Adaptive Immune Receptor Repertoire Sequencing (AIRRseq)

Barbera van Schaik
Bioinformatics Laboratory
Epidemiology & Data Science
Amsterdam UMC
b.d.vanschaik@amsterdamumc.nl

Overview

Introduction AIRRseq

* T- and B-cell receptors
* VDJ recombination
* Clonal expansion and somatic hypermutation
* Class switching
* Repertoire sequencing

Data analysis

* Pre-processing
* Identify V, (D), J and C regions
* Identify CDR3 region
* Determine clones and their frequency

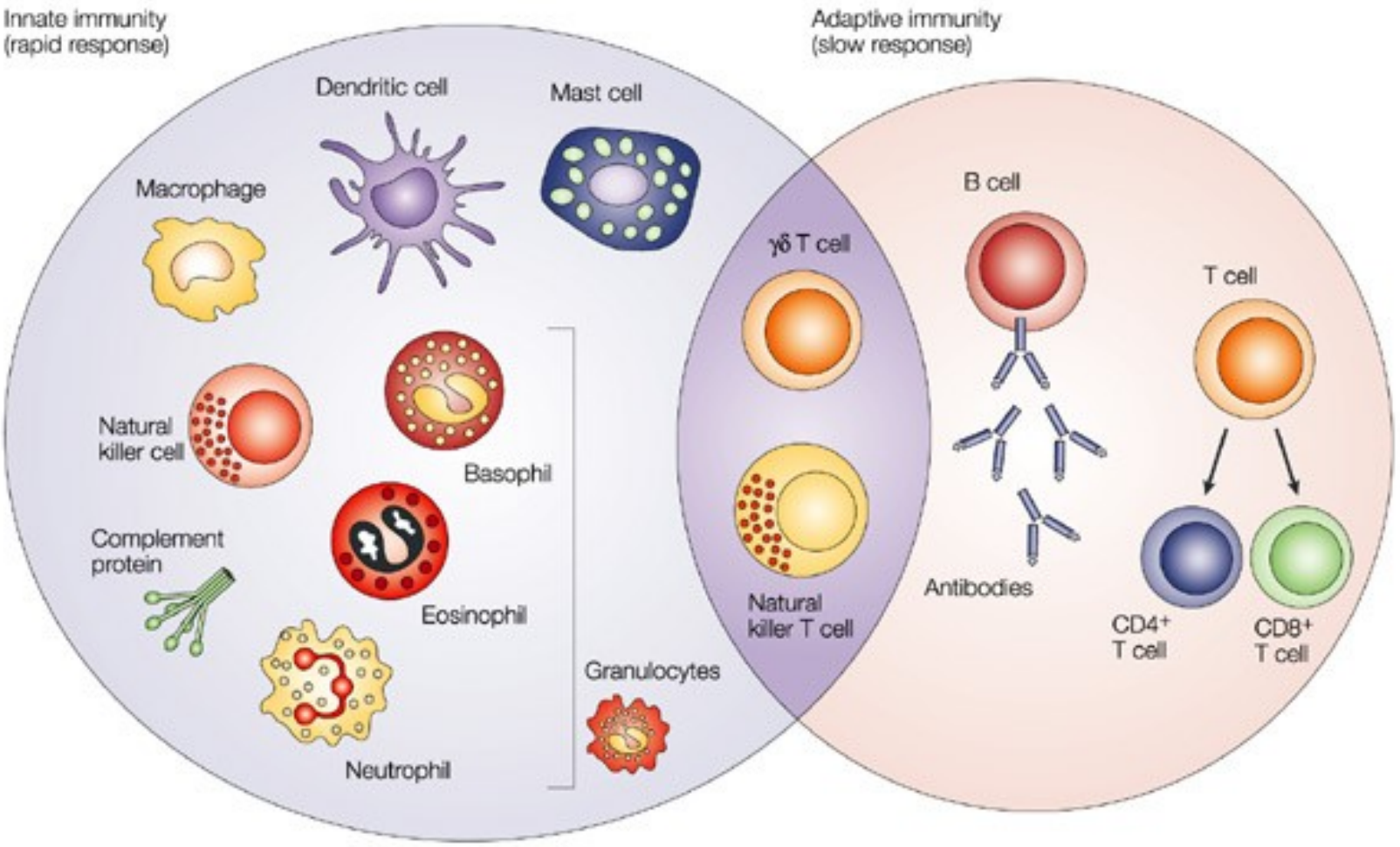
Follow-up analysis

* Dominant clones
* Compare samples
* Diversity of samples
* Shared clones between samples

Computer systems for data analysis

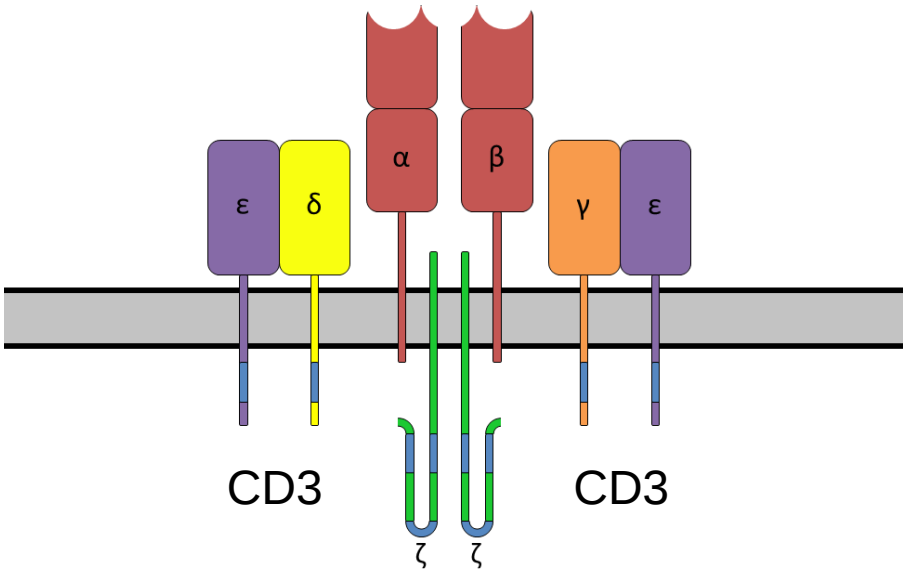
* Clusters, cloud, linux
* Overview of computer lab

Innate and adaptive immune system



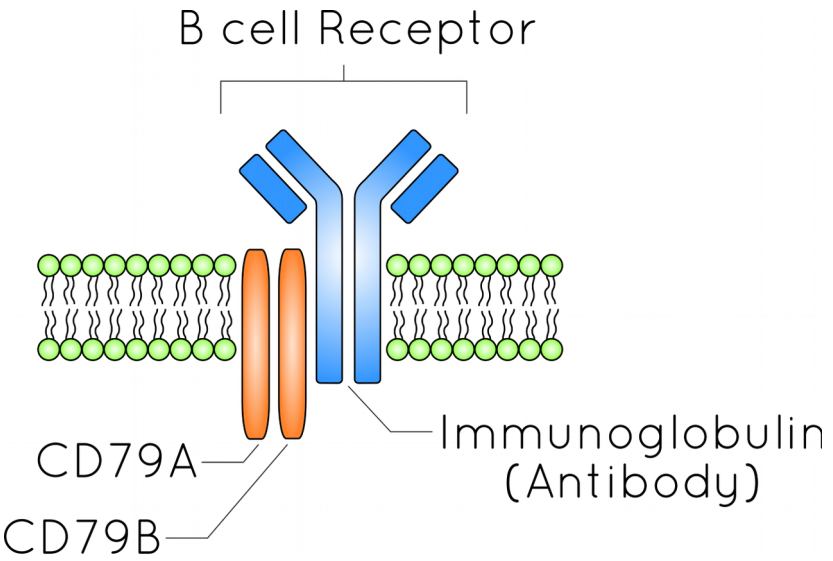
Dranoff (2004) Nature Reviews Cancer

T and B cells



T cell receptor

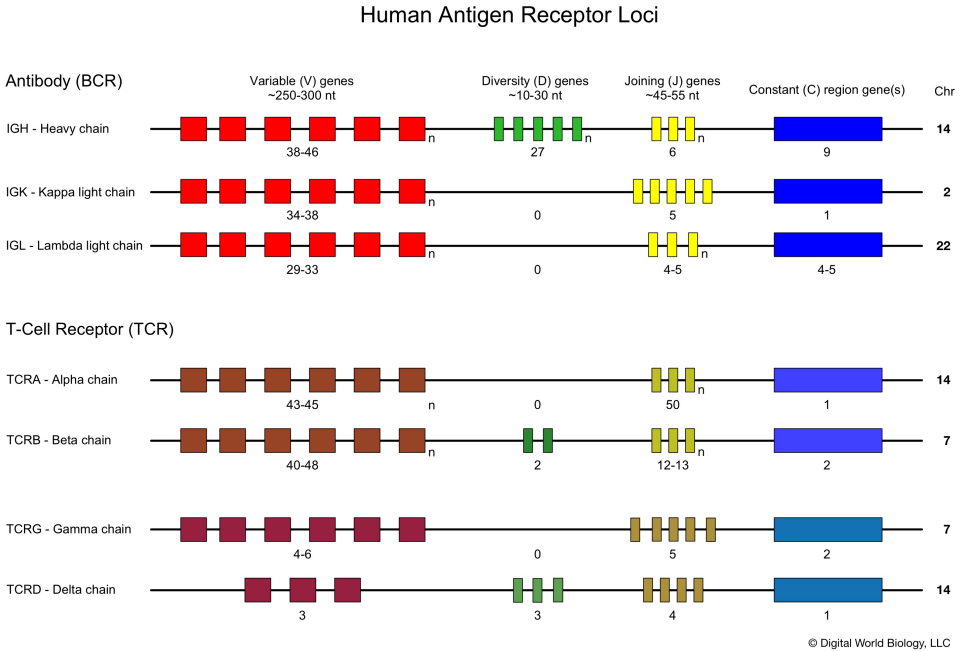
1. Part of the adaptive immune system
2. Slower, but **specific** immune response
3. Trick is the high variety of T- and B-cells: immune receptors



