Quality Control Pipelines for T cell and B cell AIRR-seq

Encarnita Mariotti-Ferrandiz
Sorbonne Université
Laboratoire i3 – UMRS959
Pitié-Salpêtrière, Paris
encarnita.mariotti@sorbonne-universite.fr

Nina Luning Prak
Human Immunology Core
Department of Pathology and Laboratory Medicine
Perelman School of Medicine
University of Pennsylvania, Philadelphia
luning@pennmedicine.upenn.edu

Outline

- Speaker introductions
- Brief introduction to Adaptive Immune Receptor Repertoire (AIRR) analysis
  - V(D)J recombination
  - Technologies
- AIRR-seq QC
  - Overview of complex sequencing workflows- need for QC
  - Sample and target amplification
  - Library prep
  - Sequencing run
  - Within sample metadata and replicate analysis
  - Between sample and sequencing run QC
  - QC platform
- Benchmarking studies and ongoing AIRR-C initiatives
- Question and Answer session

Nina

Brief overview of V(D)J recombination and immune repertoire diversification

Encarnita

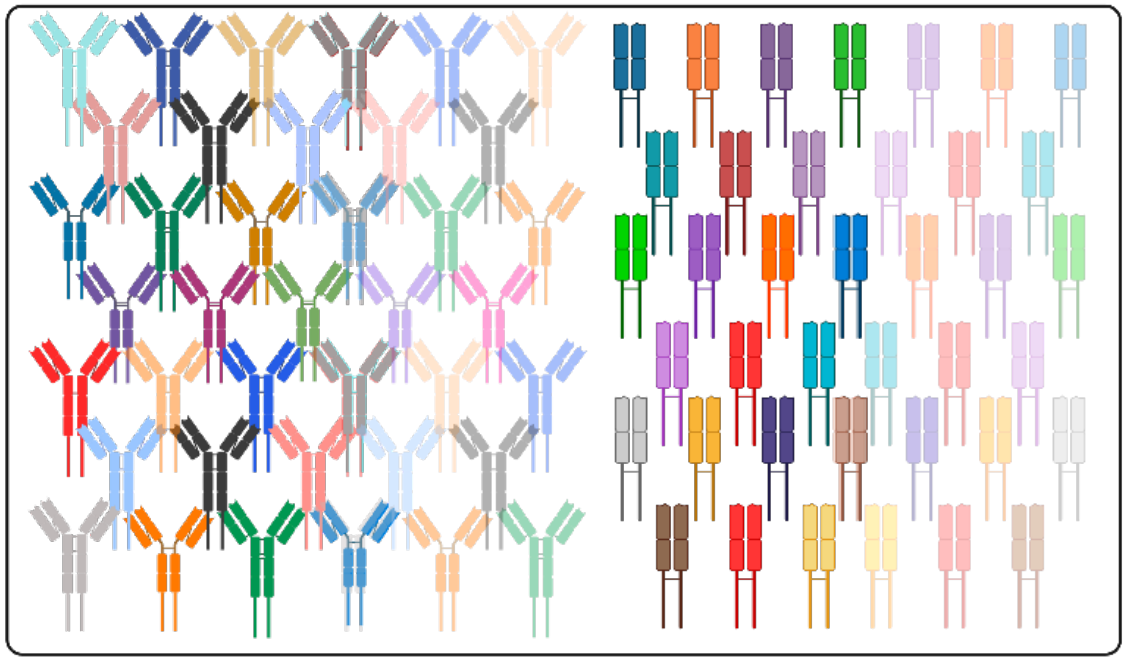
ANTI
BODY
SOCI
.ETY

AIRR
Community

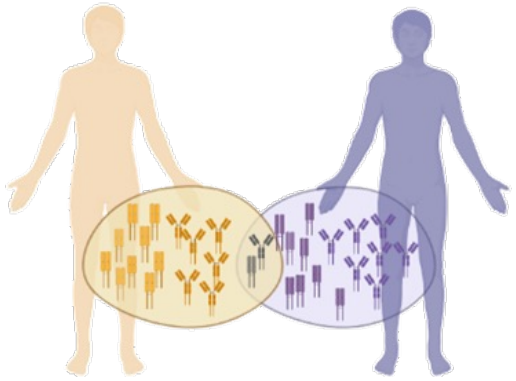
The immune repertoire, or collection of different antibody or T cell receptor rearrangements, is very large

Adaptive immune receptor repertoires (AIRR)

TCR and BCR repertoires are highly complex



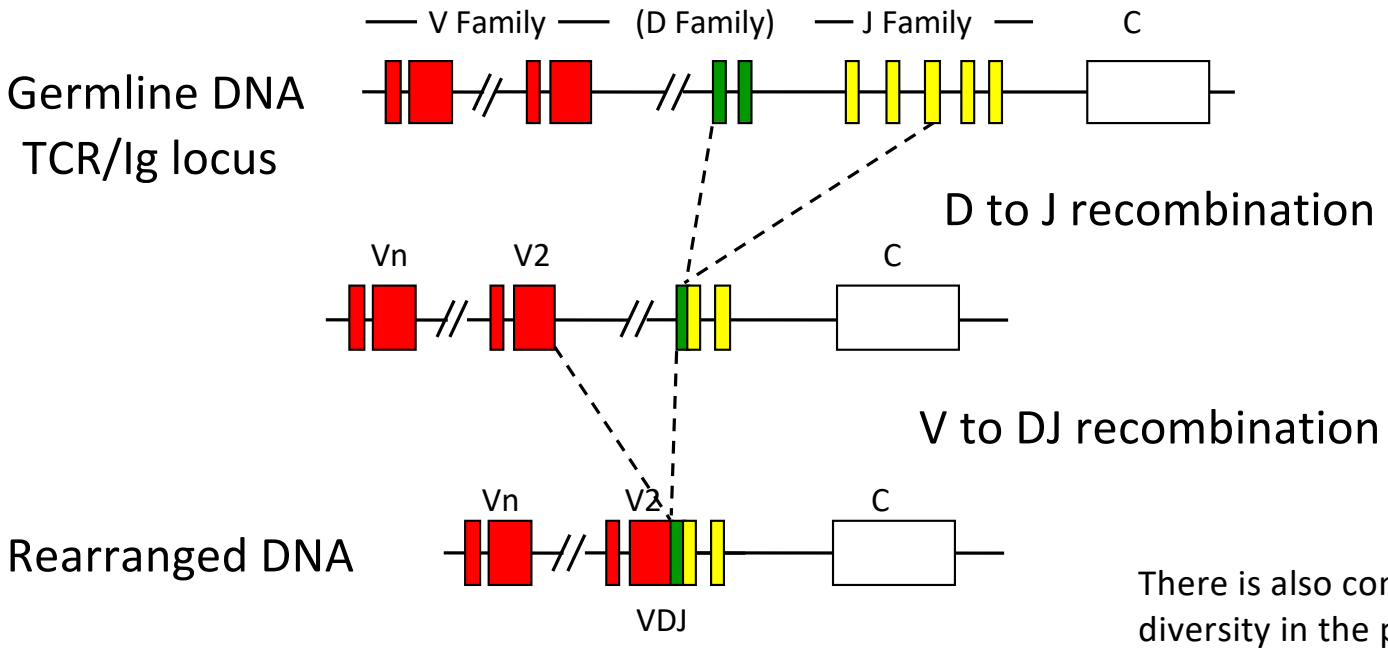
10^{19} TCRs & 10^{13} BCRs



and reflect individual immunological history

Six, Mariotti-Ferrandiz, 2013; Greiff et al, 2017; Dupic et al, 2019; Bradley & Thomas, 2019

Combinatorial diversity of the AIRR – somatic recombination of Variable, Diversity and Joining gene segments



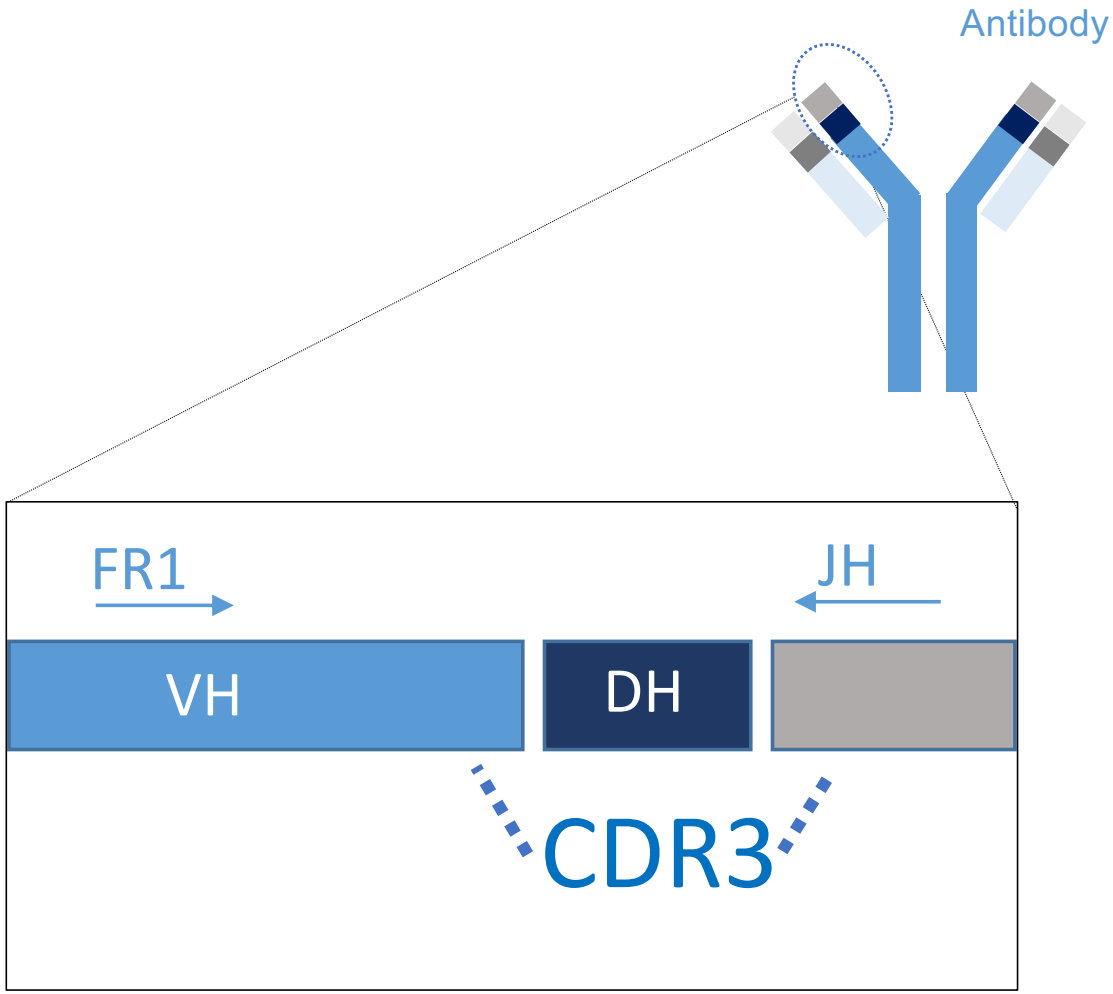
There is also combinatorial diversity in the pairing of different heavy chains and light chains

CDR3 = V(D)J junction

→ Contact with antigen for Ig

→ Contact with MHC-restricted peptide for TCR

A word about clones…



Nomenclature

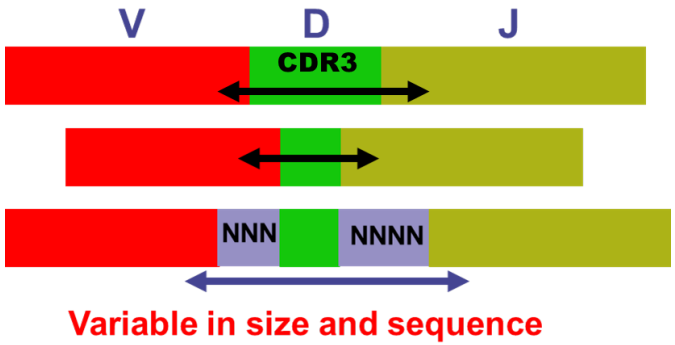
- CDR3 = third complementarity determining region
- Hypervariable sequence within the antibody heavy chain DNA sequence
- This region is used as a clonal fingerprint because two unrelated B cells (or T cells) are very unlikely to share the same sequence.

The third complementarity determining region (CDR3) has very high sequence diversity

- Combinatorial diversity = combination of V(D)J segments
- Junctional diversity = P-additions (opening of hairpin), random deletion (exonuclease), N-additions (TdT)

→The CDR3 region is thus variable in sequence and length

CDR3 = *fingerprint of the rearrangement*



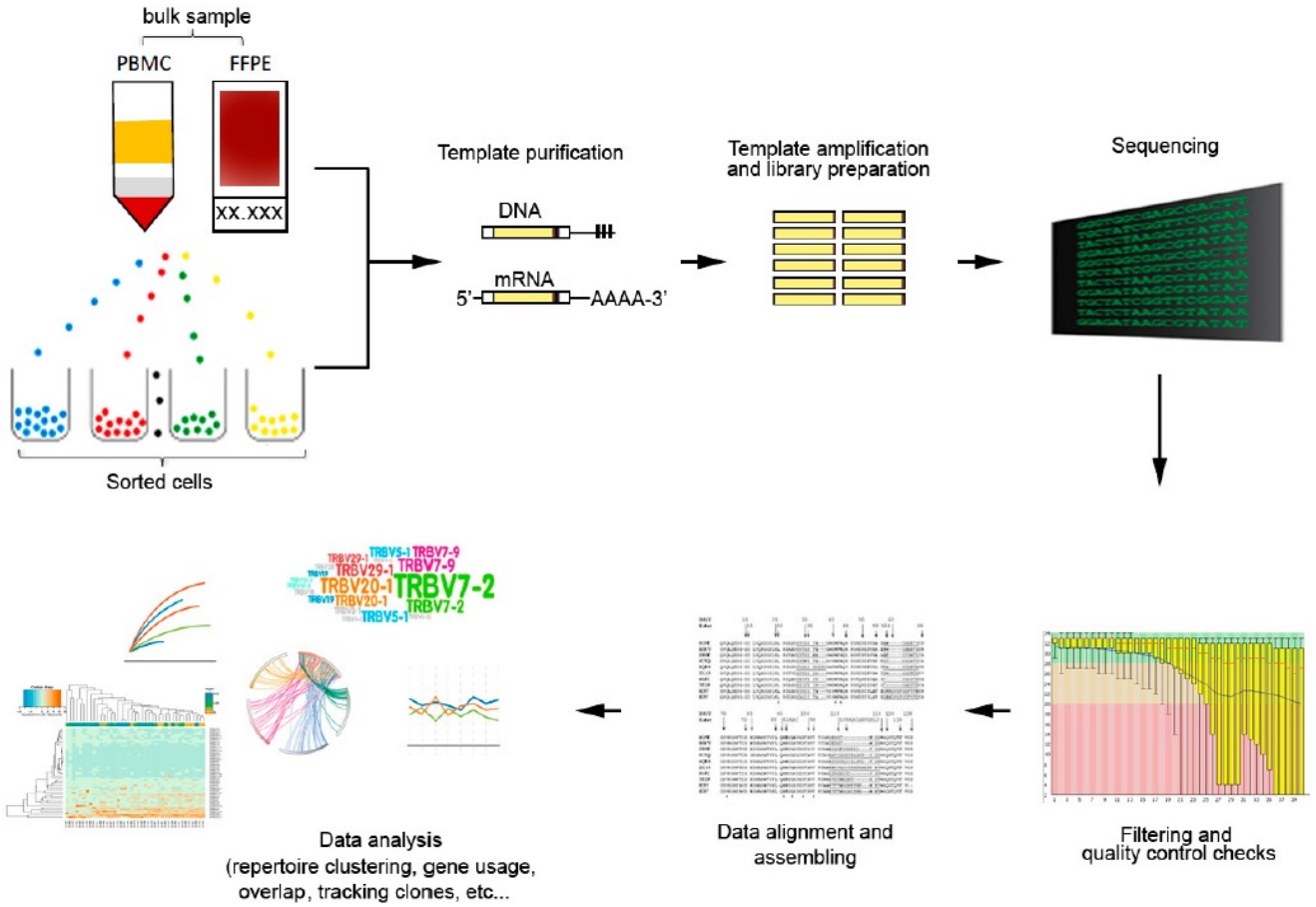
A word about repertoire analysis technology

Encarnita and Nina

ANTIBODY SOCIETY

AIRR Community

Next generation sequencing of AIRR (AIRR-seq) principle



Aversa et al, 2020, International Journal of Molecular Science

Overview of Different Methods for Immune Repertoire Profiling

Method	Questions	Advantages	Disadvantages
DNA sequencing from cells or tissues (bulk)	Repertoire diversity Identifying clonal expansions Tracking clones (ex. MRD)	High throughput (one template per cell) High sensitivity Simplest workflow (clinical use) Relatively inexpensive	Can't clone full antibody PCR bias Amplicon length
Bulk RNA sequencing, with molecular barcodes (UMI)	V gene usage Analysis of clonal evolution (study of somatic mutations within clonal lineages) Antibody heavy chain isotype information	High fidelity Moderate throughput useful for analysis of clonal lineages Minimal PCR bias	Skewed by RNA transcript abundance differences, usually best to couple with cell sorting Moderate cost
Single cell RNA sequencing	Link antibody (H+L) to single cell phenotype, transcriptional profile or antigen specificity	High fidelity (molecular barcoding) Can clone full paired antibody sequence	Expensive (kit and sequencing costs), but getting cheaper

