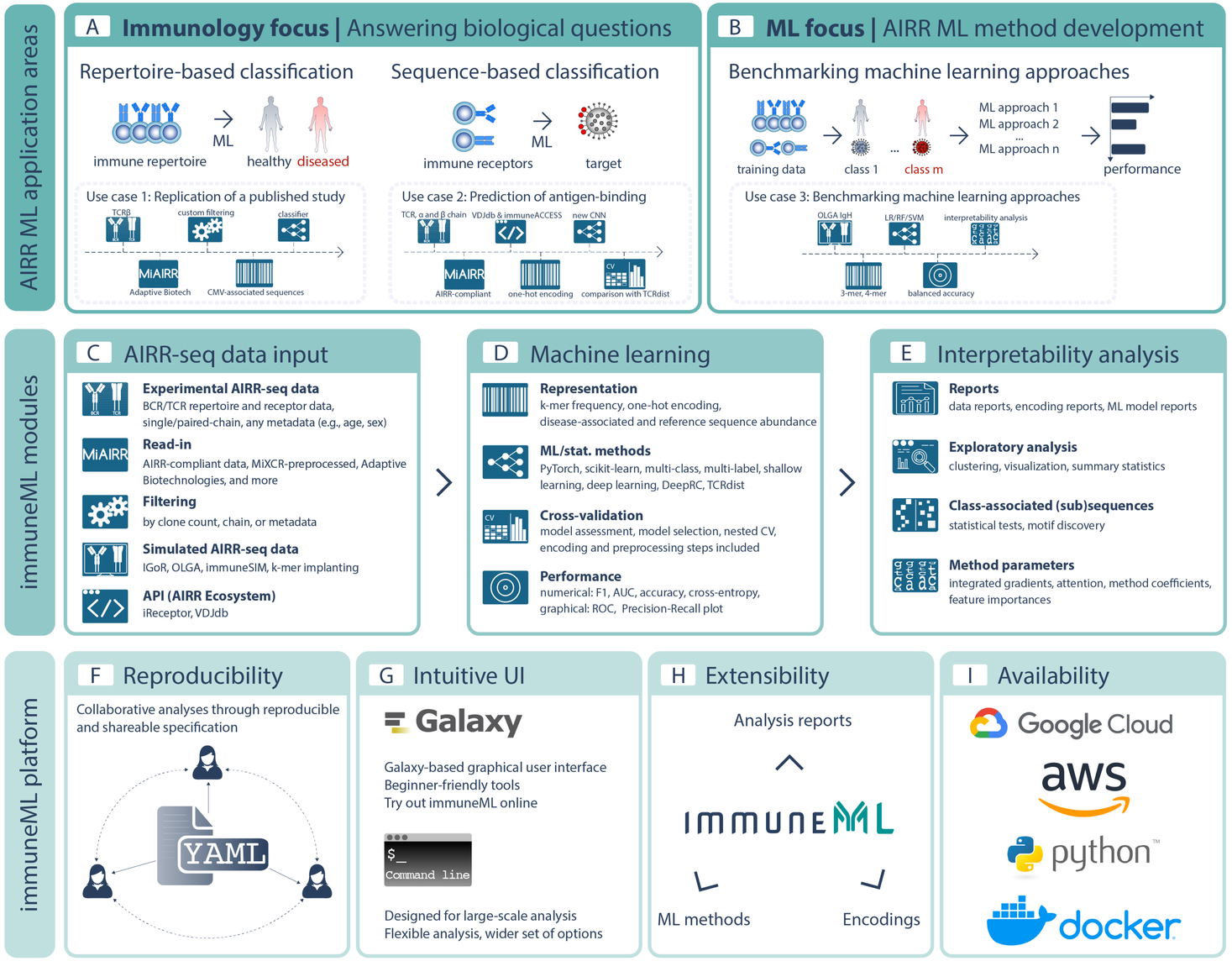
Create IGoR models and calculate the generation probability of V(D)J and CDR3 sequences.

For **installation**, **use cases** as well as a **tutorial**, please have a look at the ImmunoProbs documentation.