Wikipedia

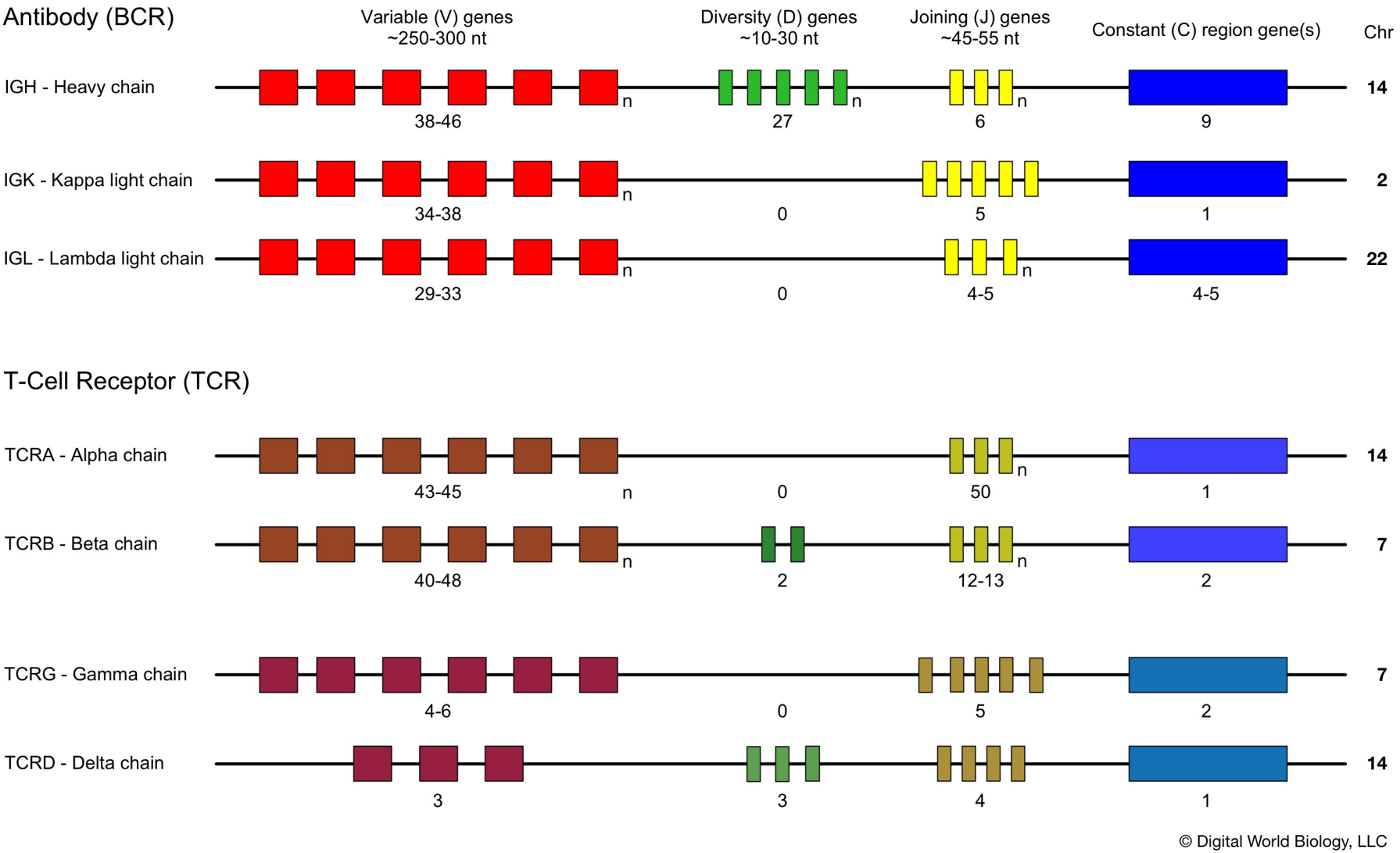
Immune receptors: V, D, J recombination

- BCR: heavy (heavy) and light chains (kappa, lambda)
- TCR: beta/alpha and gamma/delta chains