Yaari and Kleinstein. Genome Medicine (2015)
Georgiou et al. Nature Biotechnology (2014)
Papalexi et al. Nat. Rev. Immunol. (2018)

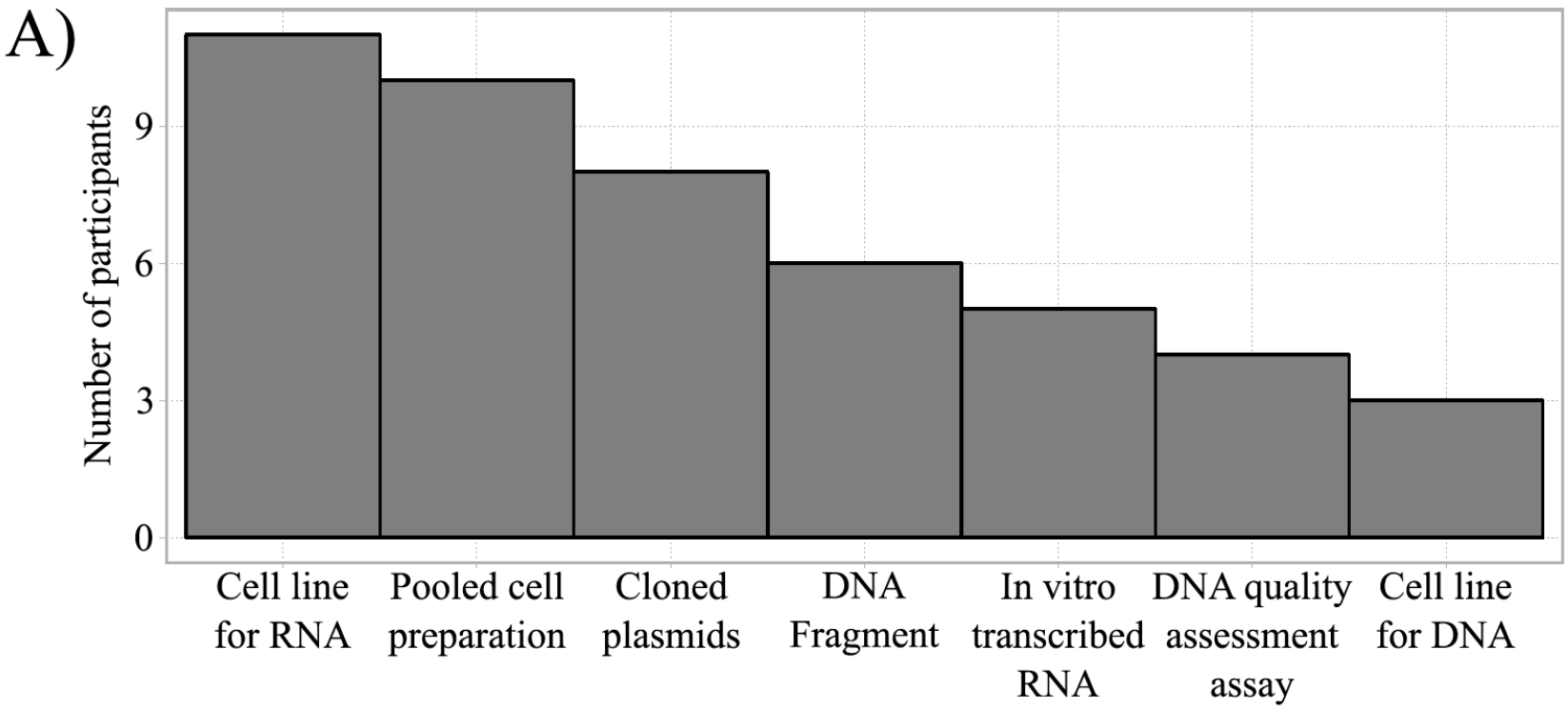
Different methods have different QC challenges

		Bulk gDNA sequencing	Bulk cDNA sequencing	Single-cell cDNA sequencing
Potential Issues	V-gene amplification bias	++	+	+/-
	V-gene annotation issues	++	+	+
	PCR and sequencing error	++	+	+/-
	Difficulty with translation of copy number to cells	+/-	++	+/-
	Degradation of template	+	++	++

Truck et al. eLife 2021

Nina

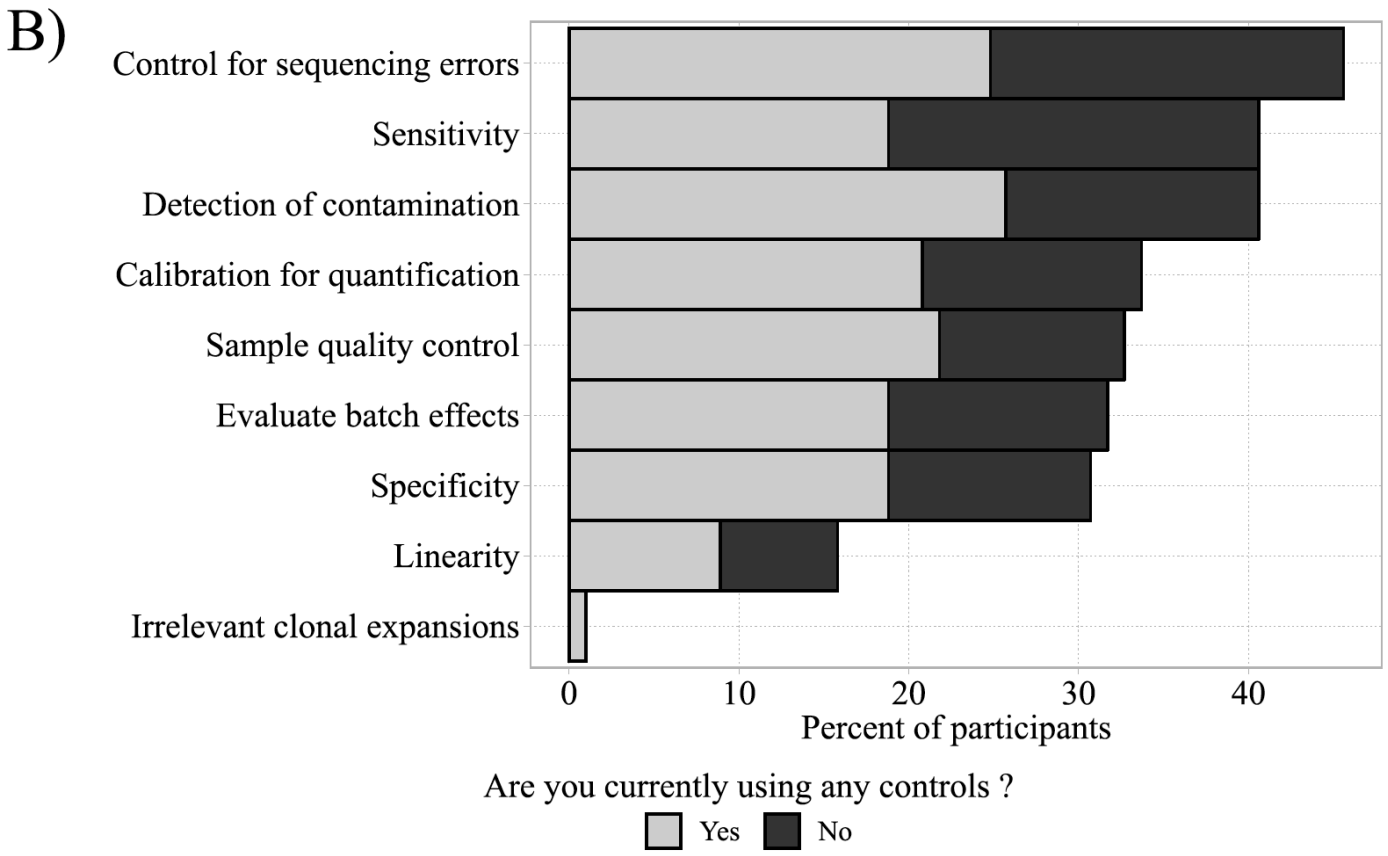
AIRR-C quality controls needed by research community



Truck et al. 2021 eLife

Nina

AIRR-C quality controls needed by research community

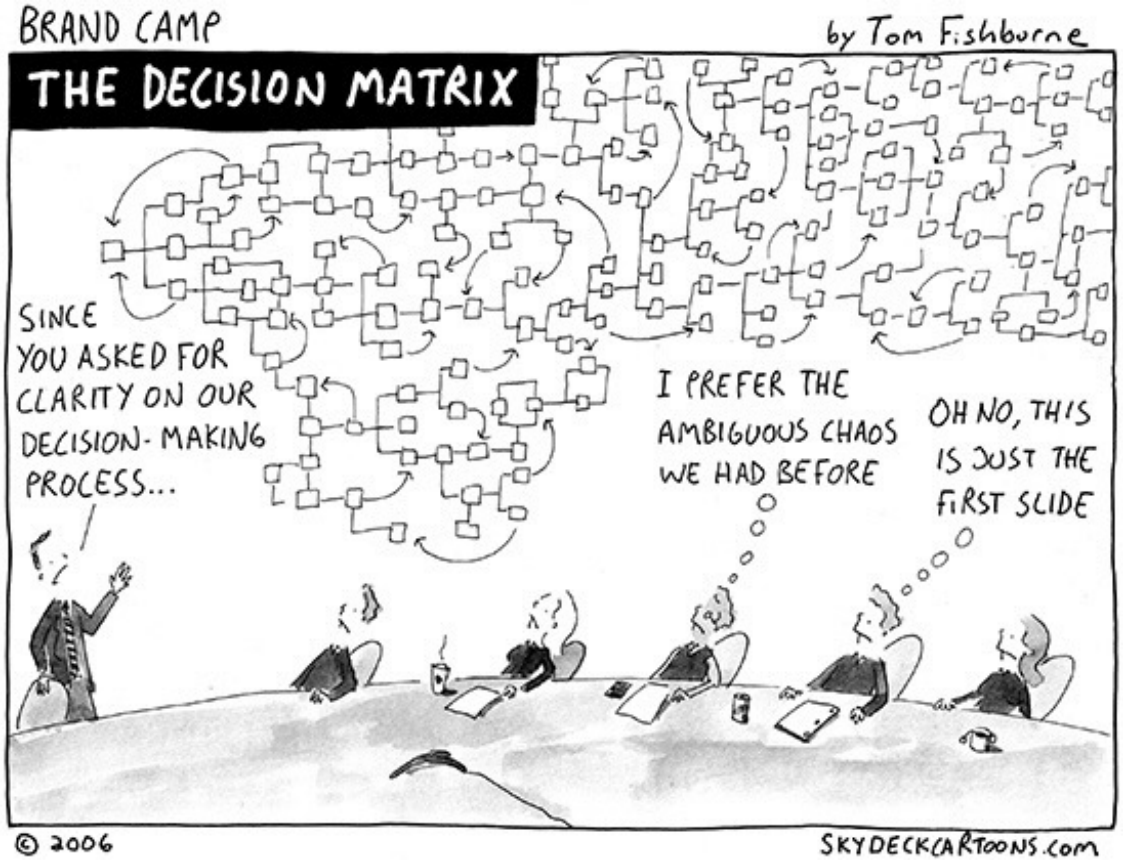


Truck et al. 2021 eLife

Nina

Need for QC: complex workflows

Nina

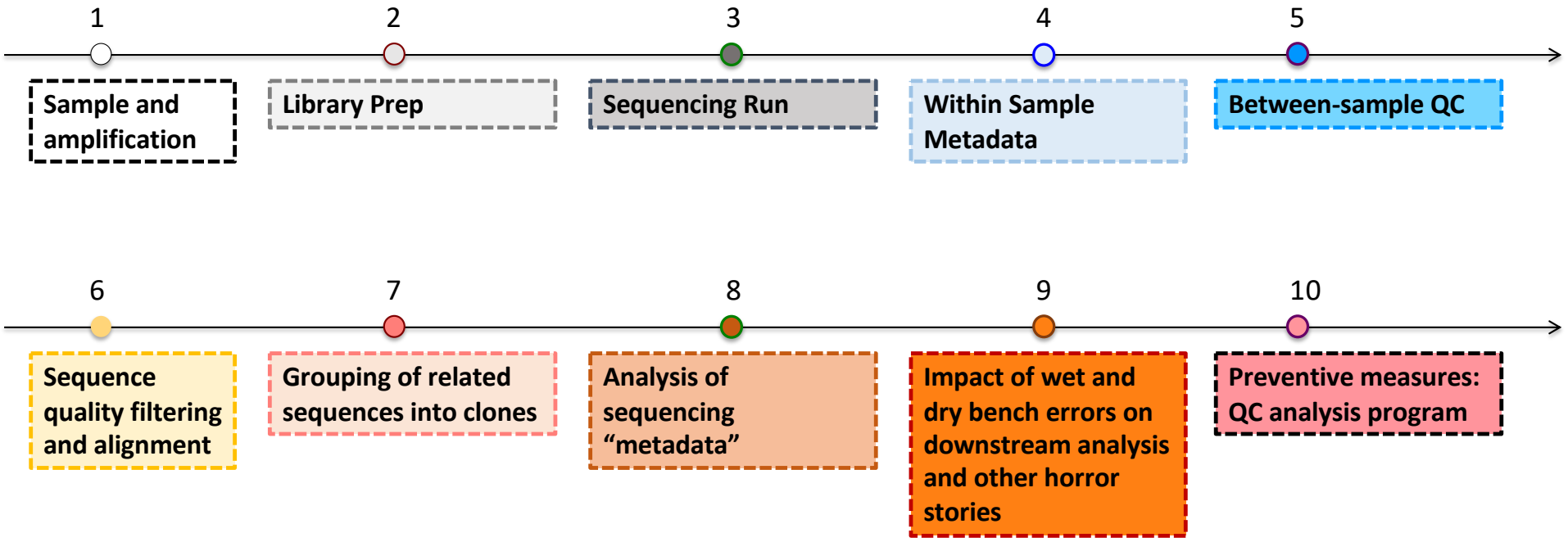


Why are workflows helpful?

Workflow with detailed procedures reduces risk of missing costly errors in real time.

Systematically checking for errors may reveal issues with the data that are not immediately obvious if you only analyze the data at the back end.

Overview of AIRR-seq workflow



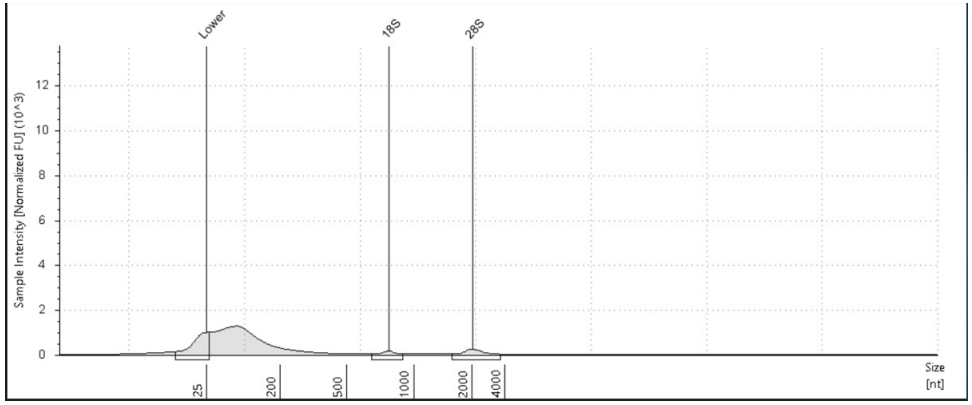
Long, serial workflows are prone to cumulative error

Need real-time monitoring of results to make sure that the results and analysis are not flawed

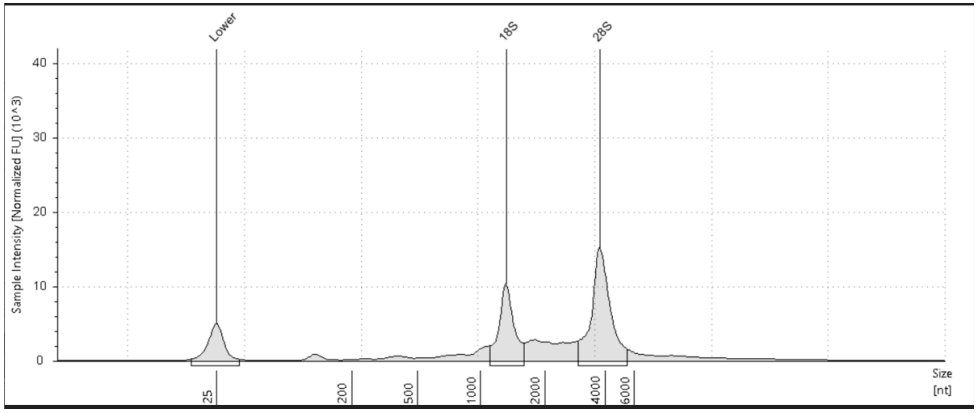
Sample QC and target amplification

Sample QC:

RNA QC by Bioanalyzer/TapeStation
Quantity by Qubit



Poor quality RNA prep
ACTION: STOP, need to re-extract and/or get new samples!