Development of a platform for AIRR machine learning

Current AIRR ML technical challenges:

- Without source code available, ML methodologies remain challenging to reproduce
- Currently researchers are developing their methodology from scratch: the code should be reusable
- The code should be flexible: it should be possible to study different data and different diseases, using the same or different models
- The structure of immune receptor data should be exploited for ML



immuneML

https://immuneml.uio.no

immuneML: an ecosystem for machine learning analysis of adaptive immune receptor repertoires

Milena Pavlović, Lonneke Scheffer, Keshav Motwani, Chakravarthi Kanduri, Radmila Kompova, Nikolay Vazov, Knut Waagan, Fabian L.M. Bernal, Alexandre Almeida Costa, Brian Corrie, Rahmad Akbar, Ghadi S. Al Hajj, Gabriel Balaban, Todd M. Brusko, Maria Chernigovskaya, Scott Christley, Lindsay G. Cowell, Robert Frank, Ivar Grytten, Sveinung Gundersen, Ingrid Hobæk Haff, Sepp Hochreiter, Eivind Hovig, Ping-Han Hsieh, Günter Klambauer, Marieke L. Kuijjer, Christin Lund-Andersen, Antonio Martini, Thomas Minotto, Johan Pensar, Knut Rand, Enrico Riccardi, Philippe A. Robert, Artur Rocha, Andrei Slabodkin, Igor Snapkov, Ludvig M. Sollid, Dmytro Titov, Cédric R. Weber, Michael Widrich, Gur Yaari, Victor Greiff, Geir Kjetil Sandve

doi: https://doi.org/10.1101/2021.03.08.433891

Pavlović and Scheffer, et al., *bioRxiv*, 2021

Necessary and interesting controls in (AIRR) machine learning

- Label shuffling
  - tests appropriateness of your class definition
  - Shuffle class labels x-times →prediction accuracy should decrease converging towards theoretical limit

- Randomize sequences
  - tests how much information is in sequence and sequence nt/aa composition
  - shuffle nt/aa order in sequences → prediction accuracy should decrease converging towards theoretical limit if sequence composition is similar between classes
- Equilibrate sequence length between classes
  - tests to what extent sequence length contributes to prediction accuracy

- Further controls: test effect of undersampling (real world robustness of classifier), evaluate feature recovery to inspect immunological meaning of classifier, hyperparameter optimization (for DL, random search might be more efficient than grid search), large enough test data sets, benchmark ML approach on simulated data where ground truth exists, balance by age, HLA, sex if possible

Summary: Classification, prediction and generation of AIRR data

- Machine learning (ML) is useful for classifying, predicting (diagnostics, repertoire-level) and generating immune repertoires (therapeutics, sequence-level)
- ML can act on entire immune receptor sequence or on subsequences (k-mers) thereof
- ML provides information on the extent to which immune repertoires capture immune information
- Measuring accuracy of ML remains a challenge, as is standardisation, reproducibility and generalizability
- Deep learning enables the capture of higher dimensional repertoire features. Its immunological interpretation, however, remains a challenge
- High-quality training and test/validation data for machine learning remains scarce. It remains also a question what are good training and test datasets
- Adjust for confounders and covariates

Brown et al., MSDE, 2019
Greiff et al., Curr Op in SysBio 2020

Future directions and Outstanding questions

Box 1. Future directions/major questions about repertoire dynamics.

Future directions

- Measurement of genetic variation in people and model organisms at B-cell receptor loci.
- Models of germinal centre dynamics that incorporate more types of data, such as B-cell receptor sequences, expression information [138], antigen availability and B-cell position.
- Phylodynamics models to evaluate spatial dynamics in germinal centres and statistical models of evolutionary descent.
- Improved models of B-cell memory formation and recall, especially those that infer the amount of competition between memory and naive responses for entry into germinal centres and between secreted antibodies and affinity-maturing B cells.
- Development of phylogenetic methodology specialized to the intricacies of B-cell receptor sequence evolution.
- Measurements of epitopes' relative immunogenicities across individuals.
- Between-species comparative analysis, especially for vaccine model organisms such as ferrets and macaques.
- Variation of B-cell response across human subpopulations, especially in response to shared exposures such as vaccines.
- Specific impacts of autoimmune checkpoints on the evolution of naive and experienced repertoires.
- Diversity and evolution of germline genes among vertebrates (i.e. evolution of presence-absence).
- Better understanding of the effects of age and co-infection, in particular, for autoimmunity and allergies.