https://digitalworldbiology.com/blog/immunoprofiling-how-it-works

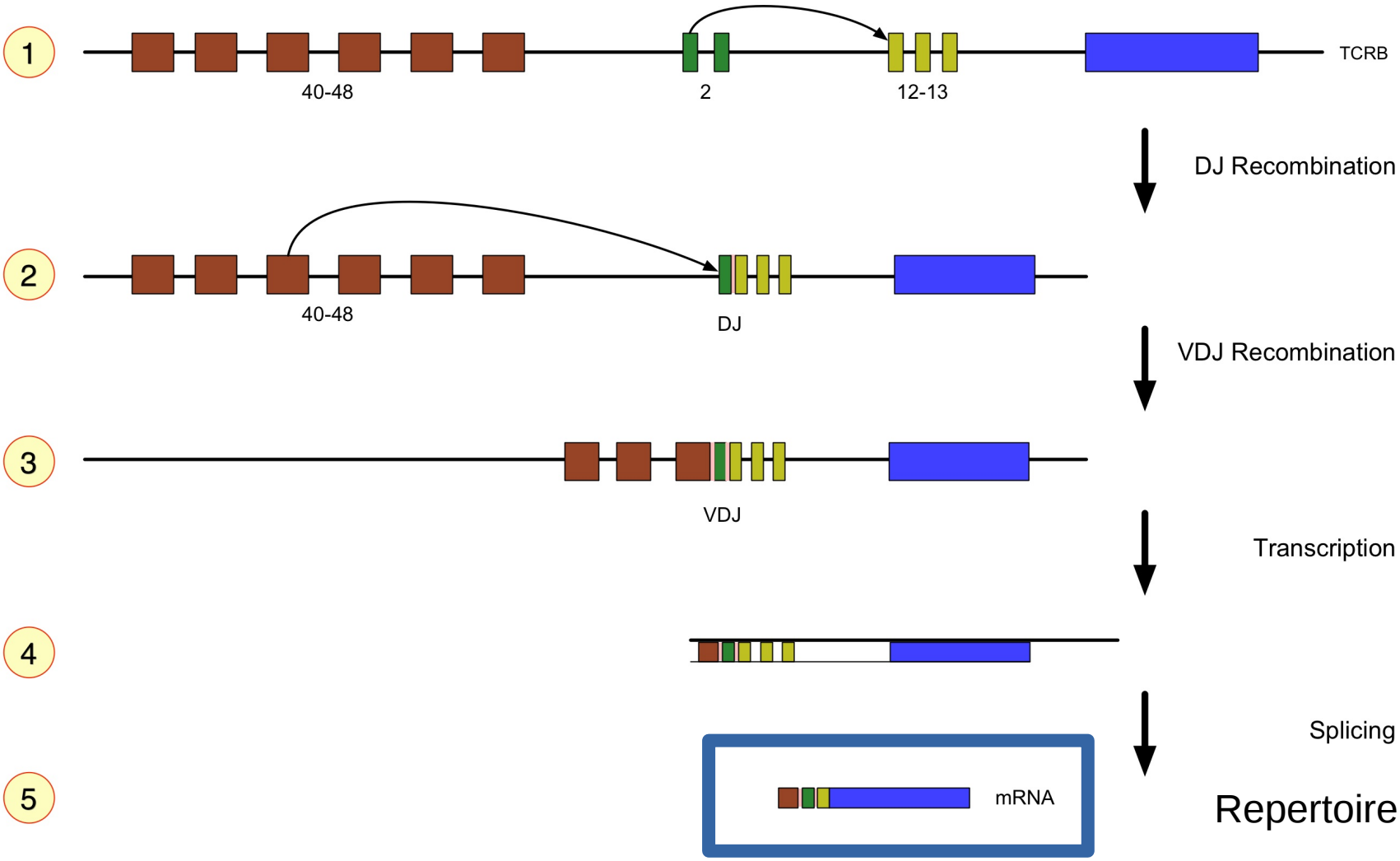
Human Antigen Receptor Loci



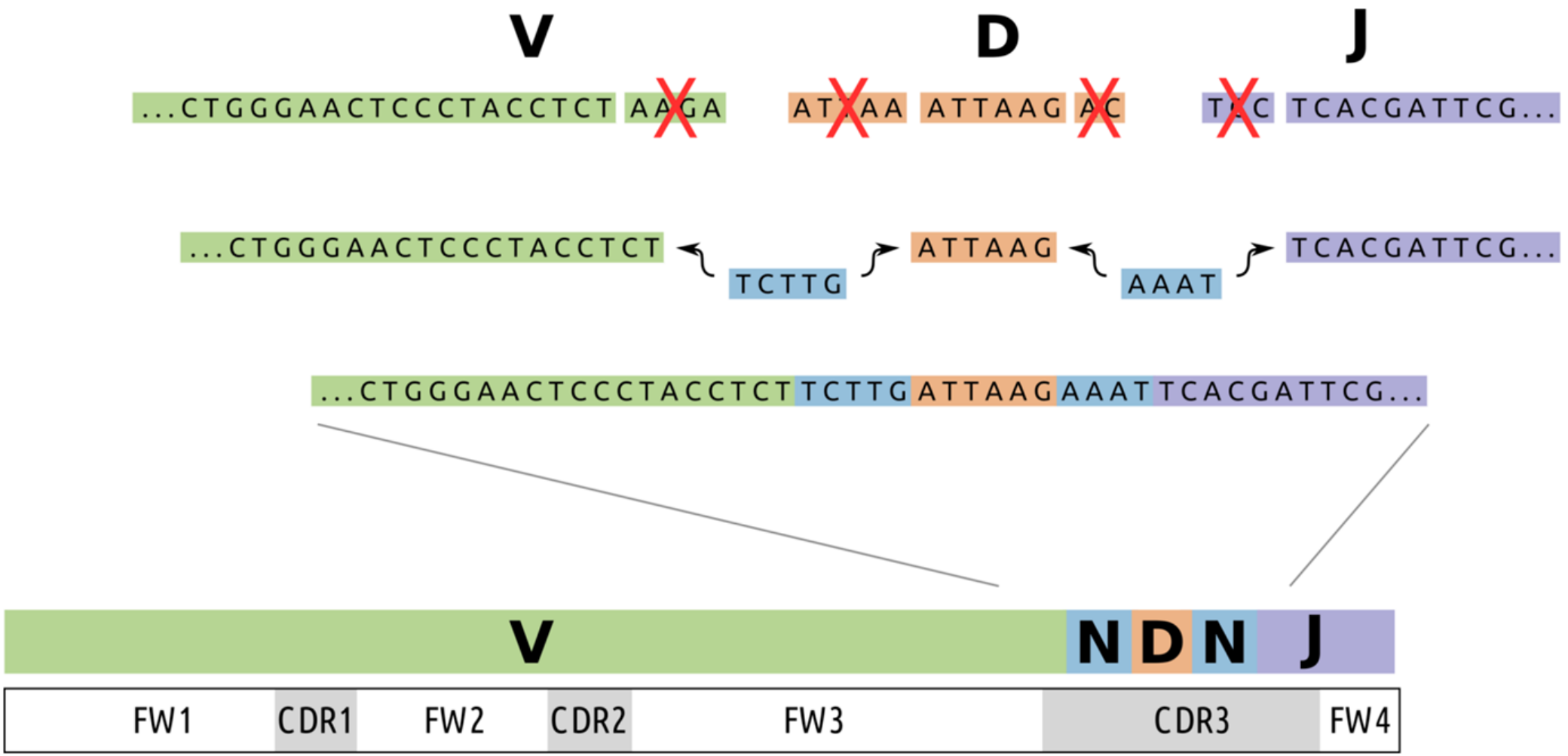
https://digitalworldbiology.com/blog/immunoprofiling-how-it-works

Antigen Receptor Formation

(general steps)

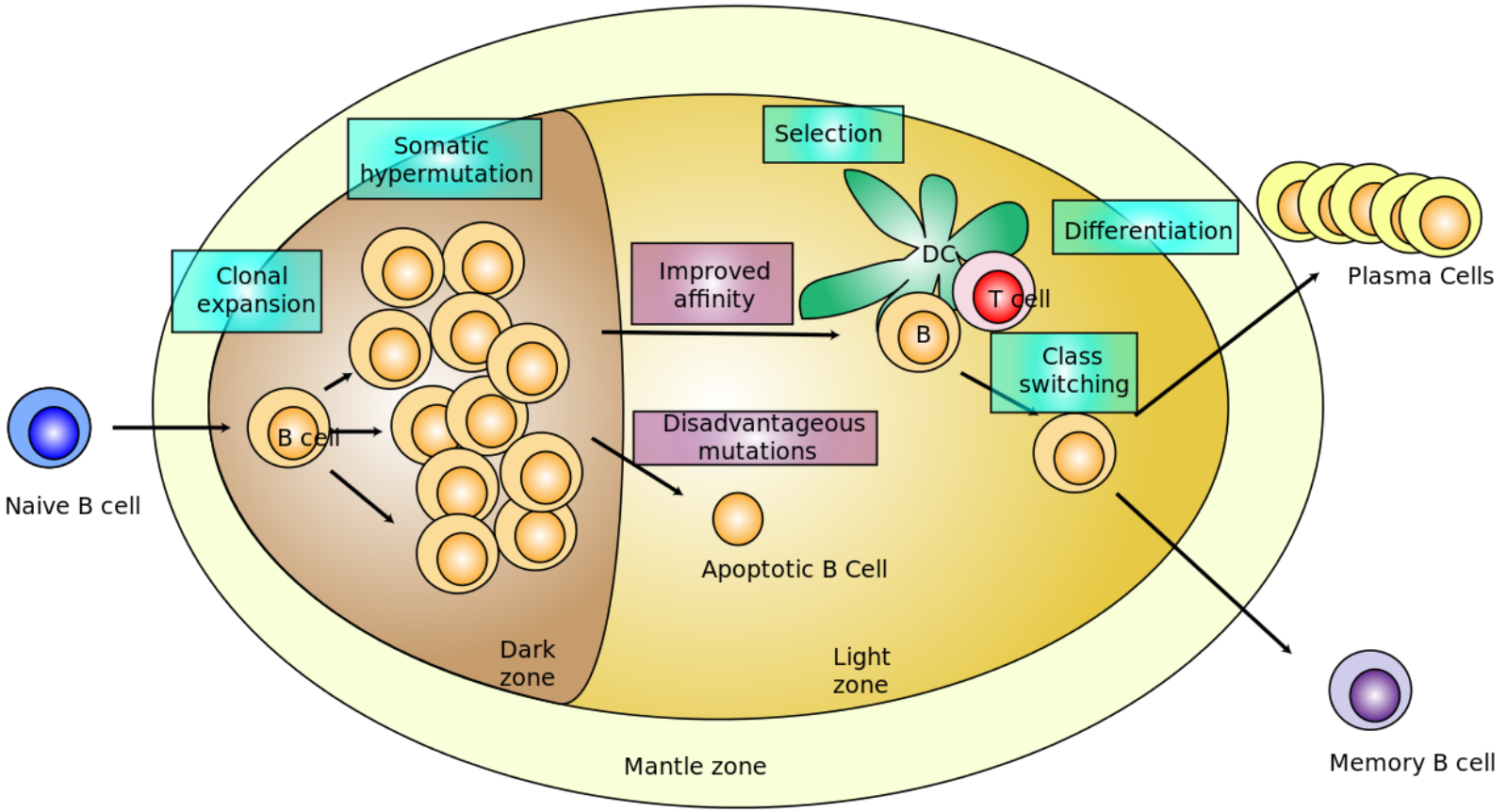


https://digitalworldbiology.com/blog/immunoprofiling-how-it-works



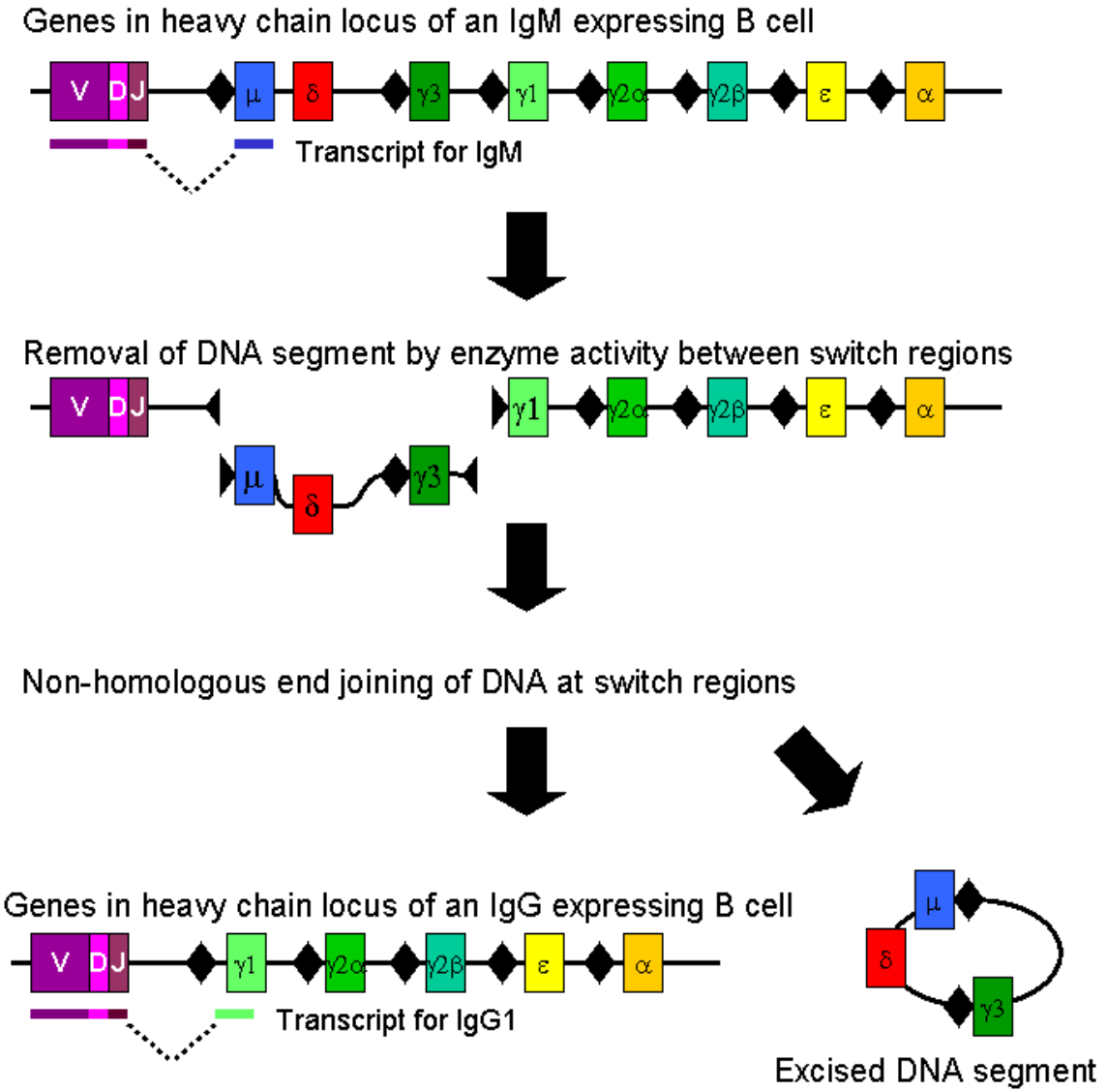
Ralph et al (2016), PloS Computational Biology