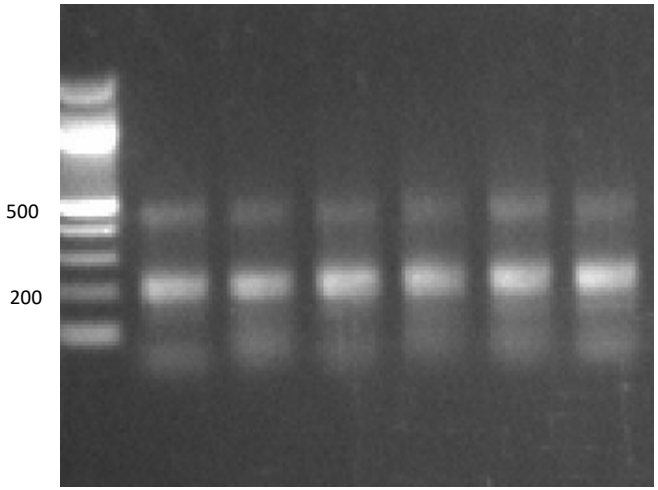


Quality assessed with a RNA integrity score (RIN for Agilent, RNA IQ on Qubit) : 0-10 scale, > 7 is already good

Amplification QC:

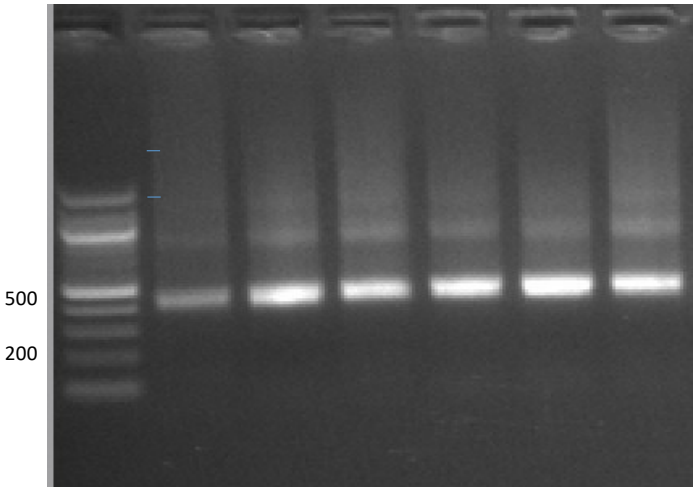
Agarose gel electrophoresis- may inform downstream purification steps

Fail



Either DNA with poor quality (FFPE) or very few B/T cells in the original samples.
ACTION: STOP! If samples are replaceable, start over. If samples are irreplaceable, need to do extra purification to remove ~200bp band and potentially modify input amount in library PCR etc.

Pass



Controls

Positive controls:

- Spleen
- Apheresis or pooled PBMC sample
- Cell line mixture
- Human/mouse DNA mixture

Experimental controls:

- Sorted follicular B cells
- Sorted class switched cells if modeling SHM
- Cell line spiked into other cells for sensitivity

Negative controls:

- Fibroblast
- Antigen negative sort
- Wild type mouse

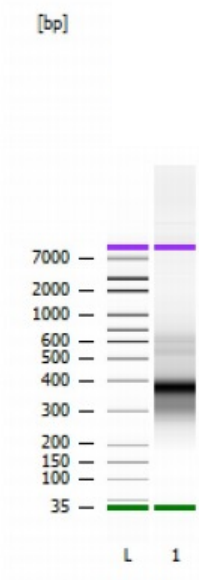
Nina

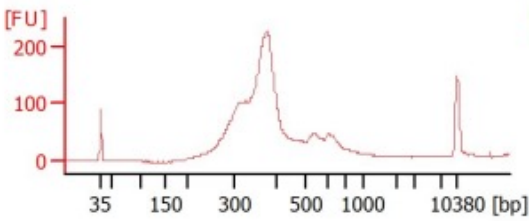
Library preparation and sequencing

Library Prep QC:

Bioanalyzer trace for library quality.
KAPA quantification RT-PCR for library quantity

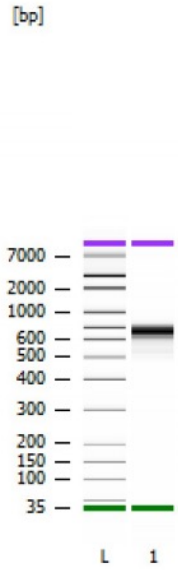
Fail

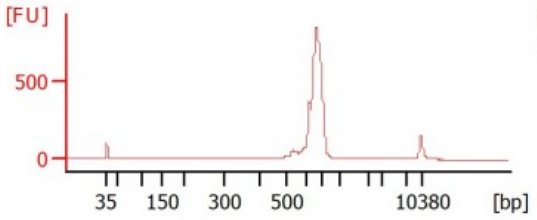




Can still work with this, but many of the reads are going to be off-target. If sample is abundant, consider re-purifying.

Pass



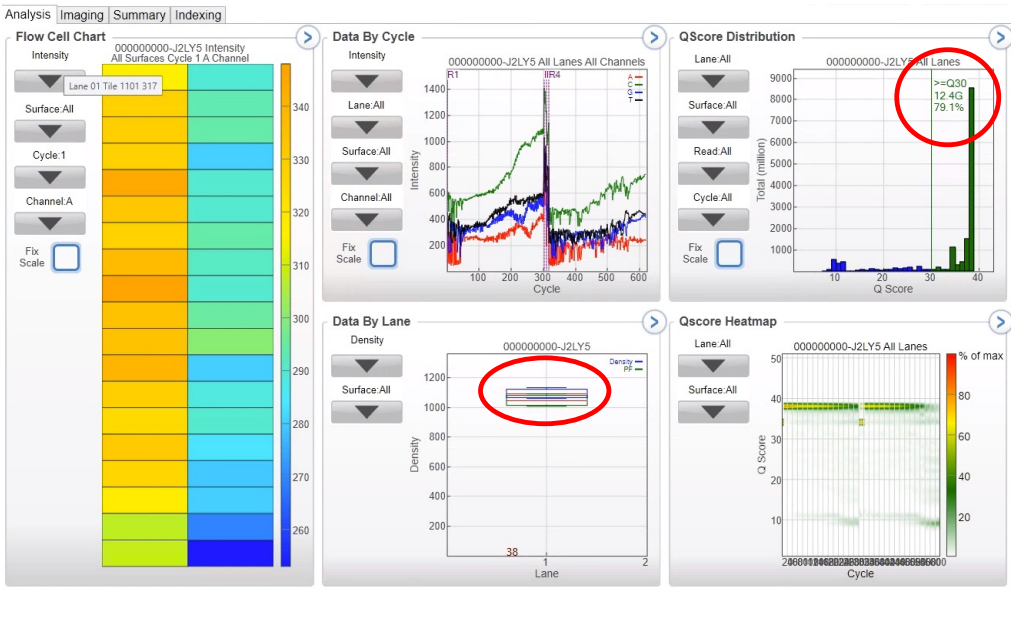


Sequencing Run QC:

Clustering density, data quality heatmap, % data pass Q30.

Fail

Pass



Can analyze higher quality sequences, but preferable to re-run the samples. May need to adjust sample input.

Within sample and replicate analysis

Within Sample QC: compare different biological replicates from the same individual

Consistency of Total reads, Valid reads, Unique sequences, Clone counts between replicates

General properties: avg CDR3 length, avg V gene identity, fraction of productive rearrangements

Fail

Replicate #	ng input	Uniques	Copies	Avg. CDR3 Length	Avg. V-identity	In-frame Fraction	Clones
1	100	4	4	53.25	0.943125	1	4
2	100	370	5001	49.8333333	0.95103333	0.83333333	6
3	100	11	49	47.1428571	0.90651429	1	7
4	100	6783	12466	48.3152542	0.93613627	0.98983051	1180
5	100	5953	10417	47.9861933	0.93667367	0.98619329	1014

Clone counts very different between replicates

Pass

Replicate #	ng input	Uniques	Copies	Avg. CDR3 Length	Avg. V-identity	In-frame Fraction	Clones
1	100	8709	29247	48.0308019	0.94436341	0.99309612	1883
2	100	7236	15517	48.195738	0.94572226	0.99368587	1267
3	100	7242	24432	47.8988219	0.94479681	0.99099099	1443
4	100	7811	16994	47.6504928	0.94459067	0.9939348	1319
5	100	8574	30837	48.0735826	0.94508239	0.99155609	1658

Clone counts are similar for a given ng input

Replicate #	ng input	Uniques	Copies	Clones
1	50	1867	61784	342
2	50	11025	41201	3951
3	50	13165	48191	3879
4	50	14090	18264	2349
5	50	13030	52075	3814

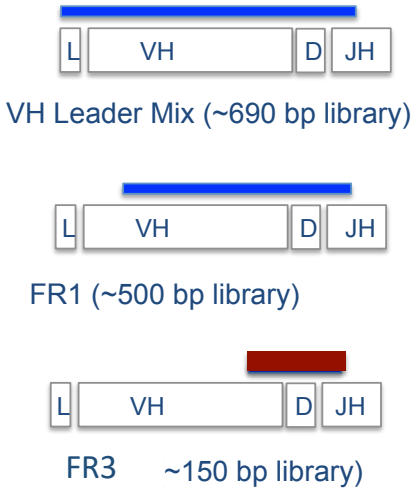
Copy counts are similar, but unique & clone counts differ

Replicate #	ng input	Uniques	Copies	Clones
1	50	15586	72791	3514
2	50	13341	53631	3936
3	50	16524	73691	3886
4	50	17719	26112	3421
5	50	14744	66146	3525

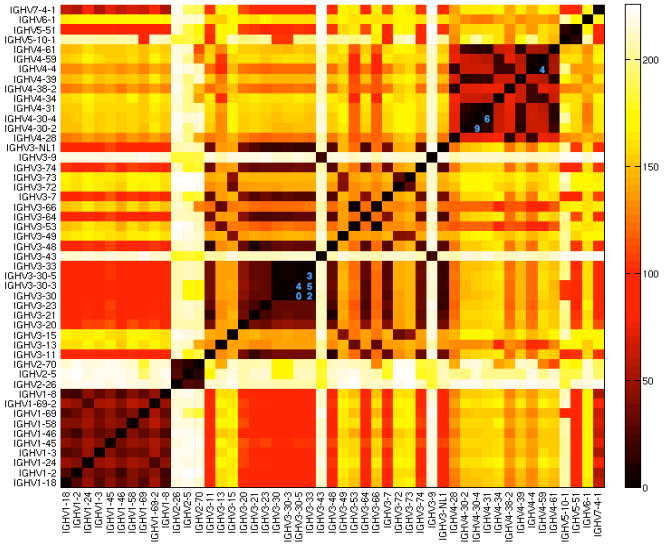
Unique & copy counts are similar

Sequence length matters: VH classification lacks resolution with short reads

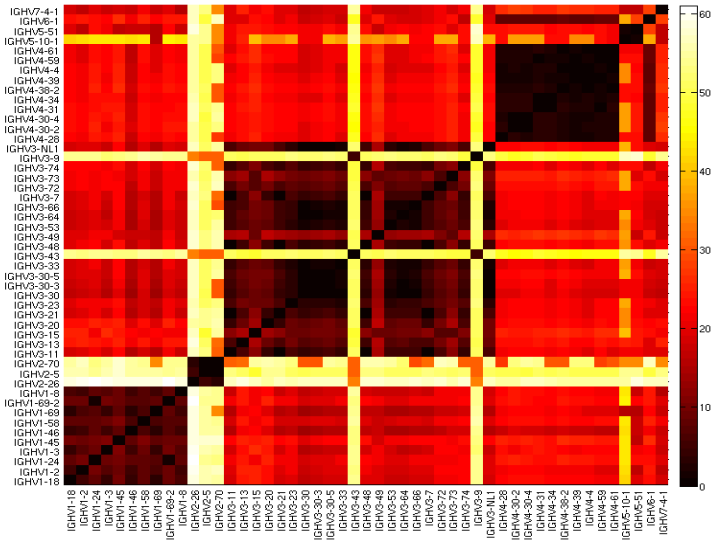
Germline VH gene similarity matrix



V gene similarity at > 150 bp



V gene similarity at 75 bp



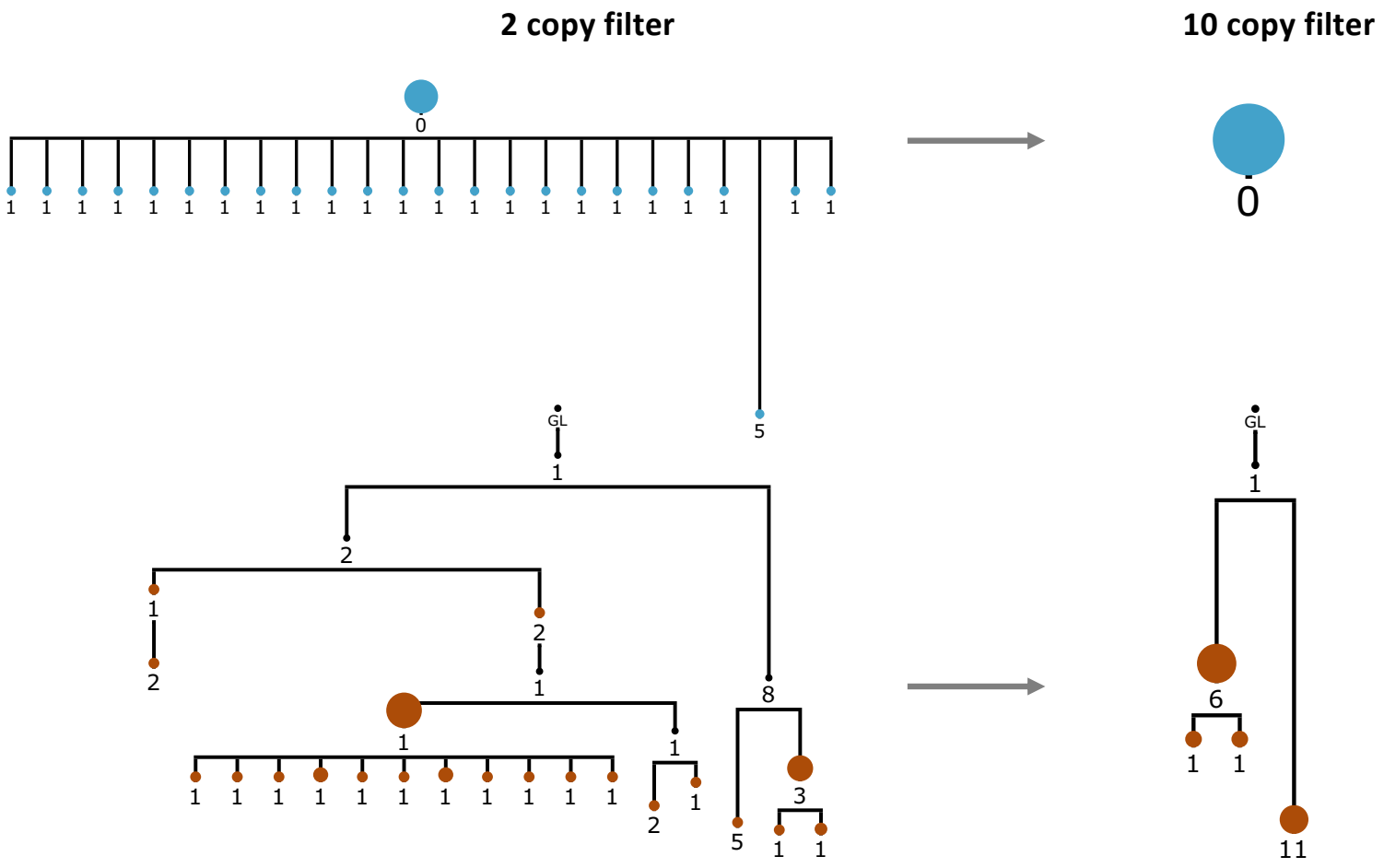
Boxes in brown are difficult to tell apart. At short read lengths, one basically has VH gene family resolution, which is sub-optimal for clonal lineage definition.