Questions

- How can we approximate the genotype to phenotype map of B-cell receptors [139]?
- What are good models of sequence-based fitness landscapes for B-cell receptors? Are pairwise interactions between sites enough, as found by the Ising versus Potts analysis in Mann *et al.* [140]?
- How does T cell help impact the general dynamics of affinity maturation and the selective pressures on specific clones?
- How do the general dynamics of affinity maturation differ between individuals and change with age?
- When two genetically identical and naive hosts are immunized to the same antigen, how do their repertoires differ genetically and phenotypically? How would differences in their naive repertoires, chance recruitment of naive B cells to the response, stochastic dynamics of affinity maturation and other factors contribute?
- Can we use immune information to infer asymptomatic infections?
- Can we relate sequences from sampled repertoires to protection?
- Can we use germline gene loci or a sample of the naive repertoire to predict an individual's responsiveness to a vaccine [141]?
- How is vaccine responsiveness affected by immune memory to other antigens?
- Can immune systems across individuals be classified into meaningful types, and can we use immune 'type' information for stratified sampling in clinical trials?
- Holding infection history constant, are differences in B-cell repertoires important for pathogen evolution [142]?

Cobey, Philo Trans B, 2015

Outstanding Questions

How large of an effect does IG polymorphism have on the development of the baseline naïve repertoire, and what types of genetic variation (CNV, coding variants, regulatory variants) matter most?

Do effects of IG genetic variants on the Ab repertoire correspond to known biases in disease and/or clinically relevant Ab responses?

Wardemann, Trends Imm, 2017

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.

Greiff, Trends Imm, 2015

How to compare the repertoires of different donors? The repertoire is shaped by multiple components (e.g., heredity, historic exposure, current exposure), so how can the noise in interindividual comparisons be reduced? Does this require normalization against the mature naïve B cell compartment?

Can antibody reactivity be predicted from sequence data? Although NGS offers unparalleled throughput it does not provide any affinity data and current recombinant expression techniques do not (yet) deliver the throughput required for large-scale screening of antibodies. While *in silico* models are often presented as an alternative, they are computationally expensive themselves but might be up for the task in the near future.

Watson, Trends Imm, 2017

Personal view: outstanding questions

- Standardization of experimental protocols
- Methods for large-scale generation of antigen-annotated AIRR data
- Merging sequence analysis with structural modelling at repertoire scale
- Improve proteomic understanding of the antibody repertoire
- Methods for analysing paired chain data
- Interpretability of machine learning approaches
- Structure of antigen-specific motifs implicated in the prediction of antigen binding and immune status

Acknowledgements

University of Oslo

Prof. Geir Kjetil Sandve
Prof. Dag T.T. Haug
Prof. Ingrid H. Haff
Prof. Ludvig Sollid
Prof. Torbjørn Rognes
Dr. Fridtjof Lund-Johansen
Dr. Rahmad Akbar
Dr. Igor Snapkov
Dr. Philippe Robert
Dr. Chakravarthi Kanduri
Dr. Ivar Grytten
Dr. Knut Rand
Dr. Gabriel Balaban
Dr. Enrico Riccardi
Dr. Mai Ha Vu
Lonneke Scheffer
Milena Pavlović
Andrei Slabodkin
Maria Chernigovskaya
Ghadi Al Hajj
Robert Frank
Thomas Minotto
Habib Bashour
Khang Le Quy

University of Florida

Keshav Motwani
Prof. Todd Brusko

Uni Linz
Prof. Günter Klambauer
Prof. Sepp Hochreiter

ETHZ Zürich

Dr. Cédric Weber
Dr. Alexander Yermanos
Prof. Sai T. Reddy

Uni Bern

Prof Andrew Macpherson
Dr. Julien Limenitakis

UCSD

Dr. Yana Safonova
Prof. Pavel Pevzner

BGI/Tsinghua-Shenzhen

Xiao Liu
Wei Zhang
Longlong Wang
Jinghua Wu
Ziyun Wan
Shiyu Wang
Kai Gao

iReceptor+

Prof. Gur Yaari
Prof. Lindsay Cowell
Dr. Scott Christley
Dr. Artur Rocha
Alexandre Almeida Costa

FHNW

Enkelejda Miho

JHU

Dr. Jeliazko Jeliazkov
Prof. Jeffrey Gray

Funding

- UiO World leading research community
- UiO Life Science
- UiO immunoHUB
- Horizon2020
- Norwegian Research Council
- The Helmsley Charitable Trust
- Norwegian Cancer Society