Clonal expansion and somatic hypermutation of B cells



Wikipedia

Class switching

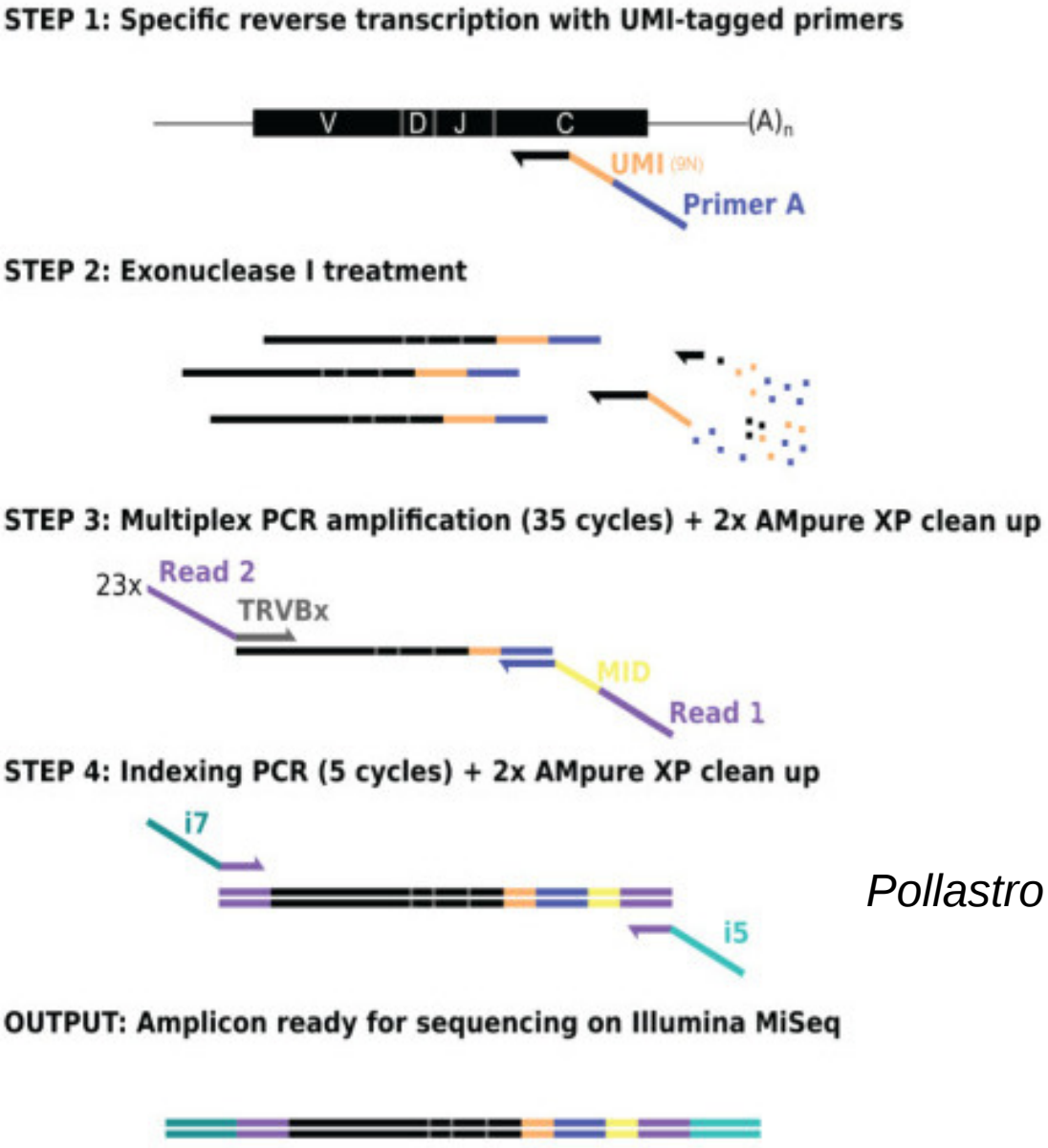


Wikipedia

Repertoire sequencing

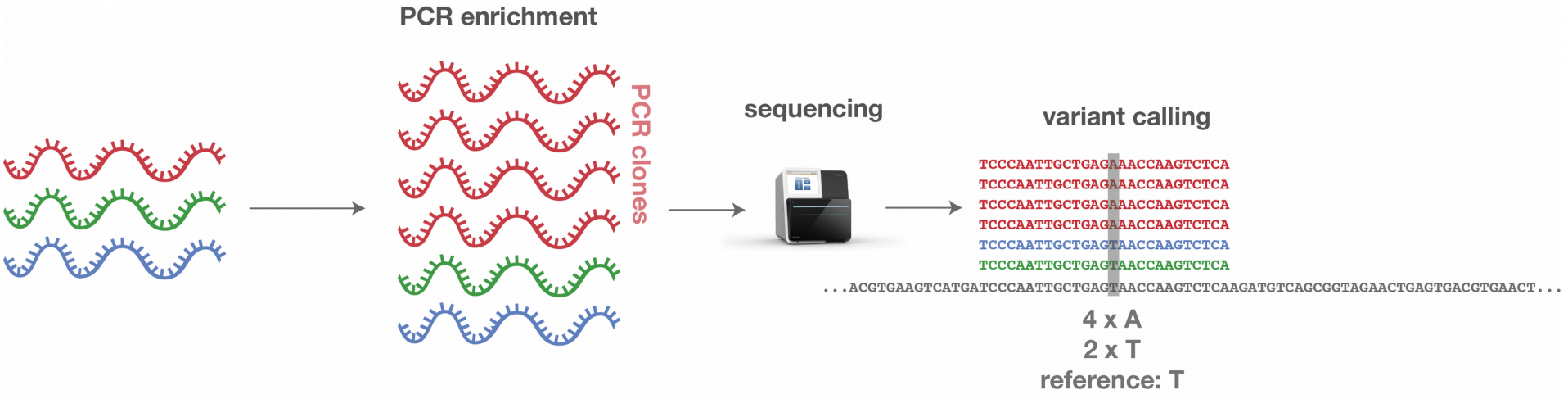
- DNA
- RNA
- Bulk (one chain)
- Single cell