Zhang et al. J. Immunol. Methods 2015

Sequencing errors

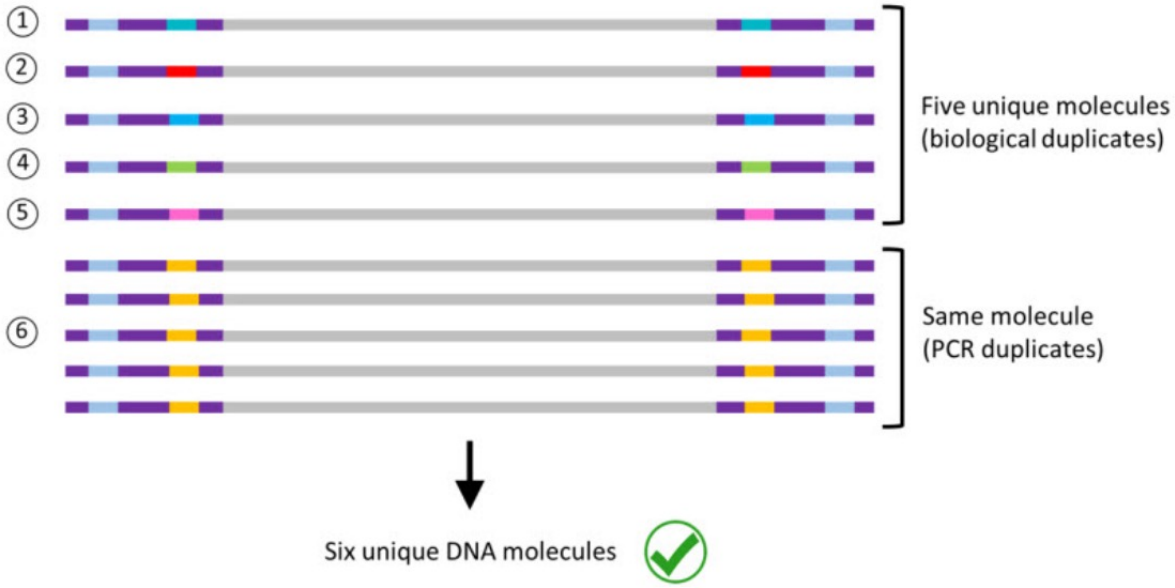
- Varies by platform, depth of sequencing
- Signs:
  - bulk gDNA unexpectedly low fraction of productive rearrangements
  - High levels of SHM (unexpected)
  - Nodes in lineage trees with single base errors
- Mitigation strategies
  - Filter data- copy number, copy fraction, number of unique sequences in replicate amplifications etc.
  - Unique molecular identifiers- create a consensus sequence for sequences that share the same molecular barcode

Effects of sequencing errors on clonal lineage analysis



UMIs- unique molecular identifiers for bulk RNA sequencing

Step 1, tag unique molecules with barcoded adapters at the cDNA synthesis step



For more detailed description of unique molecular identifiers and other Ig-seq QC, see Khan et al. Science 2016

Image from Takara Bio ThruPLEX Tag-seq

UMIs- unique molecular identifiers for bulk RNA sequencing

Step 2, error correction- generate a consensus sequence for sequences that share the same tag

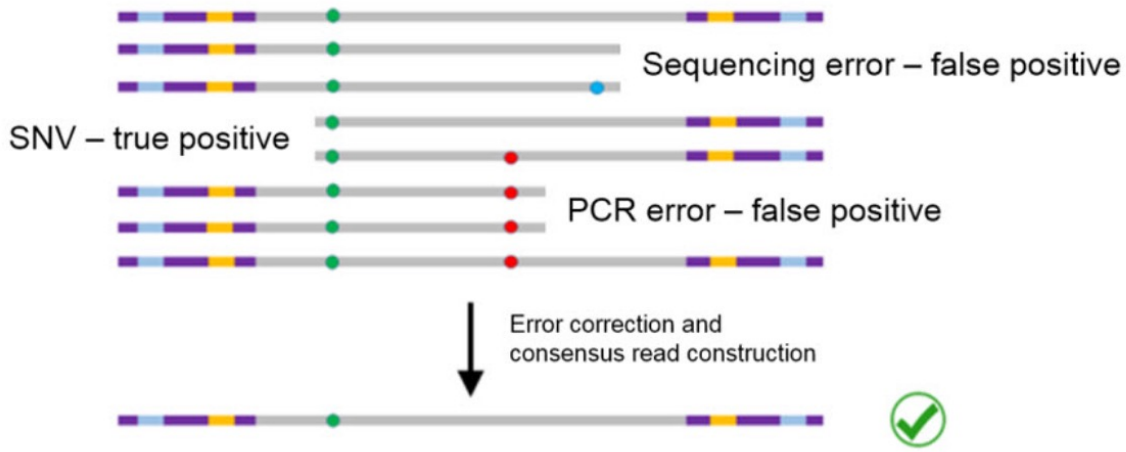
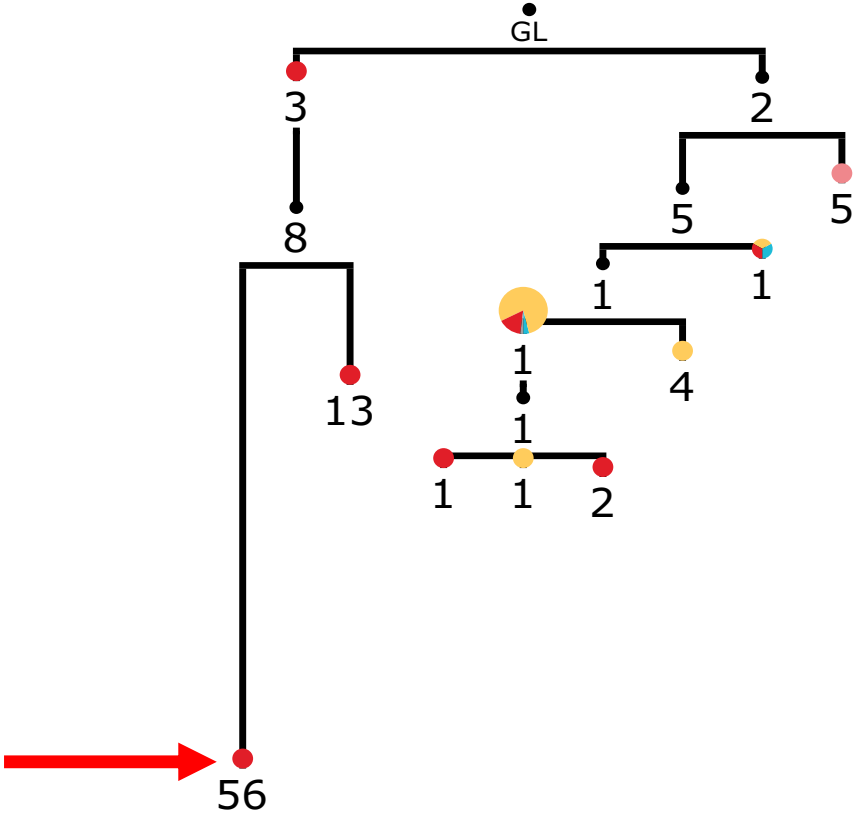


Image from Takara Bio
ThruPLEX Tag-seq

https://www.takarabio.com/about/bioview-blog/tips-and-troubleshooting/using-umis-in-ngs-experiments

Other unexpected findings

Example of a lineage tree with a distant node



What is going on?

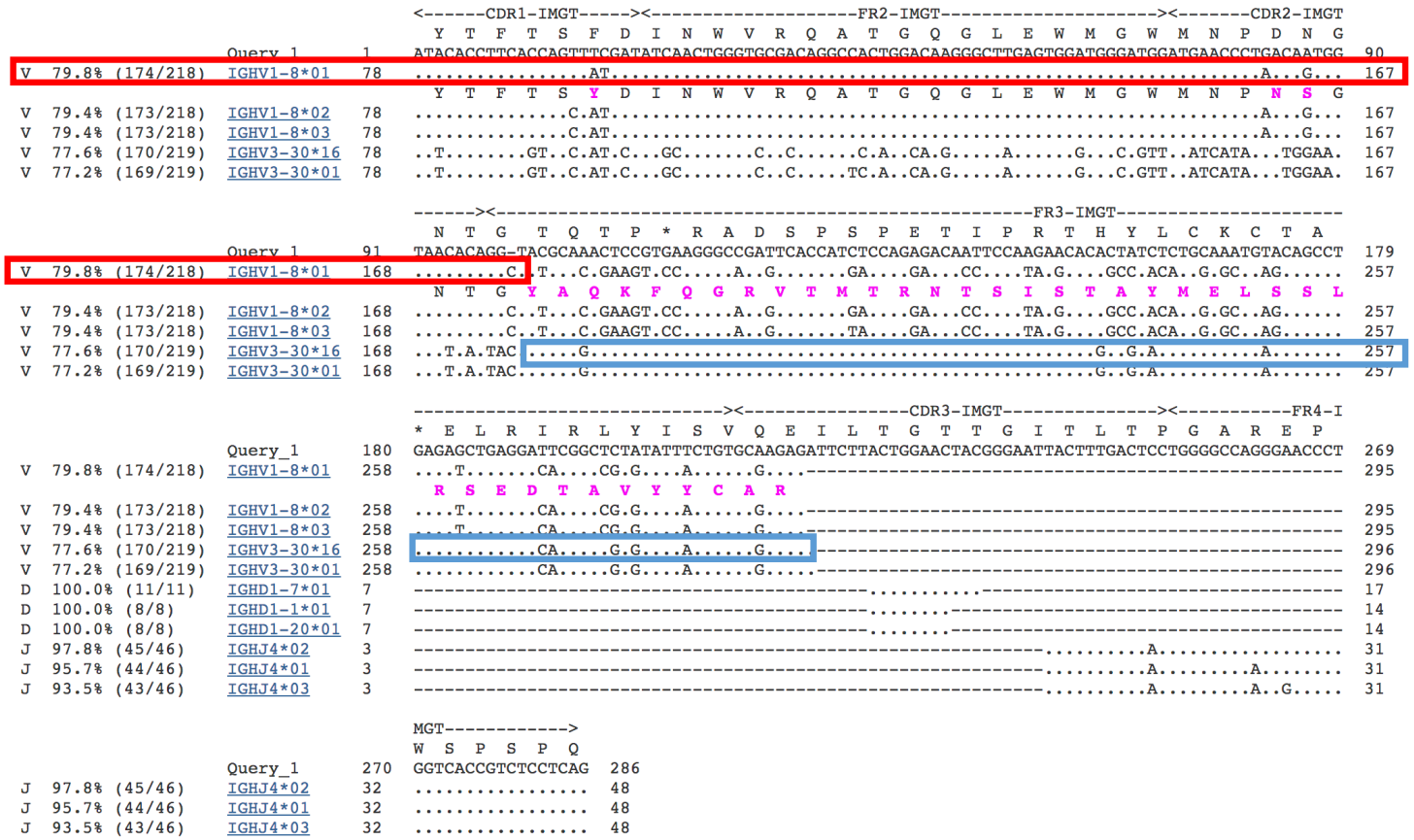
- incorrect clone collapsing algorithm?
- hybrid PCR product
- Prolonged antigen selection- but only after technical explanations have been ruled out

Hybrid PCR Products

Hybrid products can be found by aligning low-identity reads

- This sequence had low V-identity to the closest known germline (< 80%)
- Upon investigation it appears to be a hybrid PCR product of two templates, one with IGHV1-8 and the other with IGHV3-30
  - A near perfect alignment of 1-8 is shown in red for the 5’ end of the read
  - An equally high-scoring alignment of 3-30 is shown in blue on the 3’ end

Test in separate PCR amplifications, switch primer sets

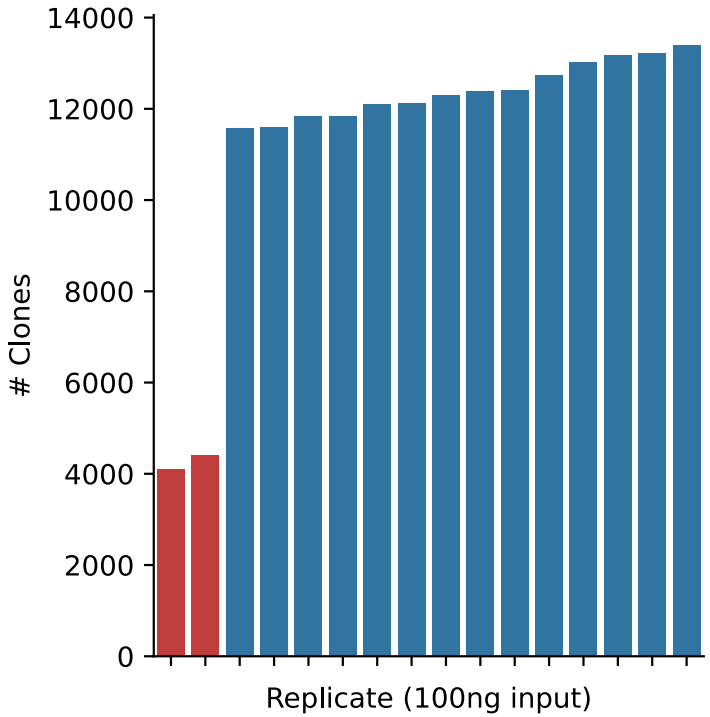


Amplification bias

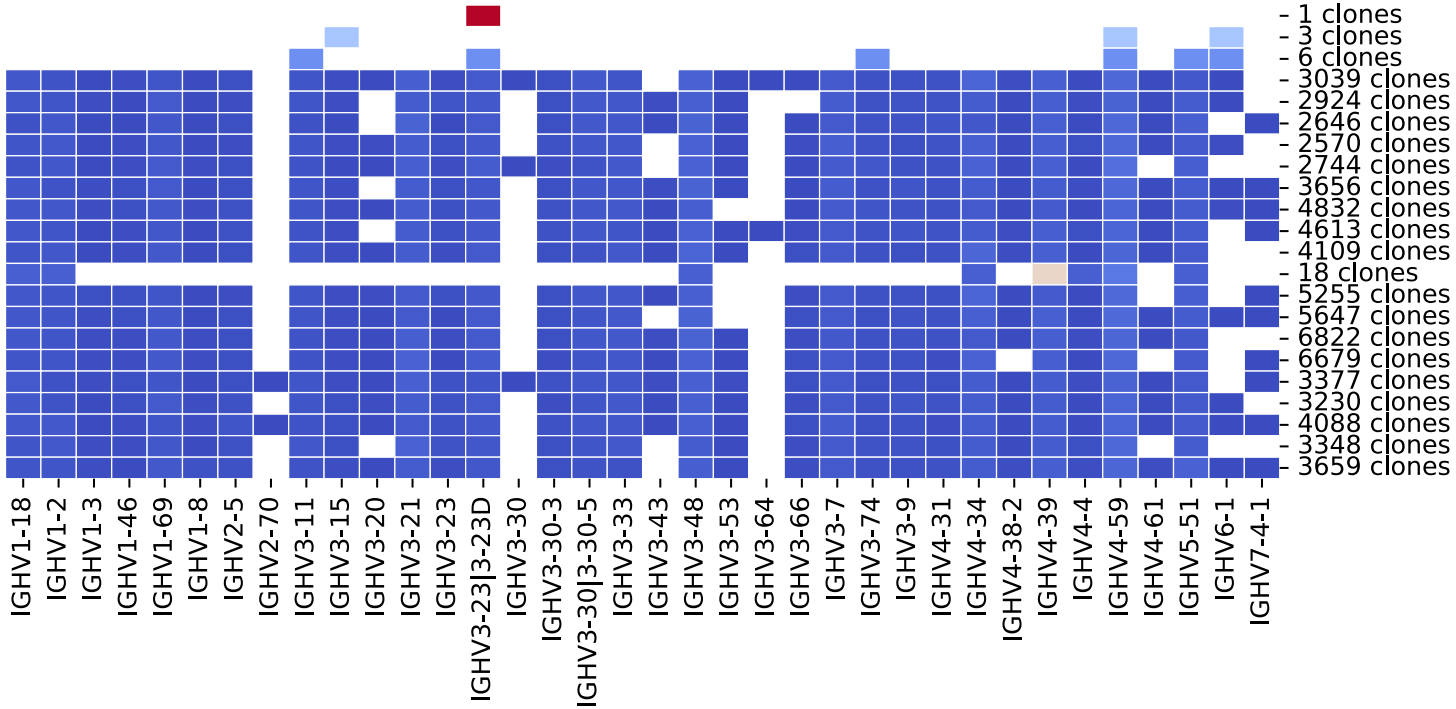
- Unexpectedly high frequency of particular VH or Vb genes?
  - Clonal expansion vs. PCR jackpot
  - Antigen-enrichment vs. amplification bias
- PCR jackpot or poor amplification of individual samples
  - look at clone counts by replicates
  - Poor sample quality or low input?
  - In-line controls to evaluate amplification within individual samples
- Poor run quality overall
  - Copy number analysis of sequencing run cell mix control
  - VH skewing of cell mix control- assess with VH usage plot

Clone count per replicate

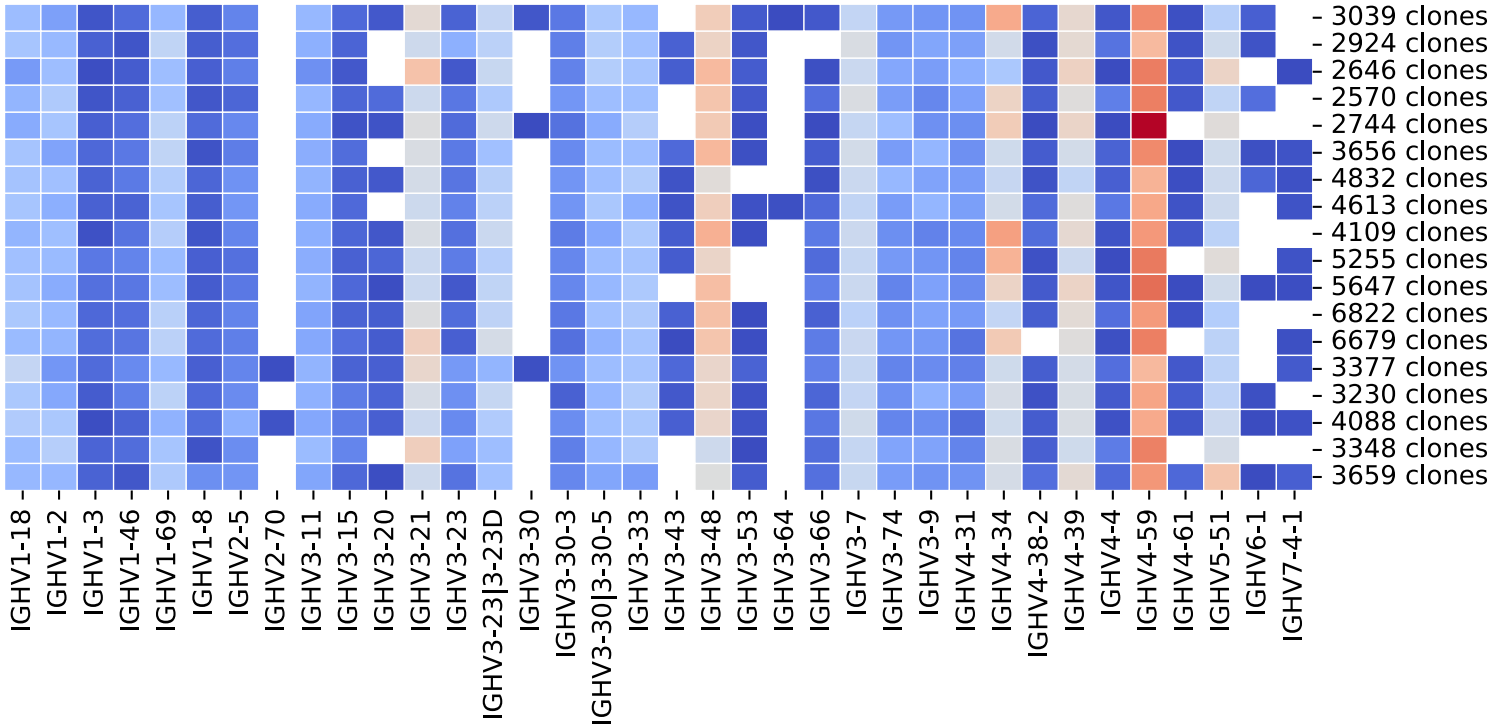
Replicate sequencing reactions from spleen gDNA from an organ donor



Amplification or sequencing failures can cause drastic changes in downstream data

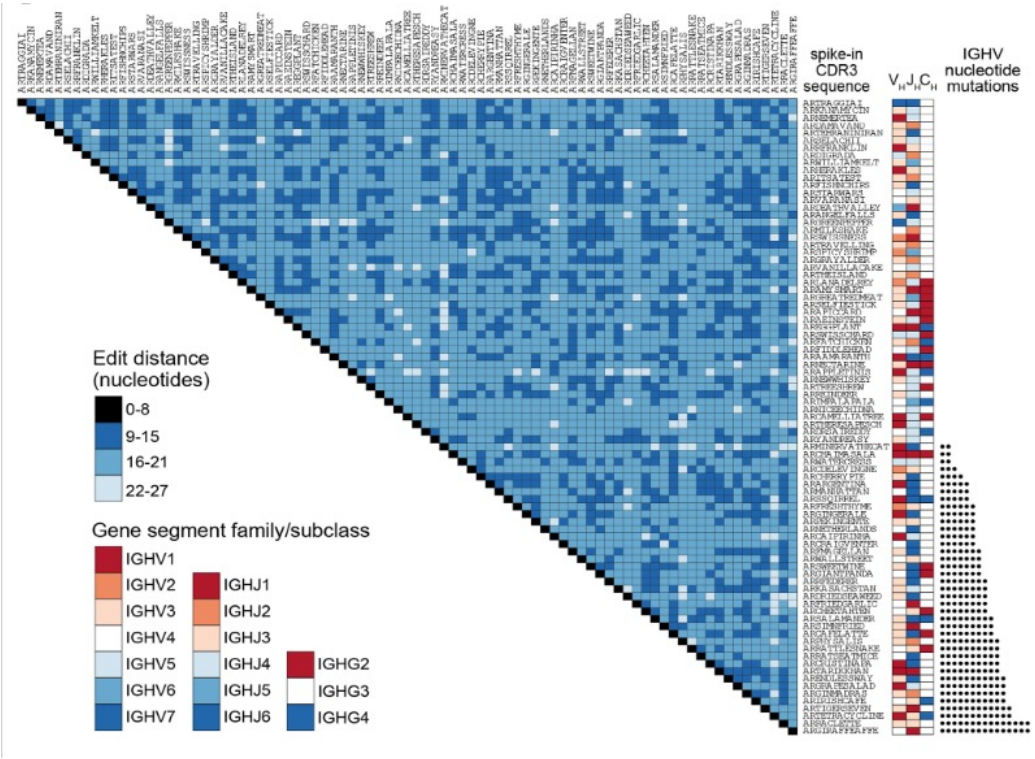


Same data with bad replicates removed and heat map re-scaled



In-line calibrators

principle: synthetic TCR or BCR



Calibrators contain V gene specific sequences and a universal sequence to evaluate their relative abundance

Friedenson et al, Frontiers in Immunology, 2018

In-line calibrators

- Example of a commercialized spike-in kit (Cellecta)

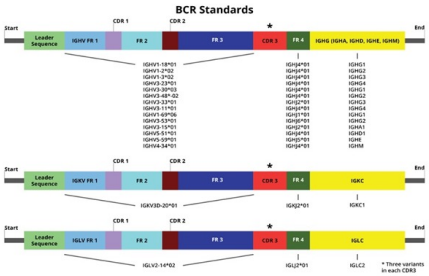
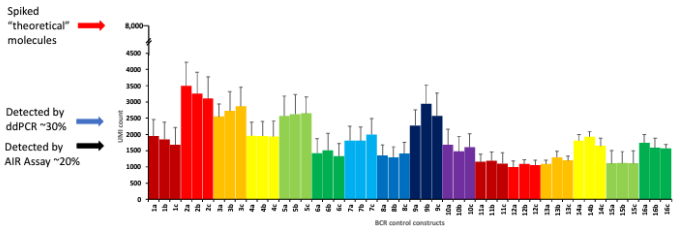


Fig 1: 48 BCR mRNA spike-in constructs represent 10 different IGHs; 1 for each IGHA, IGHD, IGHE, IGHM, IGK and IGL genes.



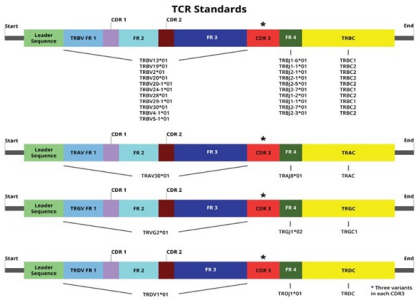


Fig 2: 39 TCR spike-in constructs represent 10 different TRBs, 1 for each of TRA, TRG and TRD genes.

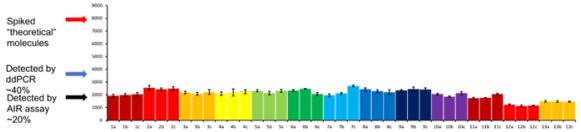
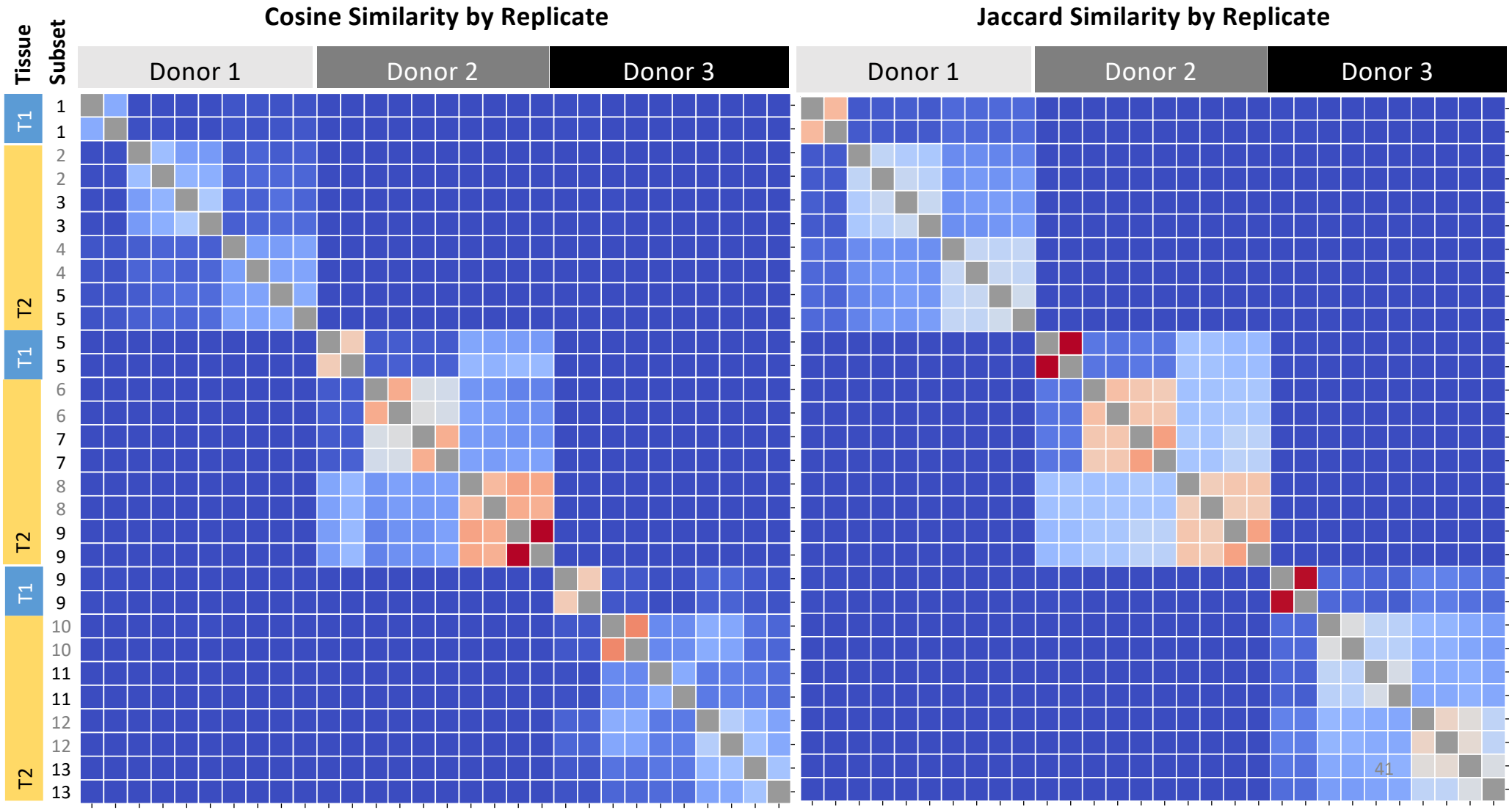


Fig 5: UMI count of 39 TCR control constructs spike-in at 8000 molecules in 50 ng of PBMC RNA

https://cellecta.com/pages/posters

Between subject and sequencing run QC

Analysis of within vs. between subject similarity expectation


Cosine Similarity by Replicate
Jaccard Similarity by Replicate
Donor 1
Donor 2
Donor 3
Donor 1
Donor 2
Donor 3
Tissue
Subset
T1
T2
T1
T2
T1
T2
1
1
2
2
3
3
4
4
5
5
5
5
6
6
7
7
8
8
9
9
9
9
10
10
11
11
12
12
13
13
41

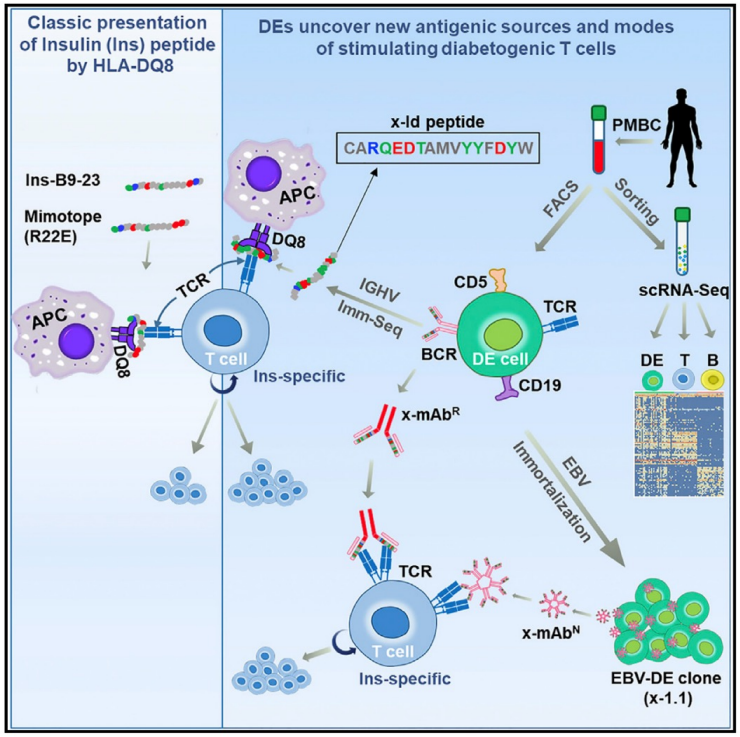
What if you see something unexpected?

A cautionary tale: rule out technical explanations before getting too excited about the biology!

Cell

A Public BCR Present in a Unique Dual-Receptor-Expressing Lymphocyte from Type 1 Diabetes Patients Encodes a Potent T Cell Autoantigen

Graphical Abstract



Authors

Rizwan Ahmed, Zahra Omidian, Adebola Giwa, ..., Chunfa Jie, Thomas Donner, Abdel Rahim A. Hamad

Correspondence

rz24@columbia.edu (R.Z.), ahamad@jhmi.edu (A.R.A.H.)

In Brief

Type I diabetes patients have unique TCR- and BCR-positive lymphocytes, in which a public BCR encodes a potent autoantigen that stimulates autologous CD4 T cells and may contribute to autoimmunity.

See Matters Arising paper by Japp et al. 2021 contradicting the major findings in this paper, including the “public” clone, which arose due to contamination

Sample contamination – data from Ahmed et al.

Inter-donor clonal overlap can indicate sample or PCR contamination

CDR3 amino-acids	Subject	Templates	Alignment
CAGGHNYGIKSYW	T1D1	14,104	caccatctccagagacaattccaagaacacgctgtttcttcaagtcaacagcctgggagctgagg acacggctgtctattacTGTGCGGGTGGACACAACTATGGTATAAAGTCCTACTGGggccaggga
	T1D2	789	---------- ----------^^^^^^^^^^----------
	T1D3	2,720	---------- ----------^^^^^^^^^^----------
CARQEDTAMVYYFDYW	T1D1	8,893	agtagacacgtccaagaaccagttctccctgaagctgagctctgtgaccgccgcagacacggccg tgtattacTGTGCGAGACAGGAGGATACAGCTATGGTTTACTACTTTGACTACTGGggccaggga
	T1D2	531	---------- --------^^^^^^^^^^---------
	T1D3	3,469	---------- --------^^^^^^^^^^---------

Identical nucleotide sequences of CDR3s are suggestive of contamination rather than independent generation of the same CDR3 sequence (very unlikely)

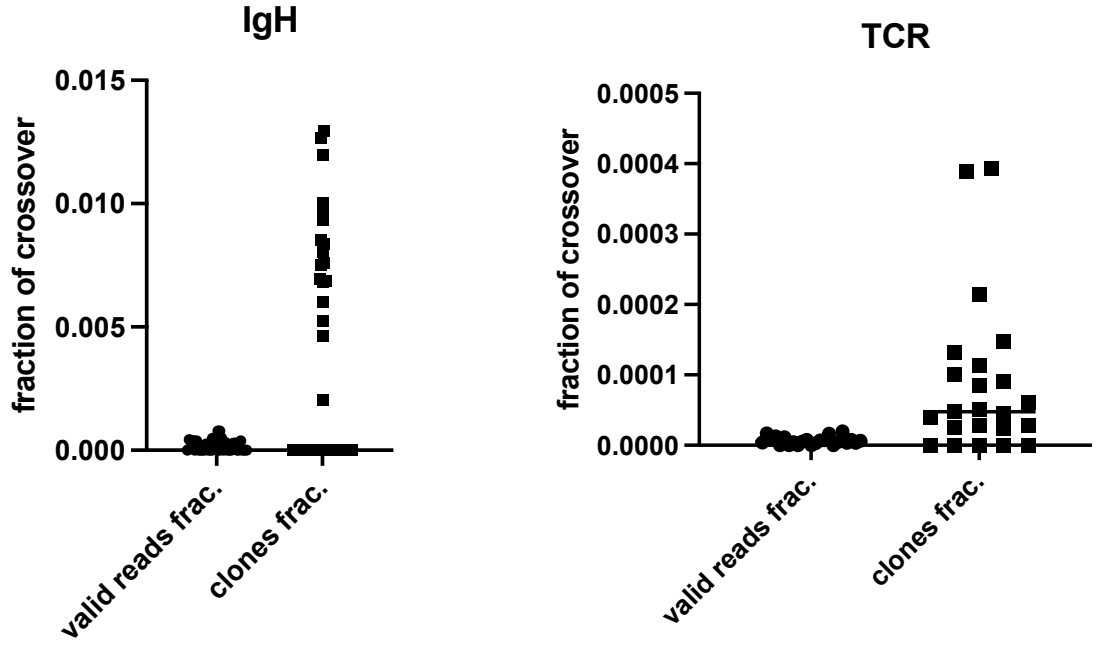
What is Index Hopping?