One of the protocols



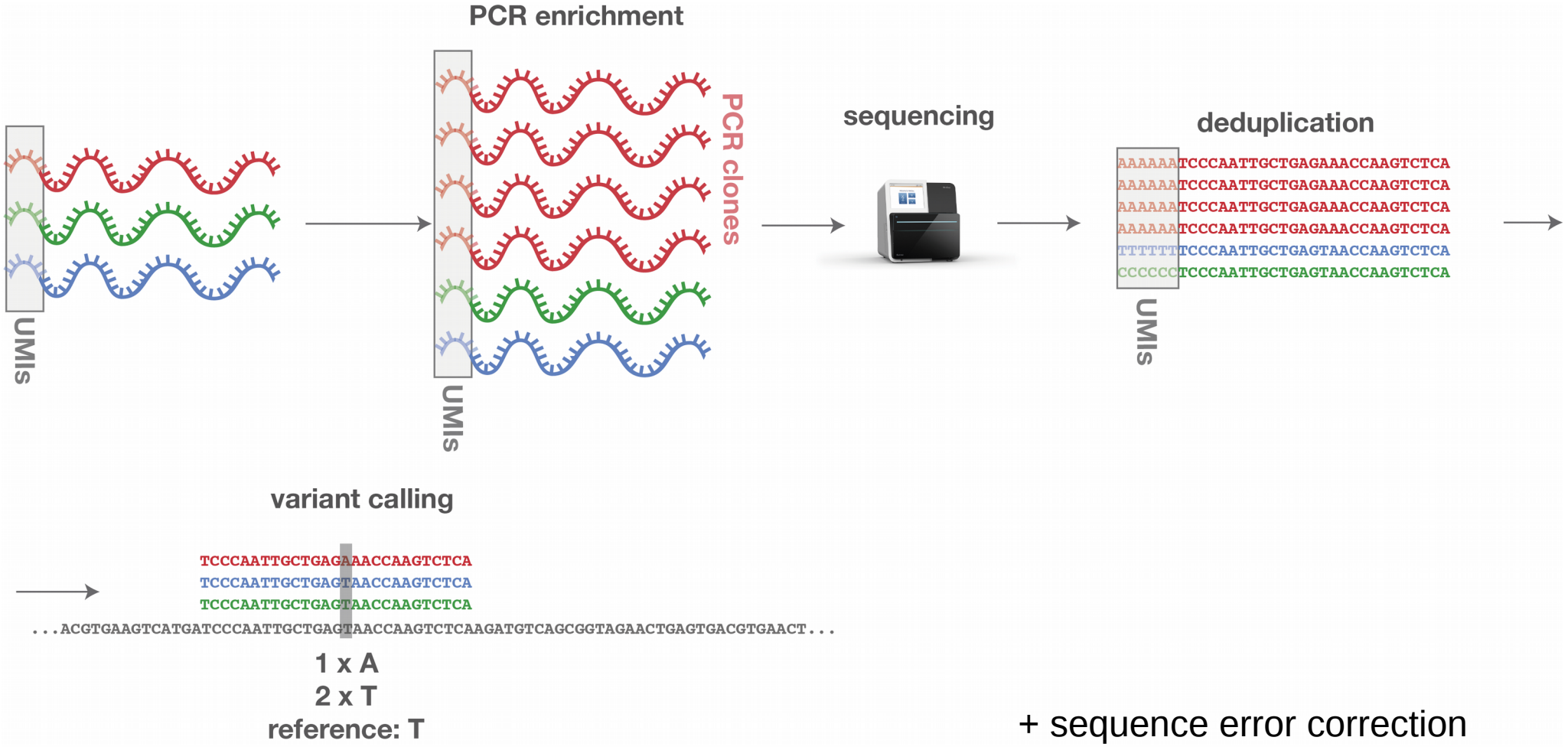
Pollastro et al (2021), Front Immunol.

Unique Molecular Identifiers (UMIs)



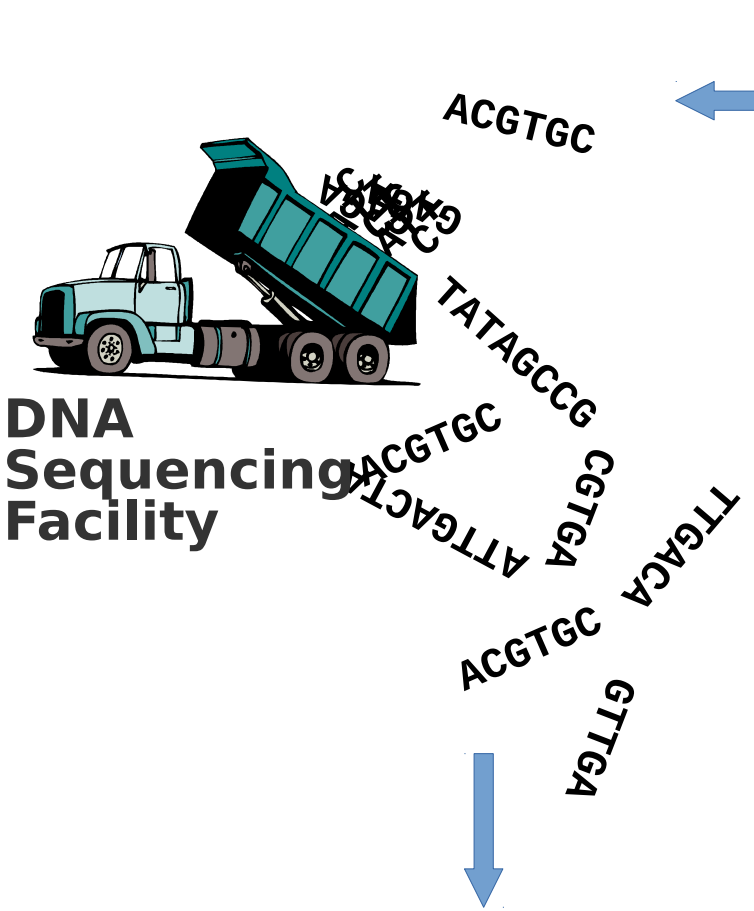
https://www.ecseq.com/support/ngs/how-can-unique-molecular-identifiers-help-to-reduce-quantitative-biases

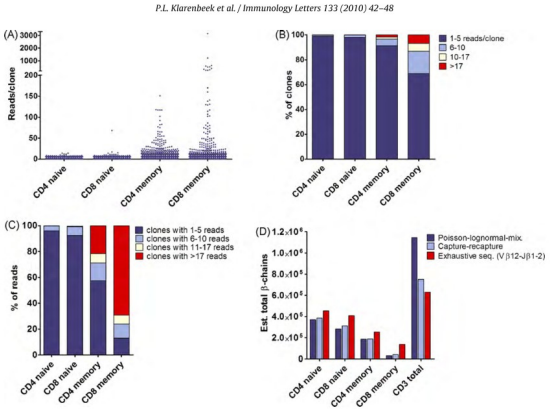
Unique Molecular Identifiers (UMIs)



+ sequence error correction

https://www.ecseq.com/support/ngs/how-can-unique-molecular-identifiers-help-to-reduce-quantitative-biases

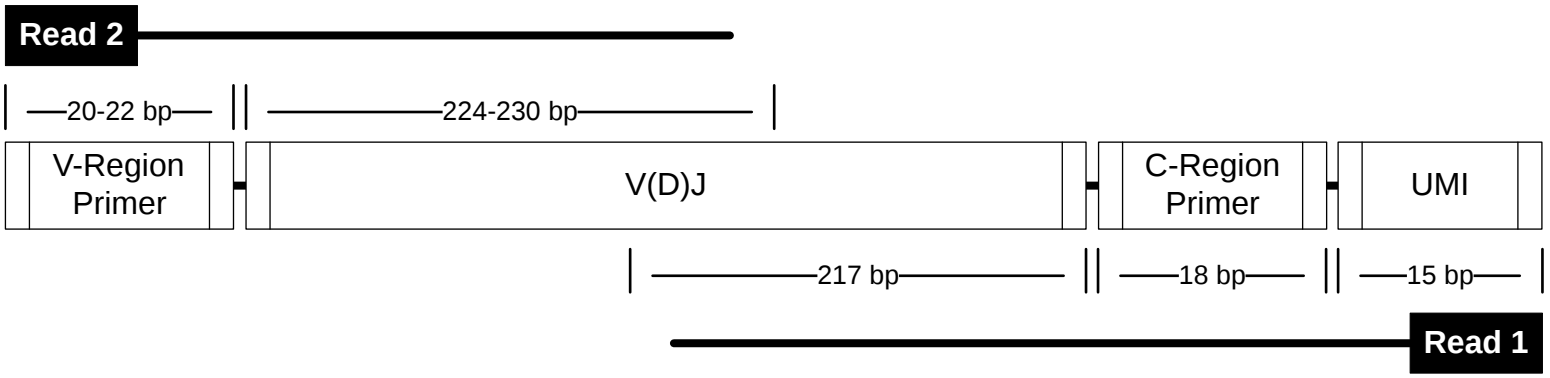

ACGTGC
TATAGCCG
CGTGA
ATTGACTA
ACGTGC
TTGACA
ACGTGC
GTTGA
DNA
Sequencing
Facility


P.L. Klarenbeek et al. / Immunology Letters 133 (2010) 42–48
(A)
Reads/clone
(B)
% of clones
1-5 reads/clone
6-10
10-17
>17
CD4 naive
CD8 naive
CD4 memory
CD8 memory
(C)
% of reads
clones with 1-5 reads
clones with 6-10 reads
clones with 11-17 reads
clones with >17 reads
(D)
Est. total β-chains
Poisson-lognormal-mix.
Capture-recapture
Exhaustive seq. (V β12-Jβ1-2)
CD3 total