Samples from an expected index are incorrectly assigned to a different index in the pool

Image from Illumina

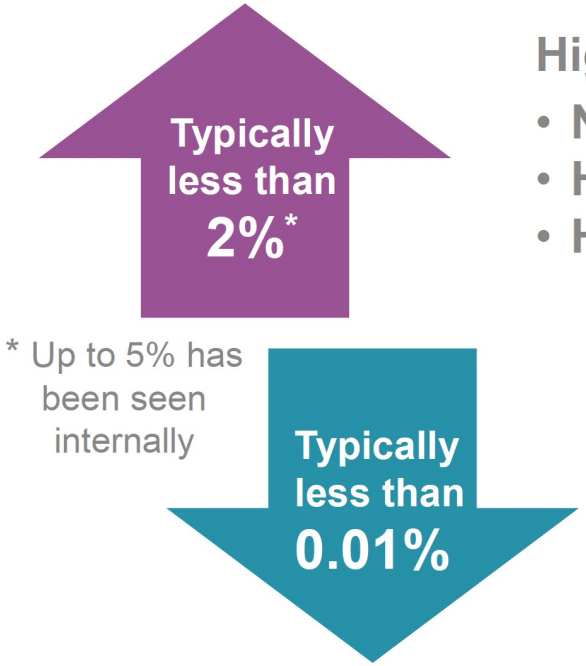
How frequent is index hopping?

A bulk gDNA sequencing run had IGH and TRB gene rearrangements in different samples. We looked for IGH in the TRB samples and vice versa.



Wenzhao Meng, unpublished data
Illumina MiSeq bulk gDNA

Index Hopping Rates By Platform



Higher Index Hopping Rates

- NovaSeq 5000/6000
- HiSeq 3000/4000
- HiSeq X

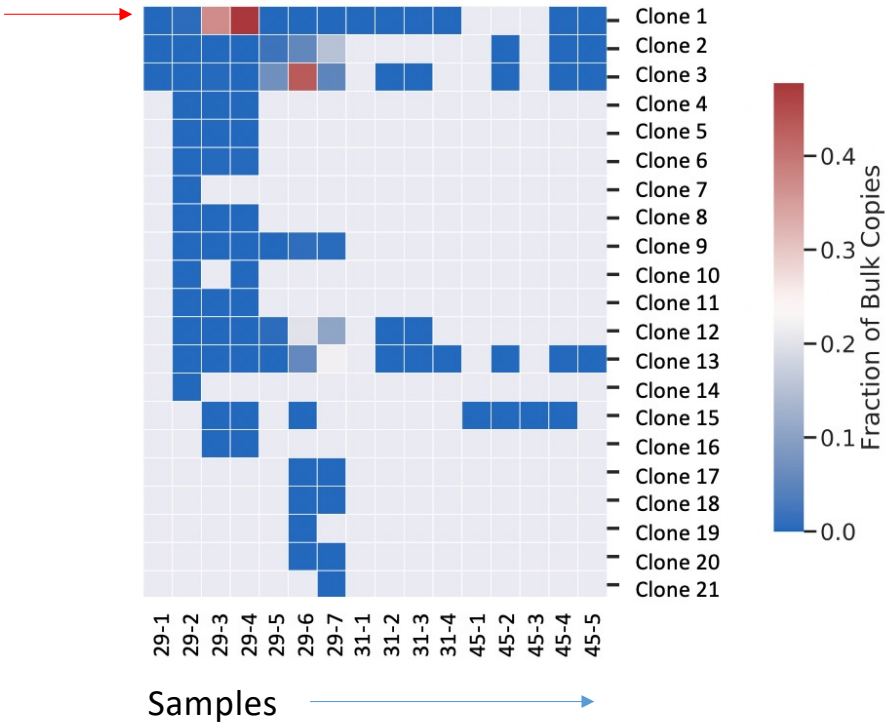
Lower Index Hopping Rates

- NextSeq 500/550
- MiniSeq
- MiSeq
- HiSeq 2500

Image from Illumina
See also Stoler et al. NAR Genomics and Bioinformatics 2021

Index hopping– A case study

Copy number variation in cross-donor clones can indicate cross-clustering



Heat map of shared CDR3 sequences

3 different subjects: 29, 31 and 45

Clonally related sequences grouped together on the basis of CDR3 AA sharing

Clones 1 and 2 are present in all three subjects

Copy number of Clone 1 is 7×10^4 in one subject, less than 100 in others

Cross clustering – A case study

Copy number variation in cross-donor clones can indicate cross-clustering

Long, identical CDR3 NTs in 3 donors

```
S1  : TGTNCACACGGGAGTNNGAGGNCGTNGCTGGGAGACTTTGACTTTTGG
S2  : TGTGCACACGGGAGTCAGAGGTCGTGGCTGGGAGACTTTGACTTTTGG
S3  : TNTGCANACGGGAGTCAGAGGTCGTGGNTGGGAGACTTTGACTTTTGG
```

Aligned full sequences were also identical

```
S3  _15     NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN  60
S1  _70560  NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN  60
S2  _76     NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN  60
            ************************************************************

S3  _15     NNNNNNNNNNNNNNNNNNTTCTCACTCAGGTCTAATGGAGTGGGTGTGGGCTGGATCCGT  120
S1  _70560  NNNNNNNNNNNNNNNNNNGTTCTCACTCAGGTCTAATGGAGTGGGTGTGGGCTGGATCCGT  120
S2  _76     NNNNNNNNNNNNNNNNNNGTTCTCACTCAGGTCTAATGGAGTGGGTGTGGGCTGGNNCCGT  120
            ****************** ********************************** ****

S3  _15     CAGCCCCCAGGAAAGGCCCTGGAGTGGCTTGCAATCATTTATTGGAATGATGATAAGCGC  180
S1  _70560  CAGCCCCCAGGAAAGGCCCTGGAGTGGCTTGCAATCATTTATTGGAATGATGATAAGCGC  180
S2  _76     CAGCCCCCAGGAAAGGCCCTGGAGTGGCTTGCAATCATTTATTGGAATGATGATAAGCGC  180
            ************************************************************

S3  _15     TACAGCCCATCTCTGAAGAGCAGGCTCACCATCACCAAGGACACCGCCAAAAAGCAGGTG  240
S1  _70560  TACAGCCCATCTCTGAAGAGCAGGCTCACCATCACCAAGGACACCGCCAAAAAGCAGGTG  240
S2  _76     TACAGCCCATCTCTGAAGAGCAGGCTCACCATCACCAAGGACACCGCCAAAAAGCAGGTG  240
            ************************************************************

S3  _15     GTCCNGACAATGACCAACATGGACCCTGTGGACACCGCCACATATTACTNTGCANACGGG  300
S1  _70560  GTCCTGACAATGACCAACATGGACCCTGTGGACACCGCCACATATTACTGTNCACACGGG  300
S2  _76     GTCCTGACAATGACCAACATGGACCCTGTGGACACCGCCACATATTACTGTGCACACGGG  300
            **** ********************************************* * ** *****

S3  _15     AGTCAGAGGTCGTGGNTGGGAGACTTTGACTTTTGGGGCCAGGGAACCCTGGTCACCGTC  360
S1  _70560  AGTNNGAGGNCGTNGCTGGGAGACTTTGACTTTTGGGGCCAGGGAACCCTGGTCACCGTC  360
S2  _76     AGTCAGAGGTCGTGGCTGGGAGACTTTGACTTTTGGGGCCAGGGAACCCTGGTCACCGTC  360
            ***  **** *** * ********************************************

S3  _15     TCNTCAG  367
S1  _70560  TCCTCAG  367
S2  _76     TCCTCAG  367
            ** ****
```

Conclusion: most likely due to cross-clustering in flow cell or low-level contamination

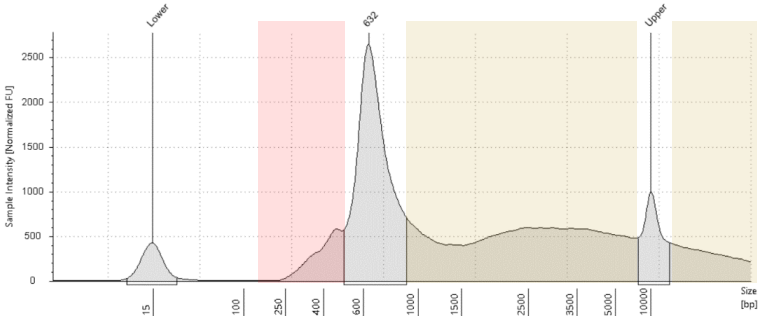
Causes and remedial actions for index hopping

- Causes
  - Index adapters are contaminated during or prior to library prep
  - Samples are mixed up and associated with the wrong index adapters
  - Experimental designs that can be associated with higher rates of index hopping: use of ligation-based library prep method (PCR-free Tru-seq), storage of sequencing library at RT (store at -20^{o}C), use of a patterned flow cell, inclusion of samples with very different copy numbers
- Remedies
  - Use a non-patterned flow cell
  - Use samples with similar sequencing depth (preferably higher diversity)
  - Use of free adapter blocking agent

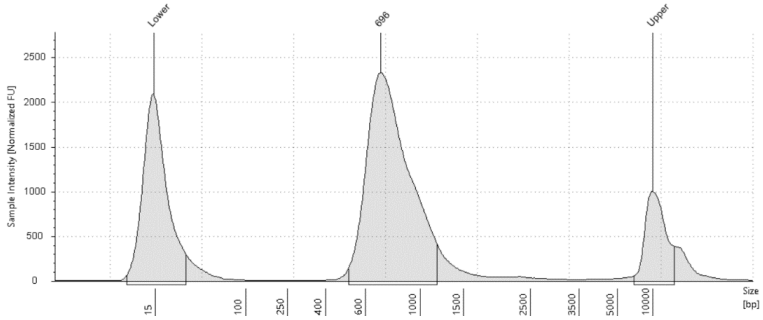
High level purification to remove small and large fragments

96 multiplexed libraries for sequencing

Before purification

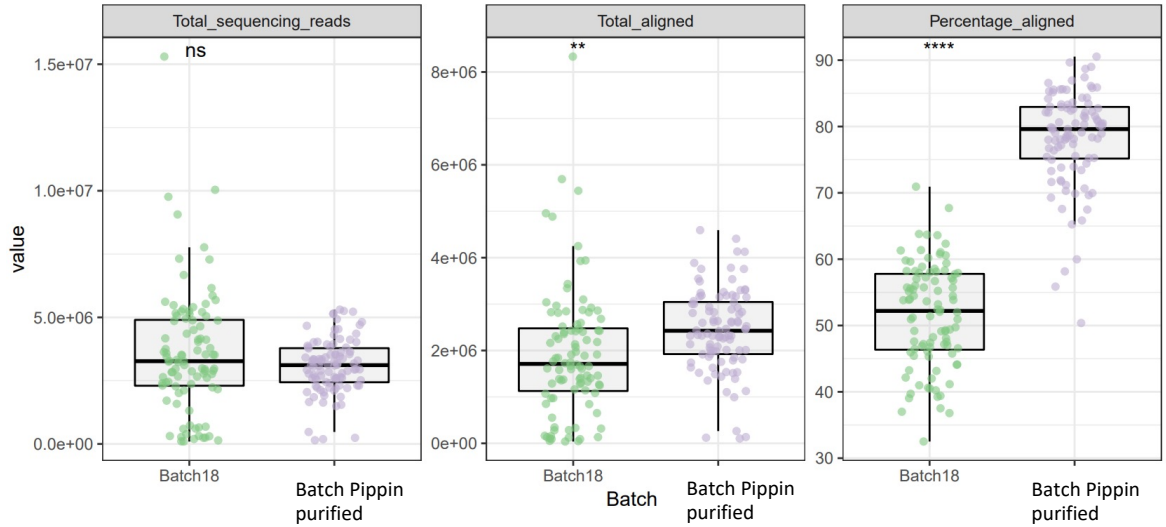


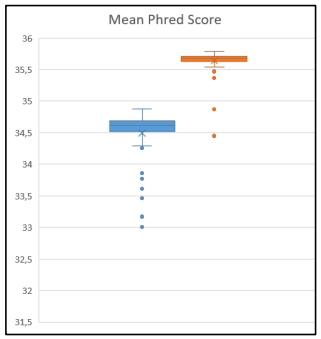
After purification



Small and large fragments are absent from purified samples

Pippin prep purified libraries are better aligned

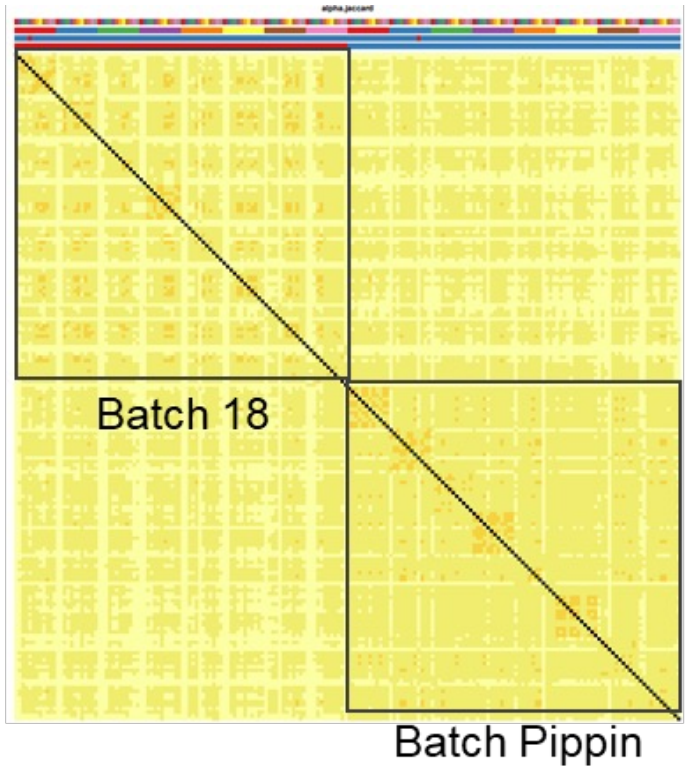




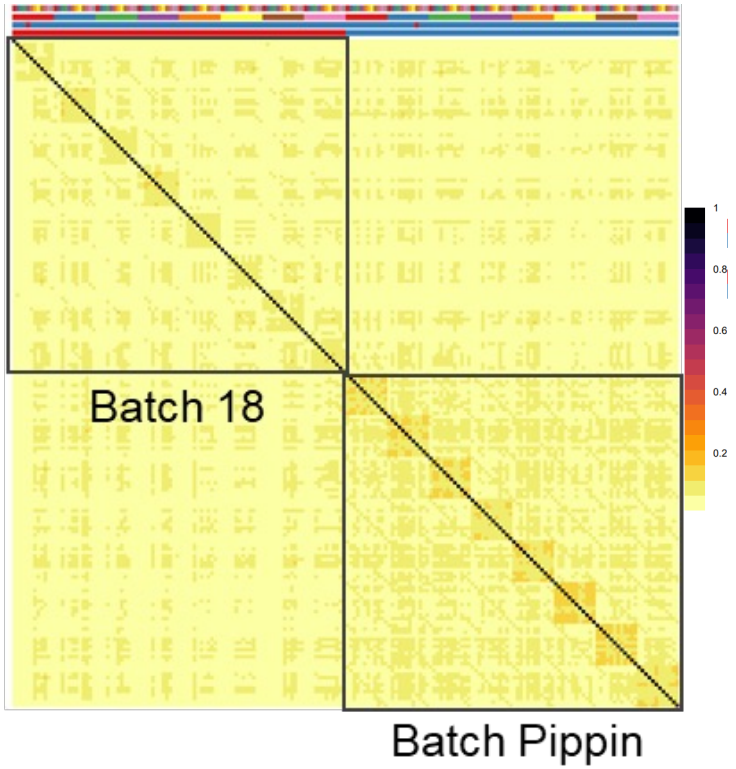
- No differences in sequenced reads per lane
- **66% more aligned reads in Batch19**
- 1 point higher Phred score

Pippin purification has no impact on index hopping

α Jaccard



β Jaccard



Jaccard similarity
Samples sorted by lane and reverse.

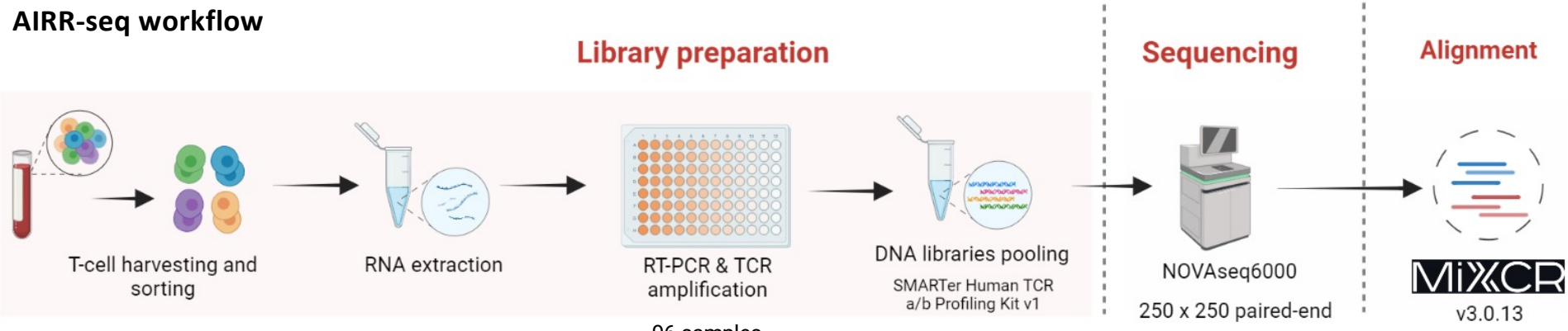
Clear index hopping on both batch.

Need to integrate such complex workflow to identify issues:
An example of a QC data analysis pipeline

Encarnita

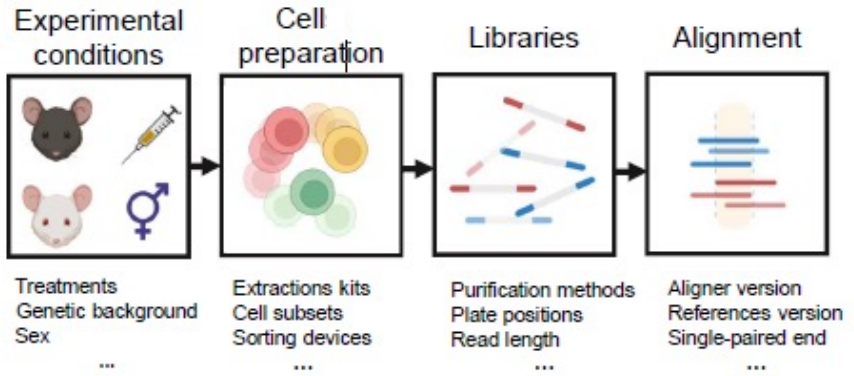
AIRR-seq QC: what do we need to know?

AIRR-seq workflow



96 samples

Record the full history of your data:



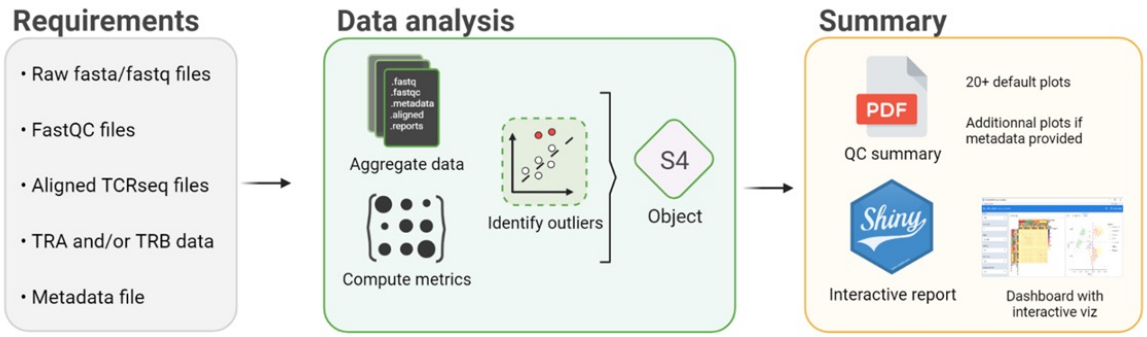
All these variables can influence on your results

Needed to distinguish biological differences from technical biases

Integrated quality control pipeline

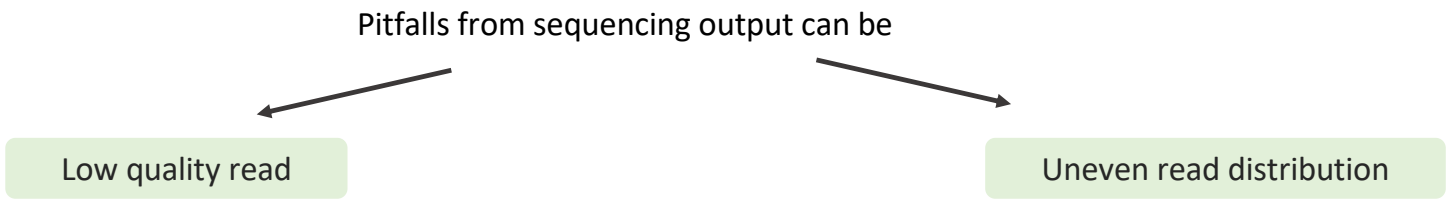
QtCR is a QC pipeline provides a **comprehensive and tunable** quality control (QC)

- Provides **global** and **in-depth metrics** to evaluate sample quality
- Combines both **raw** and **aligned** reads QC
- Currently developping an **user-friendly graphic interface** for at-a-glance analysis



1. Global raw reads' quality assessment

.meta .fastqc



1. Global raw reads' quality assessment

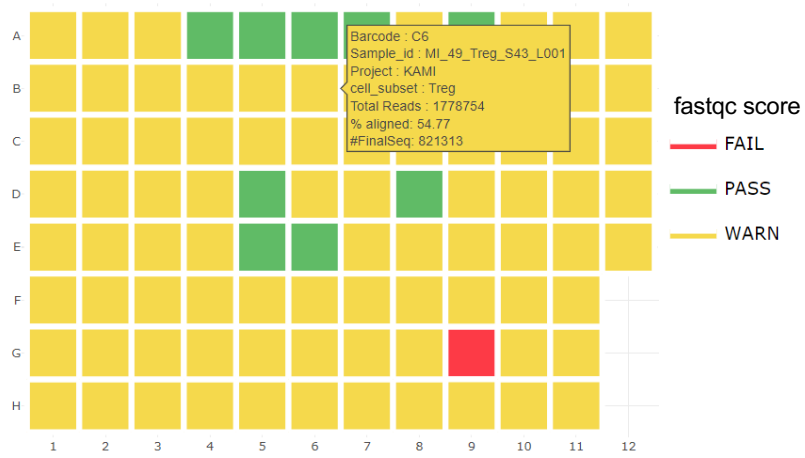
.meta .fastqc

Pitfalls from sequencing output can be

Low quality read

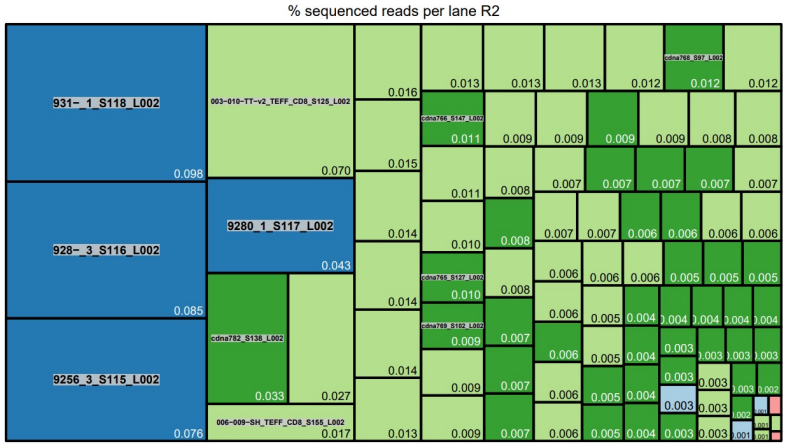
Uneven read distribution

fastqc : Per base sequence quality



This run has poor sequence quality, regardless of sample position

Treemap of raw reads distribution



Top 10 samples represent 40% of the sequenced reads

Complicated samples (low input material, library quantification issues…): flow-cell overloading if libraries were underestimated

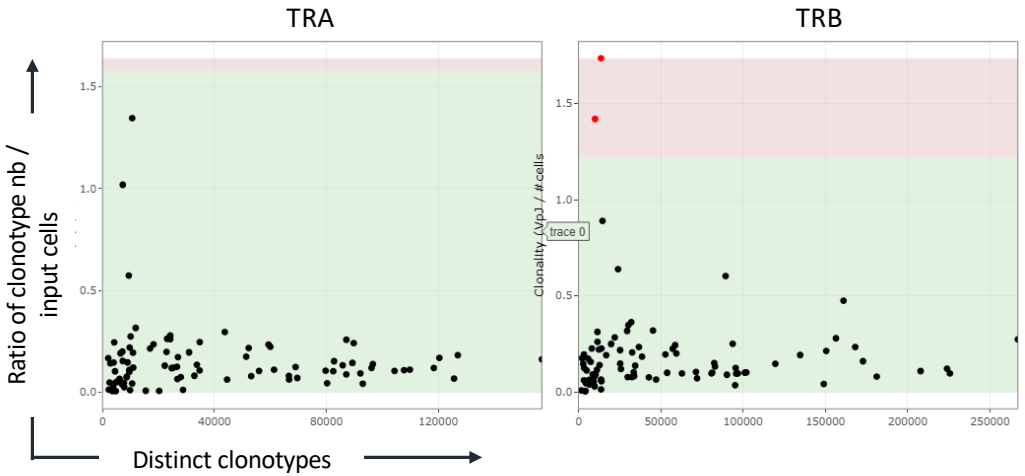
Pooling issue here: few libraries underestimated, took over the run!

2. Consistency between data and biology

.meta . aligned

a. Consistency between cell input and clonotype output

You are not supposed to have more unique clonotypes than initial cells



2 samples (in red) with low initial cell number have an irregular high number of TRB but not TRA clonotypes

Amplification issue: kit lot, primers….

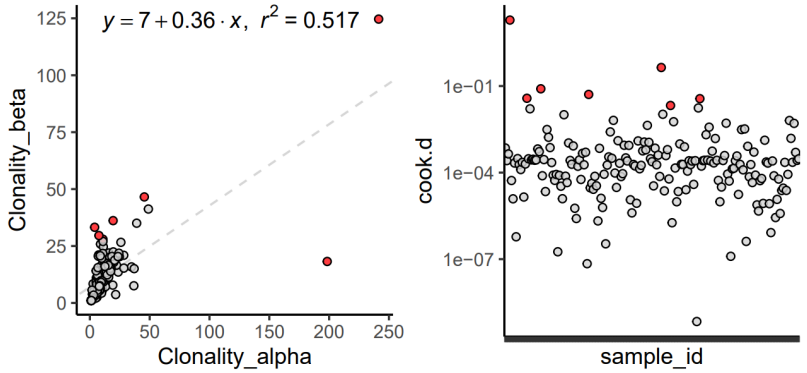
2. Consistency between data and biology

.meta . aligned

b. Consistency between TRA/TRB clonality

We do not know how **alpha** and **beta** chains **clonality** are related at the RNA level, however we can spot samples that are far from the others.

Linear regression Cook distance



Cook distance evaluates how a point influences regression line.

Here the 7 samples highlighted in red are above cook's threshold and will require further checks

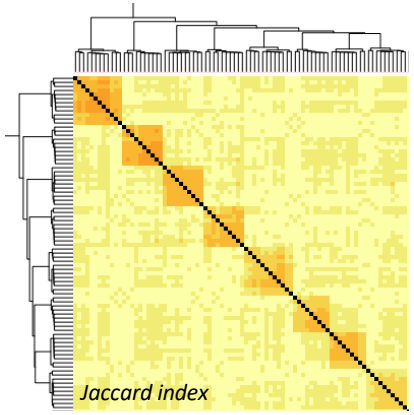
3. Identifying contamination

.meta . aligned

Investigating contamination is a long and tedious work.

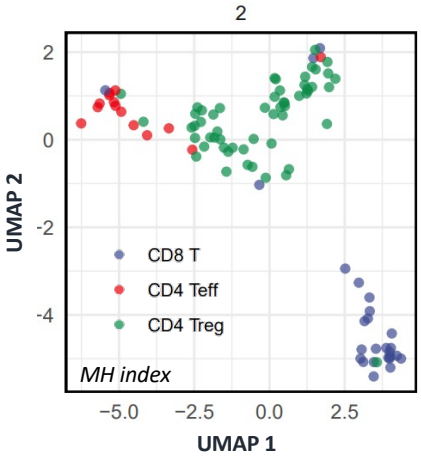
You can spot it with jaccard / mh indexes.

Run-scaled contamination



Necessity to identify common variables

Sample swapping



A few samples were most likely swapped

This works well when you already know what variables explains your data.

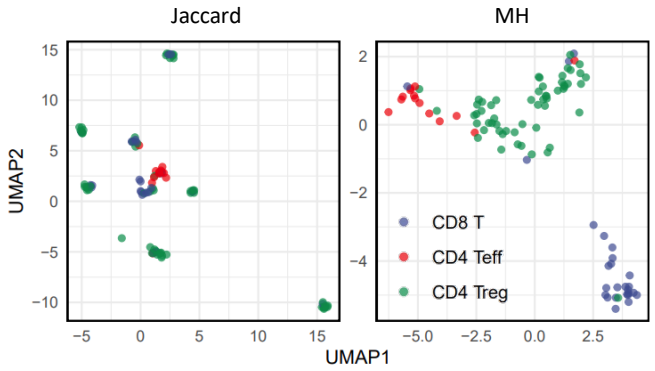
With **dozens** of variables, how do you screen all of them ?

3. Identifying contamination

.meta . aligned

Principal Variance Component Analysis (PVCA)

PVCA is used to identify batch effects by combining PCA and mixed linear models.
It screens sources of variability (variables) in datasets



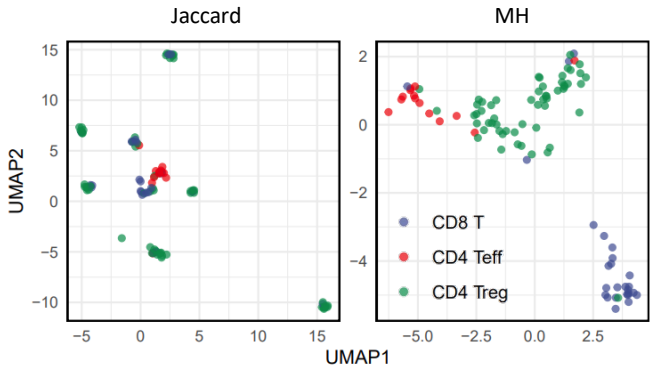
UMAP is driven by T-cells subsets for MH, but jaccard relies on other **unknown** variables

3. Identifying contamination

.meta . aligned

Principal Variance Component Analysis (PVCA)

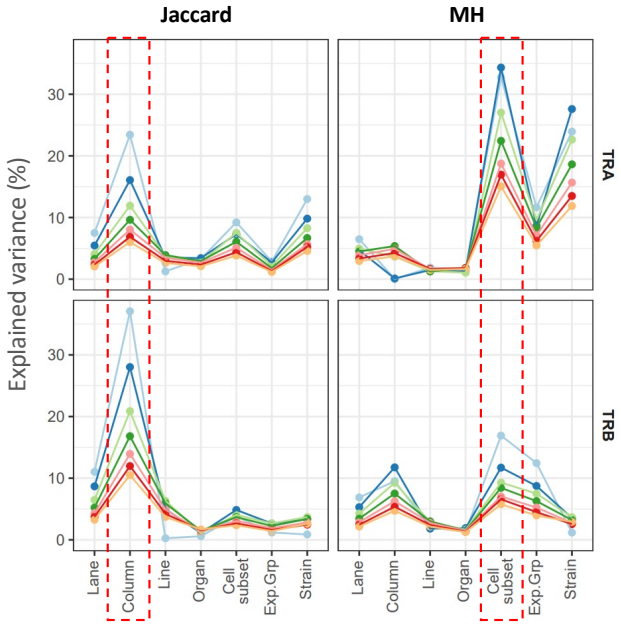
PVCA is used to identify batch effects by combining PCA and mixed linear models.
It screens sources of variability (variables) in datasets



UMAP is driven by T-cells subsets for MH, but jaccard relies on other **unknown** variables