Data analysis

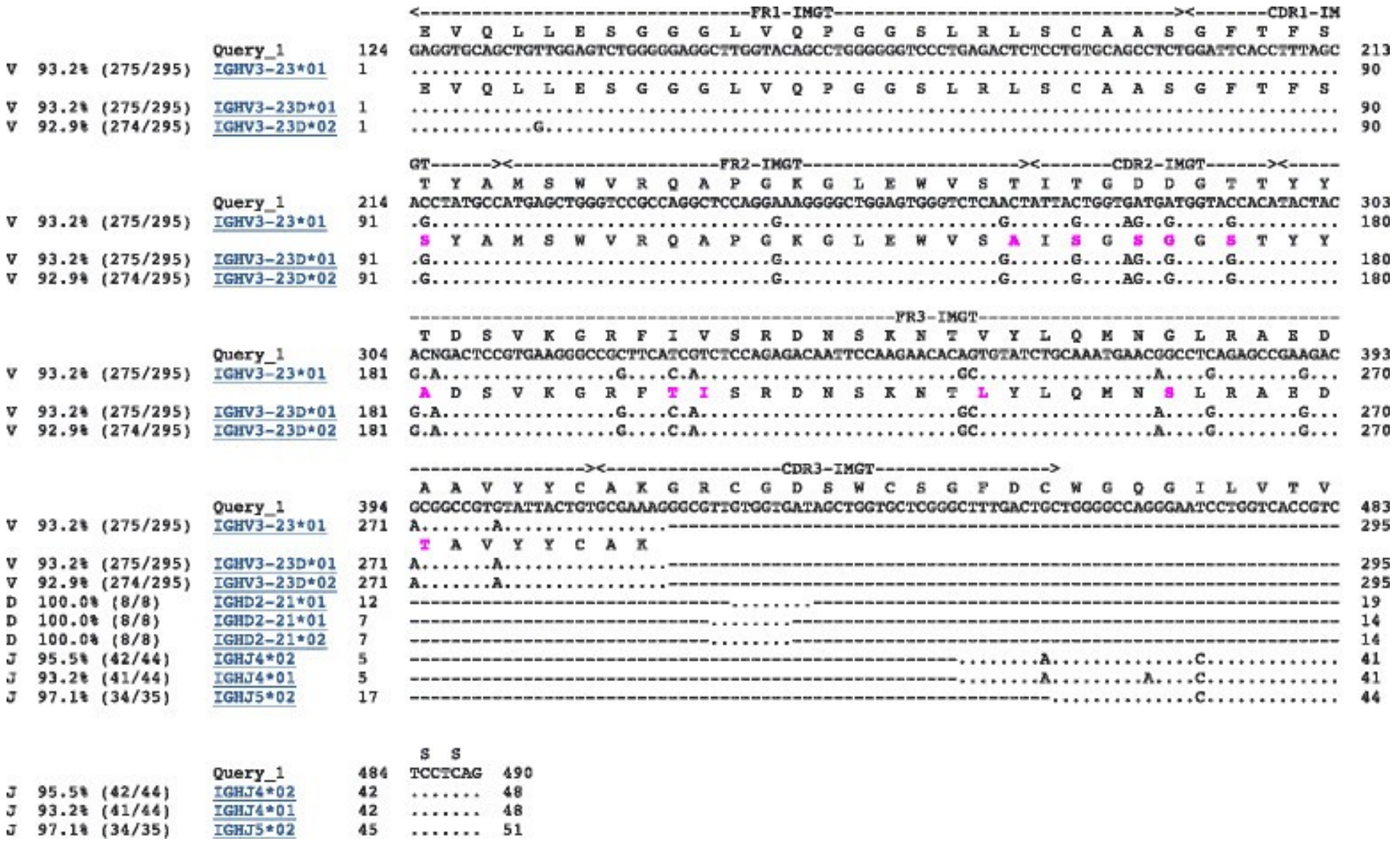
Pre-processing

- Identify barcodes
- Identify UMIs
- Mask primer sequences



https://presto.readthedocs.io/en/stable/workflows/Stern2014_Workflow.html

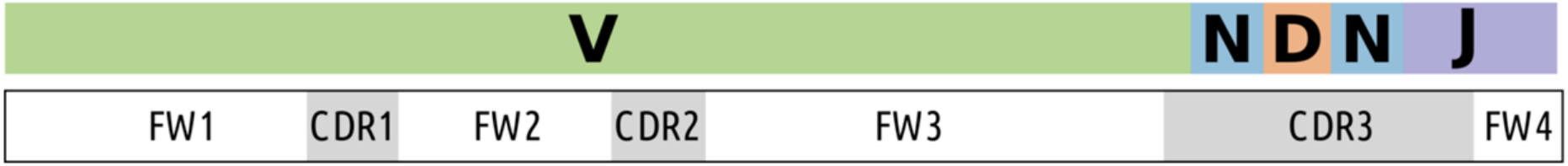
Identify V, (D), J and C regions



Toby et al (2016), BMC Bioinformatics

Identify CDR3 sequence

- Two common methods for this
  - Deduce from alignment
  - Search for conserved motifs at end of V and start of J region



Define (sub)clones

Sub clones

V, J and CDR3-peptides

V, J and CDR3-nucleotides

Entire nucleotide sequence

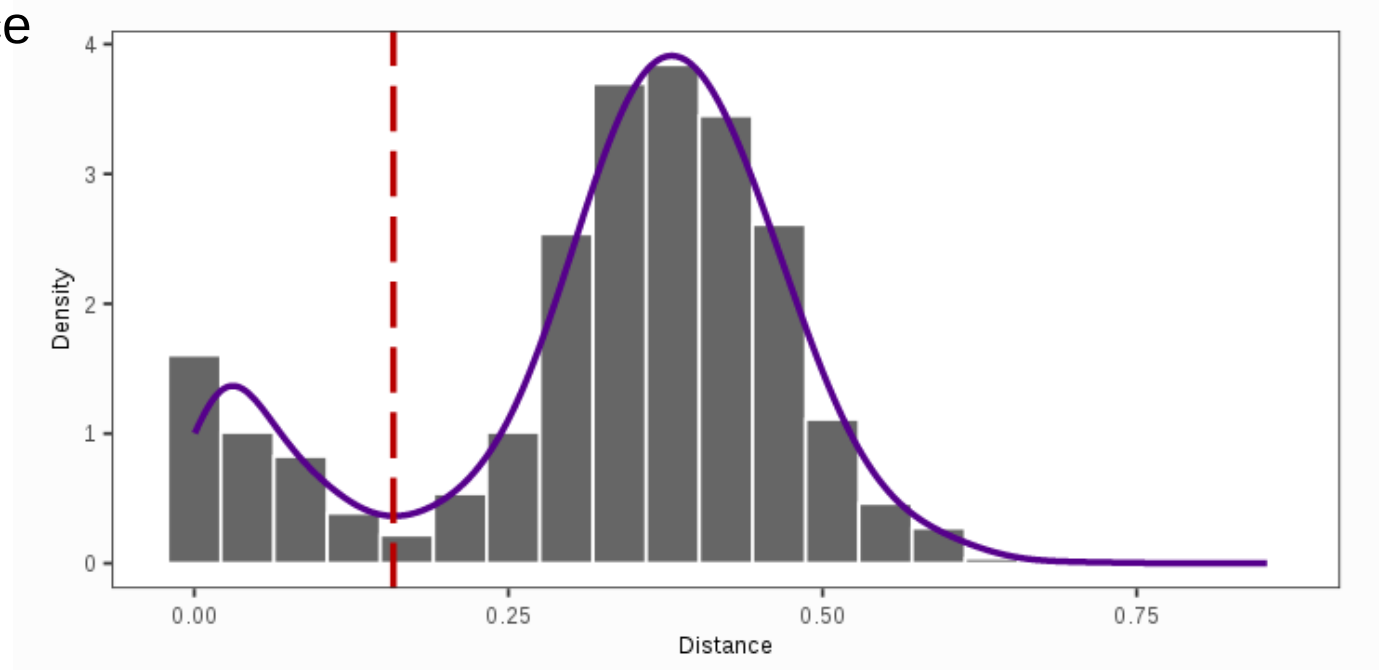
Entire protein sequence

Clonal families (clones)