PVCA

- Column
- Line
- Organ
- Cell-subset
- Strain
- …



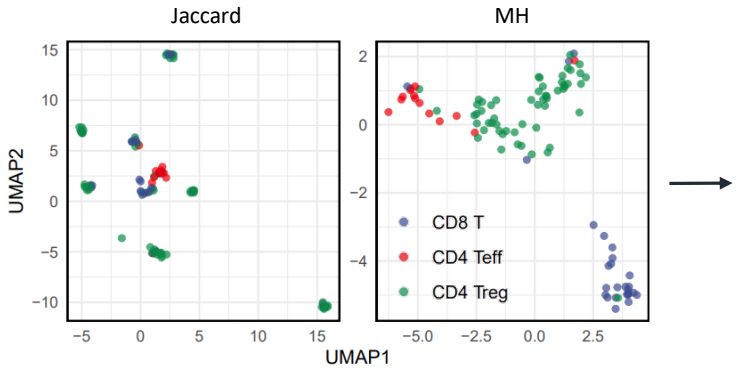
PVCA identifed **column** as main driver of jaccard and **cell subset** as main driver of MH

4. Identifying contamination

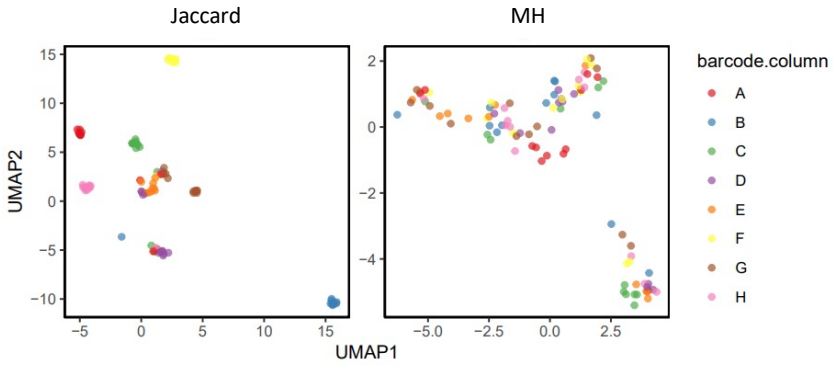
.meta . aligned

Principal Variance Component Analysis (PVCA)

PVCA is used to identify batch effects by combining PCA and mixed linear models.
It screens sources of variability (variables) in datasets



UMAP is driven by T-cells subsets for MH, but jaccard relies on other **unknown** variables



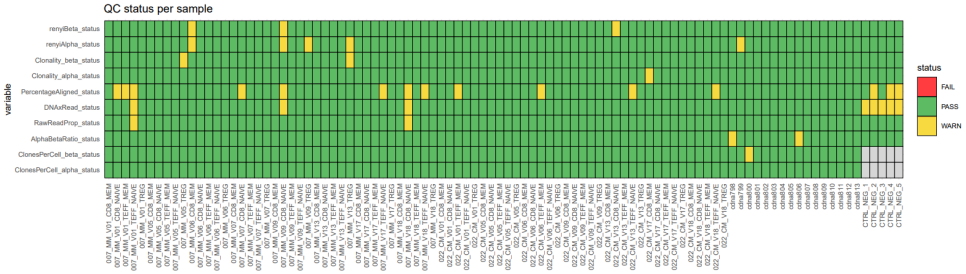
PVCA is a method suited to identify batch effects parameters at the scale of a sequencing run.

4. Combine and aggregate QC results

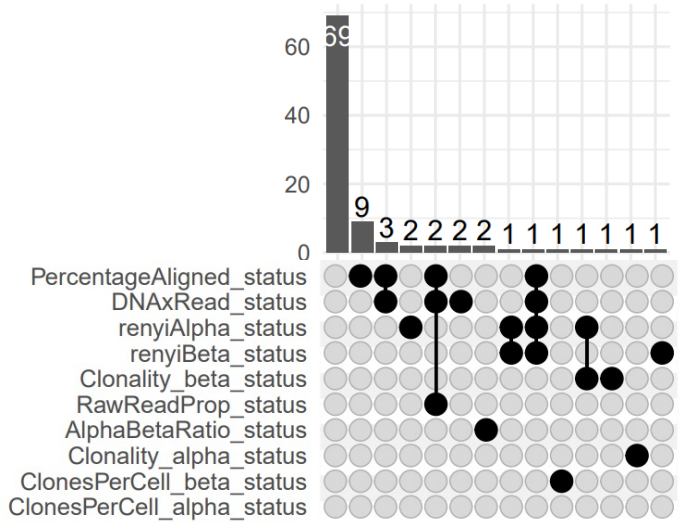
.meta | .fastqc | .reports | .aligned

» Easily access QC data

sample_id	ClonesPerCell_alpha	ClonesPerCell_beta	AlphaBetaRatio	RawReadProp	DNAxRead
007_MM_V01_CD8_MEM	✔PASS	✔PASS	✔PASS	✔PASS	✔PASS
007_MM_V01_CD8_NAIVE	✔PASS	✔PASS	✔PASS	✔PASS	✔PASS
007_MM_V01_TEFF_MEM	✔PASS	✔PASS	✔PASS	✔PASS	✔PASS
007_MM_V01_TEFF_NAIVE	✔PASS	✔PASS	✔PASS	✖WARN	✖WARN
007_MM_V01_TREG	✔PASS	✔PASS	✔PASS	✔PASS	✔PASS
007_MM_V05_CD8_MEM	✔PASS	✔PASS	✔PASS	✔PASS	✔PASS



» Identify patterns of failures



Benchmarking studies and ongoing AIRR-C initiatives

Encarnita and Nina

Benchmarking of T cell receptor repertoire methods

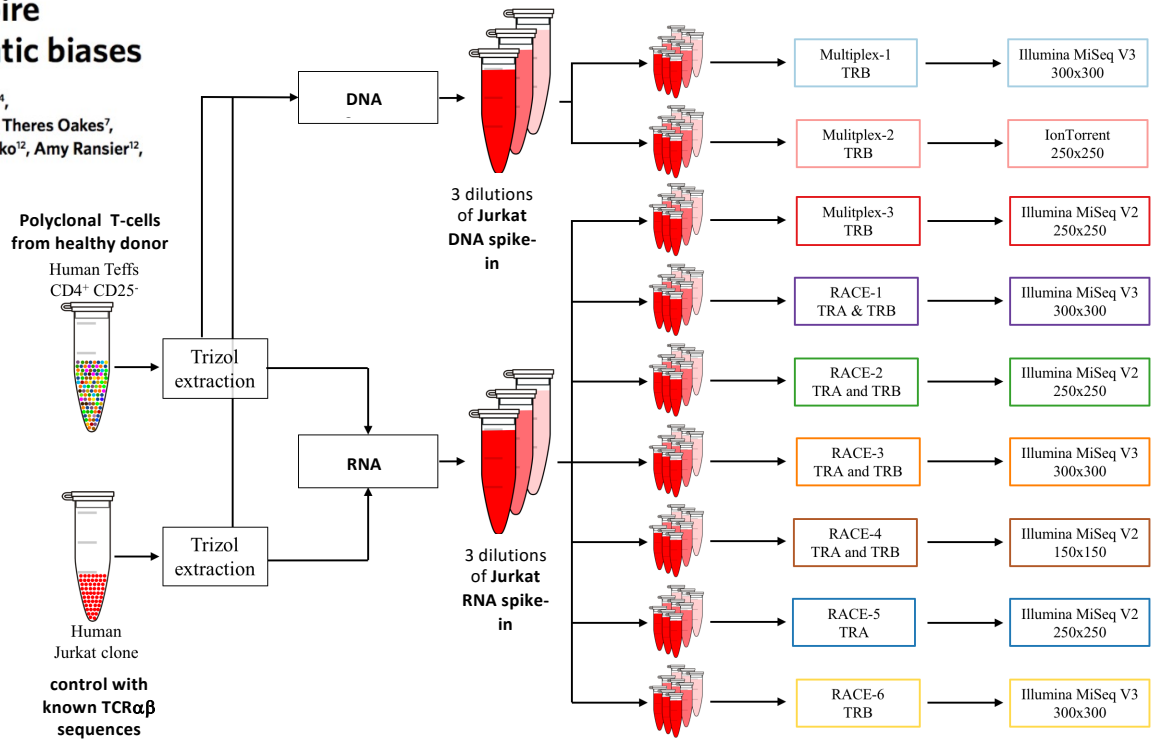
nature biotechnology
ANALYSIS
https://doi.org/10.1038/s41587-020-0656-3
Check for updates

Benchmarking of T cell receptor repertoire profiling methods reveals large systematic biases

Pierre Barennes[1,2], Valentin Quiniou[1,2], Mikhail Shugay[3,4,5], Evgeniy S. Egorov[4], Alexey N. Davydov[6], Dmitriy M. Chudakov[3,4,5,6], Imran Uddin[7], Mazlina Ismail[7], Theres Oakes[7], Benny Chain[7], Anne Eugster[8], Karl Kashofer[9], Peter P. Rainer[10,11], Samuel Darko[12], Amy Ransier[12], Daniel C. Douek[12], David Klatzmann[1,2] and Encarnita Mariotti-Ferrandiz[1,2]✉

Pierre Barennes
PhD student

- Most used methods (publications)
- Voluntary based (challenge their method)
  - Academic labs
  - Commercially available kits
    - Service providers
    - Reagent providers



N_{tot}=81 replicates

Encarnita

Summary of TCR method benchmarking

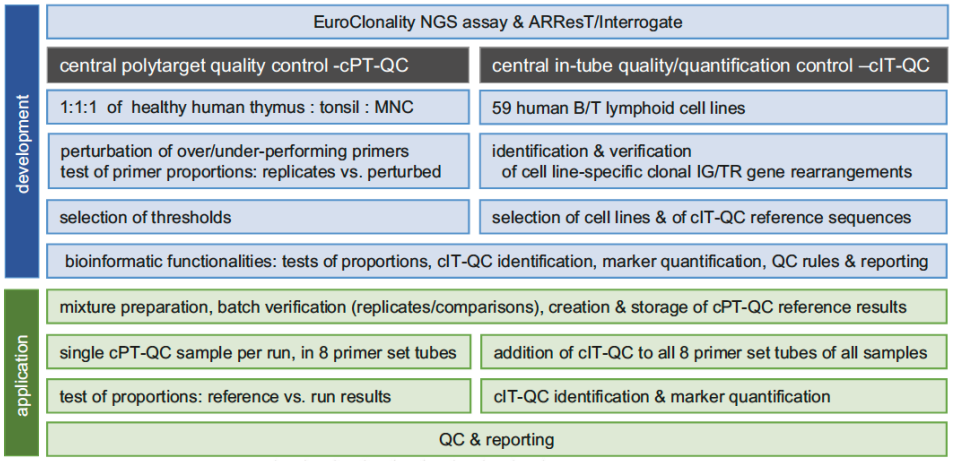
Table 1 | Comparative performance of the nine TCRseq molecular methods

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

The lower the score the better the method

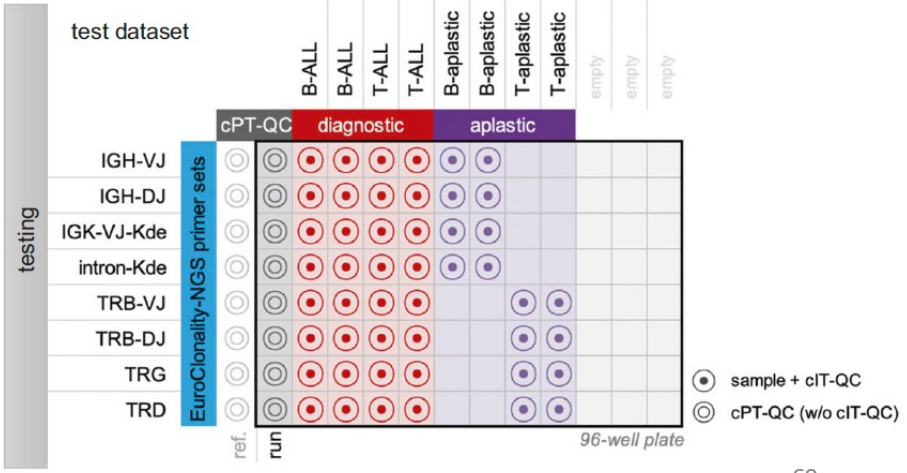
Euroclonality consortium: quality control and quantification of TR and IG rearrangements

H. Knecht et al.



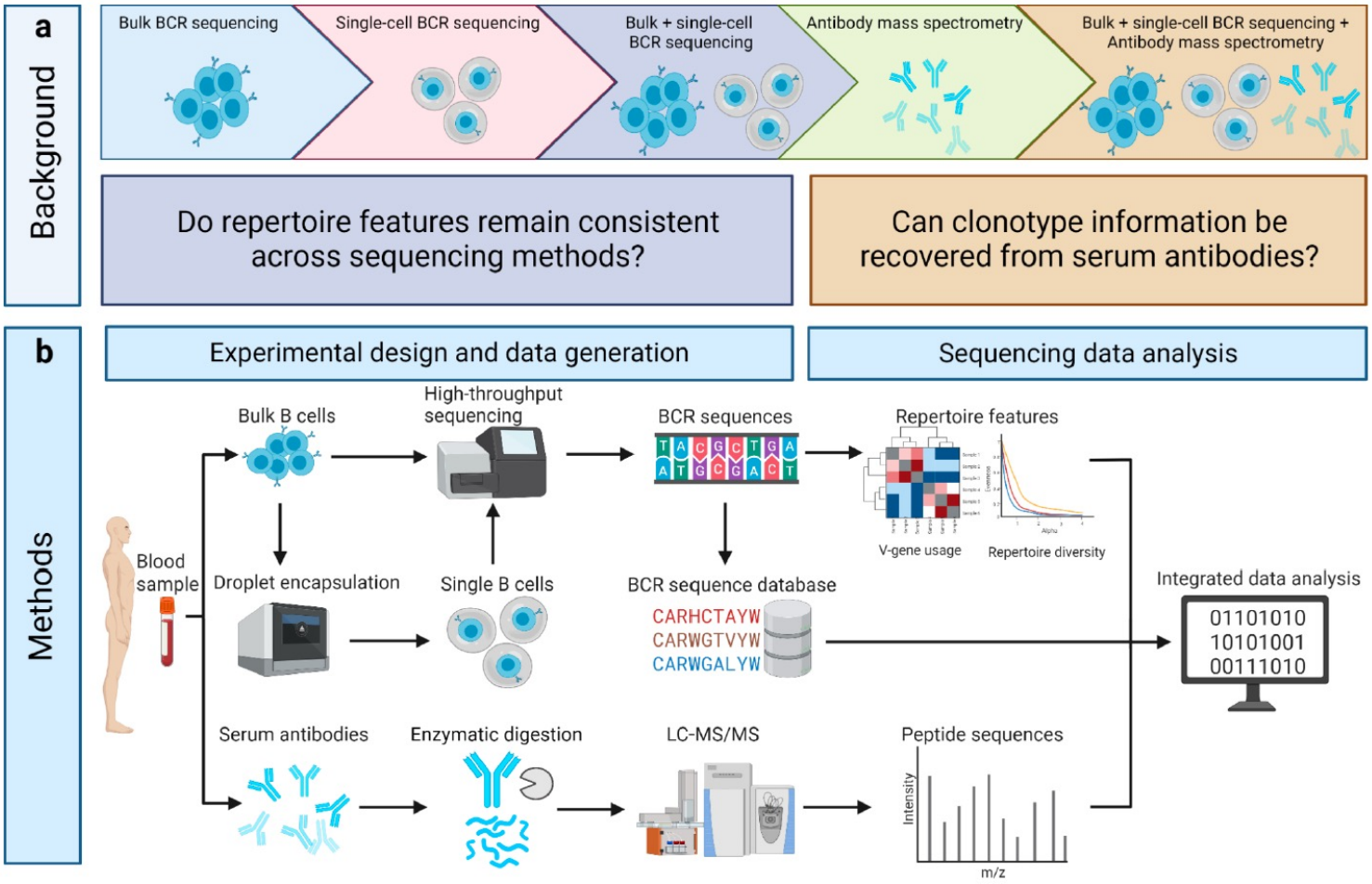
Knecht et al. 2019 Leukemia

1. A central polytarget QC (cPT-QC) consisting of a standardised mixture of lymphoid specimens, representing a full repertoire of IG/TR genes. It serves to assess performance biases or unusual amplification shifts in a sequencing run by tracking primer usage and comparison with stored reference profiles.
2. A central in-tube quality/quantification control (cIT-QC) consisting of human B and T cell lines with well-defined IG/TR rearrangements. The cIT-QC is directly added to a sample to undergo concurrent library preparation and sequencing, acting as in-tube qualitative and quantitative standard that is subjected to the same technical downstream variables.



Nina

Benchmarking and integrating human B-cell receptor genomic and antibody proteomic profiling



Quy et al. bioRxiv preprint 2023

Ongoing AIRR-C initiatives related to AIRR-seq QC

AIRR-seq special issue Methods in Molecular Biology 2453, Springer protocols (Anton Langerak, editor)

- Chapter 21: Quality control: chain pairing precision and monitoring of cross-sample contamination (Brandon DeKosky and AIRR Community)

AIRR-C Biological resources working group

- FDA BCR-SEQC project (led by Wenming Xiao)
  - in –depth characterization of cell line mixtures as controls for IGH
  - Benchmarking studies with different AIRR-seq methods

- Analysis of unique molecular identifiers
  - depth of coverage required for different applications
  - error correction methods

Nina

Question and answer session

Encarnita and Nina