1 amino acid difference

85% sequence identity

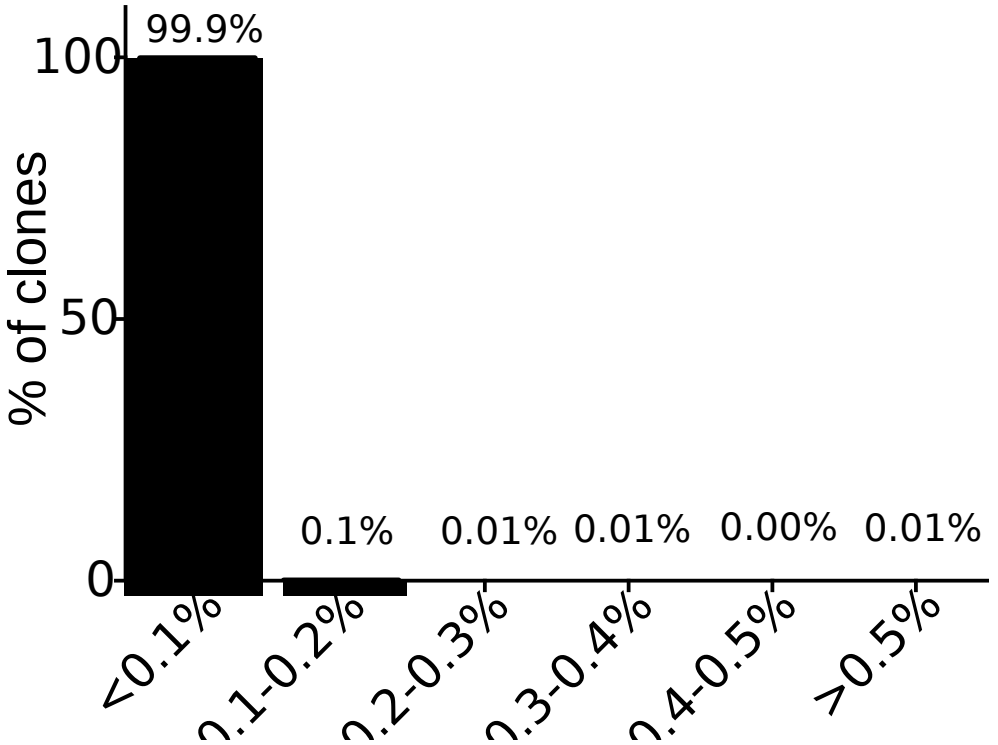
Dynamic threshold



https://changeo.readthedocs.io/en/stable/examples/cloning.html

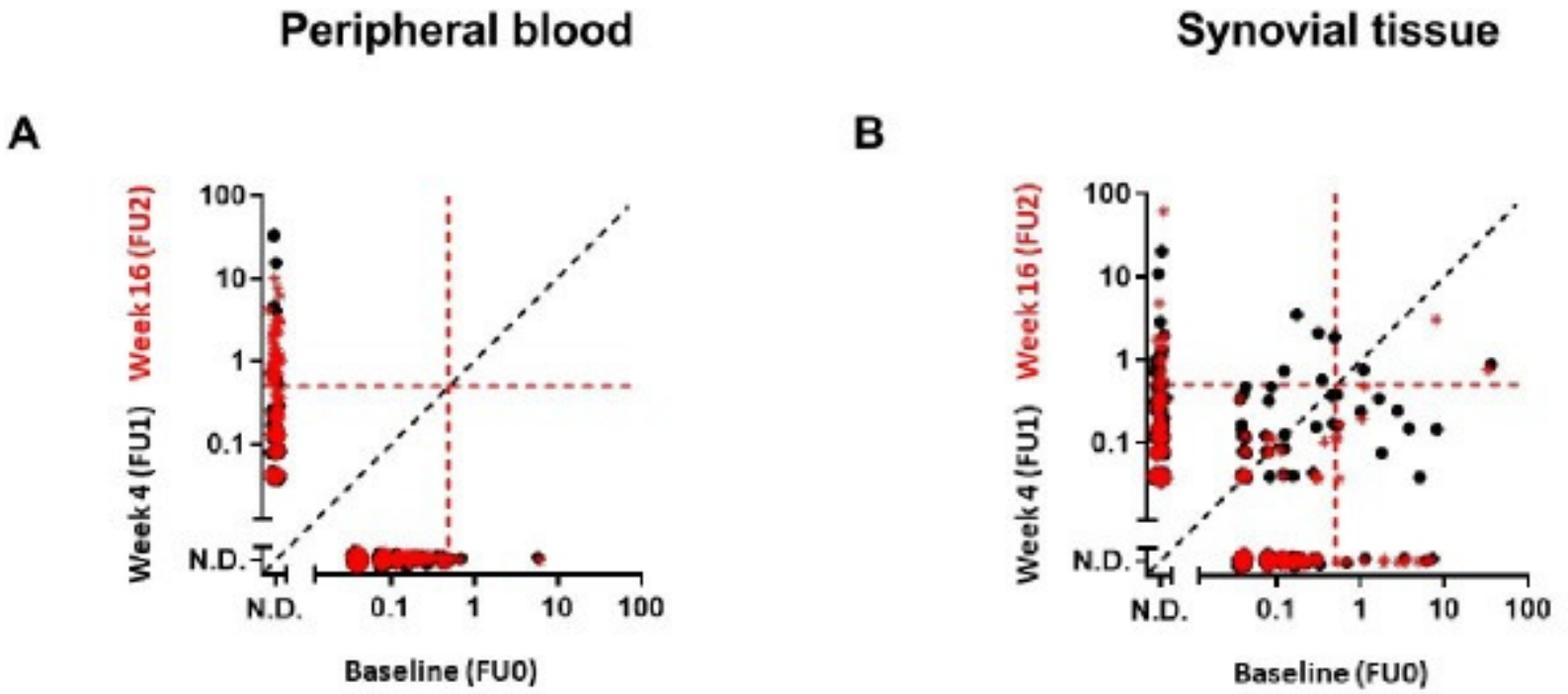
Follow-up analysis

Dominant clones



- Defined as clones with percentage reads (or UMIs) > 0.5% compared to total sample
- This threshold was supported by naive sorted T cells

Compare samples



Pollastro et al (2019), ARD

Diversity

Diversity

Many different plants
+
Equal frequencies
= diverse

Species richness
Evenness

Few types of plants
+
Dominated by lawn
= less diverse

Diversity indices

Shannon entropy

$$H_1 = -\sum_s p(s) \ln p(s).$$

Renyi entropy

$$H_\beta = \frac{1}{1-\beta} \ln \left[\sum_s p(s)^\beta \right]$$

Hill diversity

$$D_\beta = \exp[H_\beta].$$

Inverse Simpson

$$D_2 = 1/\sum_s p(s)^2.$$

Gini-Simpson

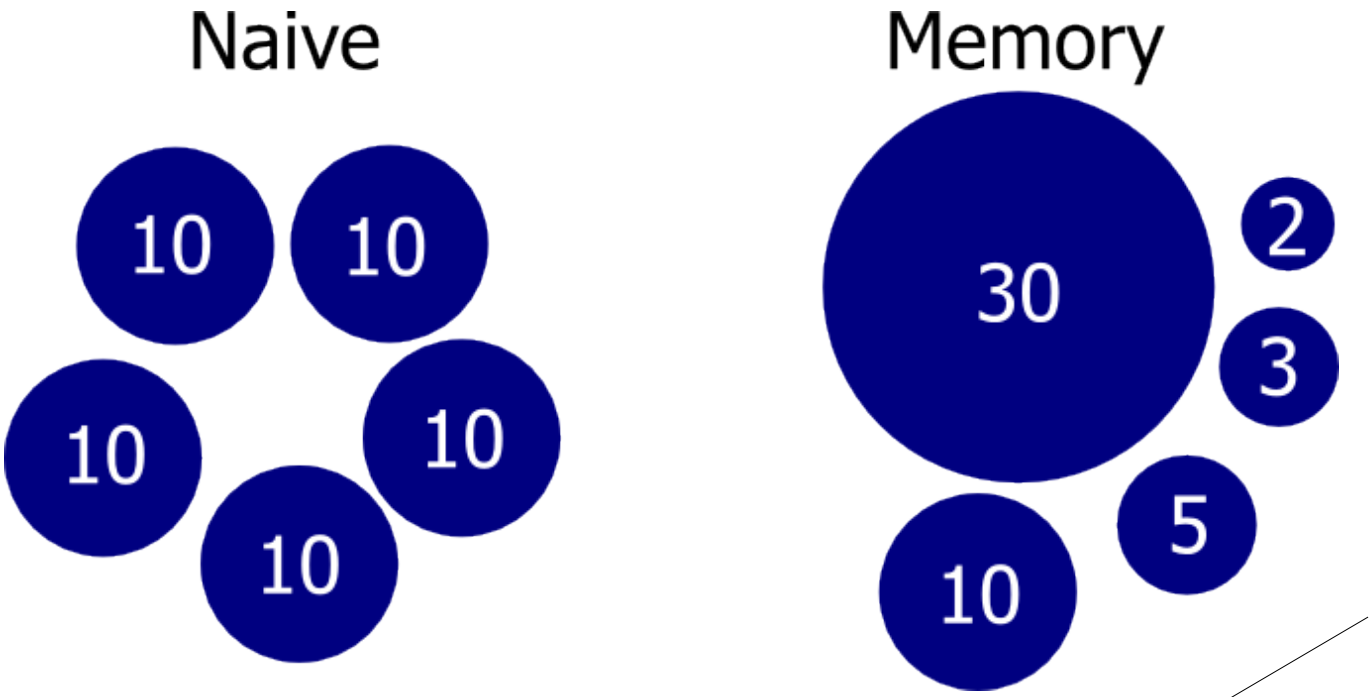
$$1 - 1/D_2.$$

p(s) is probability, frequency or abundance of species

D0 is species richness

Mora and Walczak, 2016

Example immune cells



Inverse Simpson

$D_2 = 1/\sum_s p(s)^2.$

Simpson's index

$H_2 = 5 * (10/50)^2 = 0.20$

Inverse Simpson

$D_2 = 1/0.20 = 4.99999$

Simpson's index

$H_2 = (30/50)^2 + (10/50)^2 + (5/50)^2 + (3/50)^2 + (2/50)^2 = 0.4152$

Inverse Simpson

$D_2 = 1/0.4152 = 2.4085$

Chance that you draw two similar species

Naive more diverse than memory compartment

Generalized formula for entropy and diversity

- ß = 0 : insensitive to frequency
- ß = low: emphasis on rare species
- ß = high: emphasis on dominant species

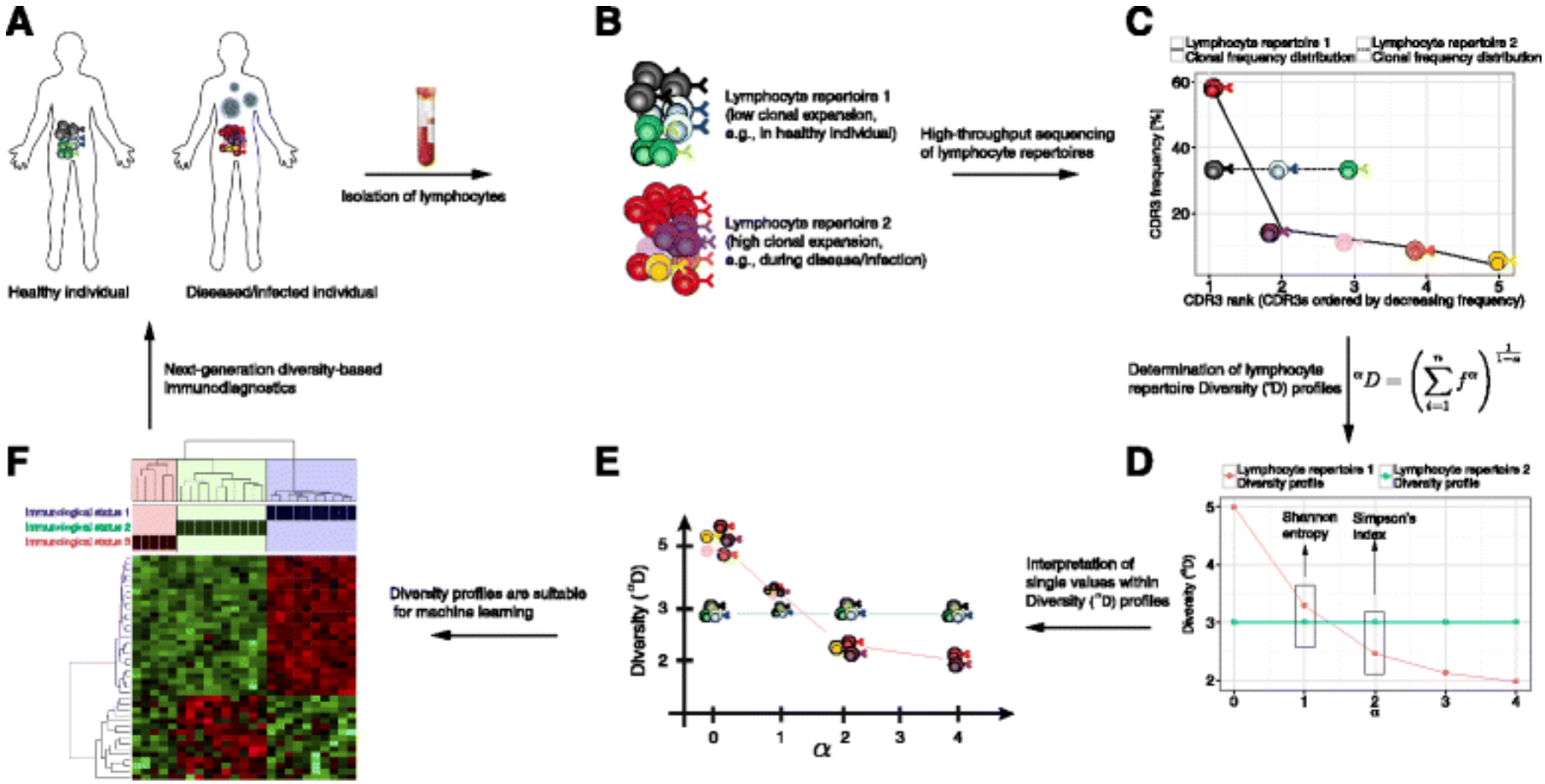
$$H_\beta = \frac{1}{1-\beta} \ln \left[\sum_s p(s)^\beta \right]$$

Renyi

$$D_\beta = \exp[H_\beta].$$

Hill diversity

Diversity profiles

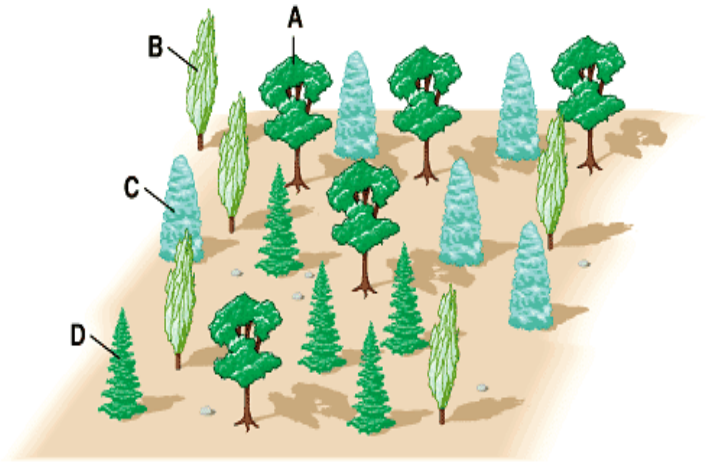


Greiff et al (2015), Genome Medicine

Shared clones

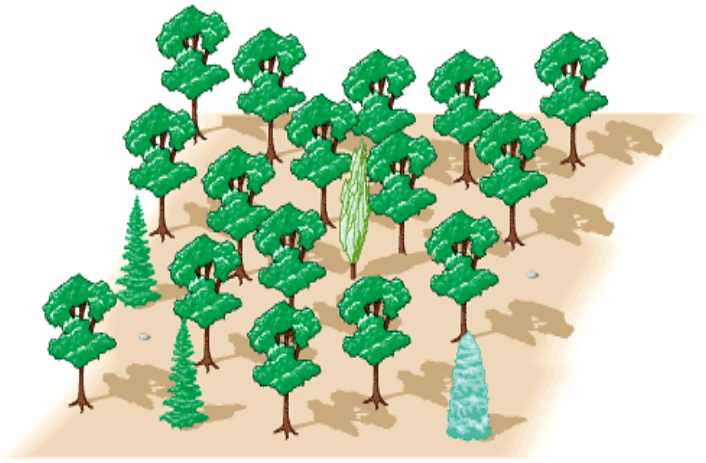
Similarity indices

- Inverse of distance measure
- E.g. determine overlap in species richness (species count, without abundance)
- Several similarity indices exist that do take abundance into account



Community 1

A: 25% B: 25% C: 25% D: 25%

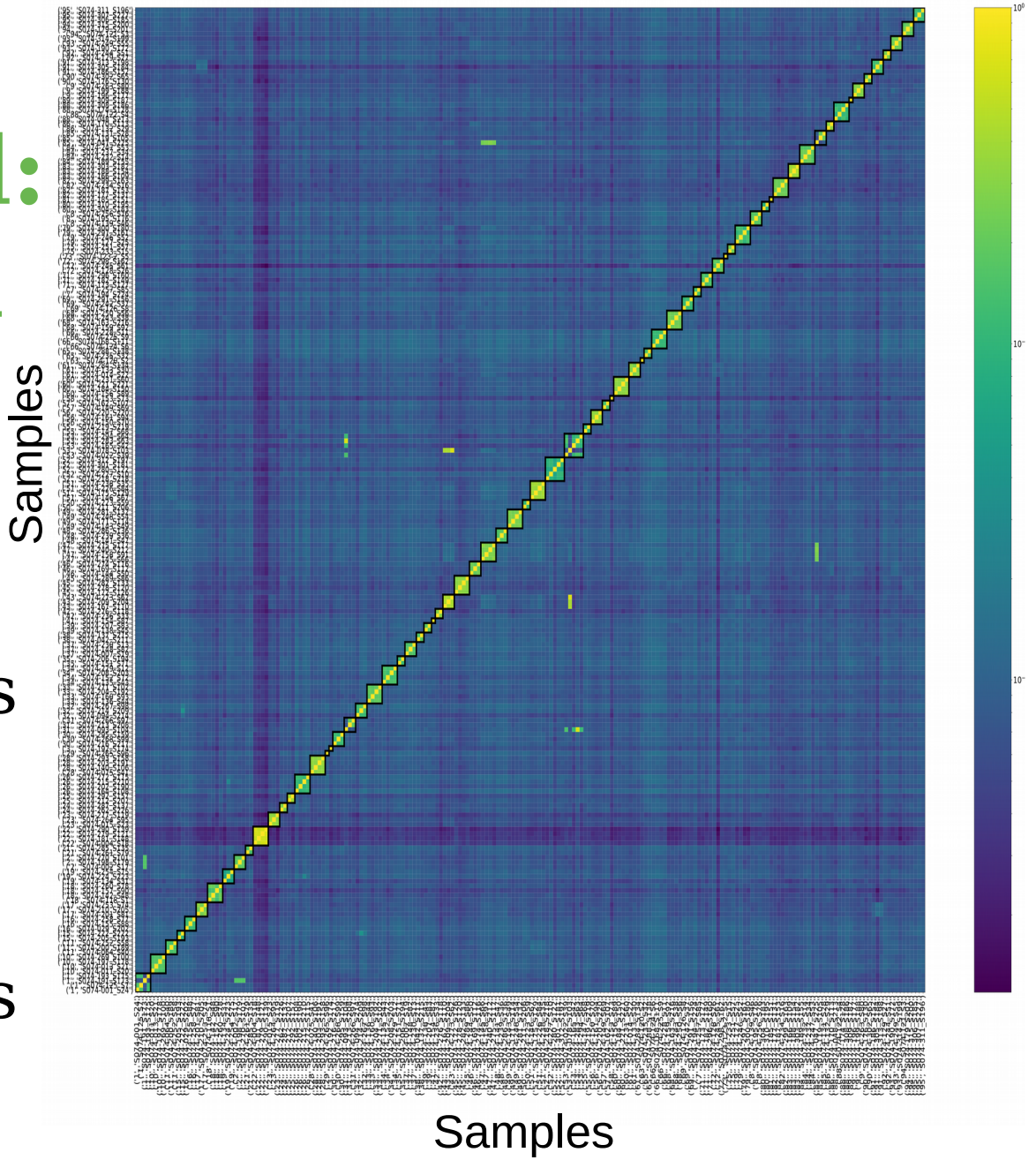


Community 2

A: 80% B: 5% C: 5% D: 10%

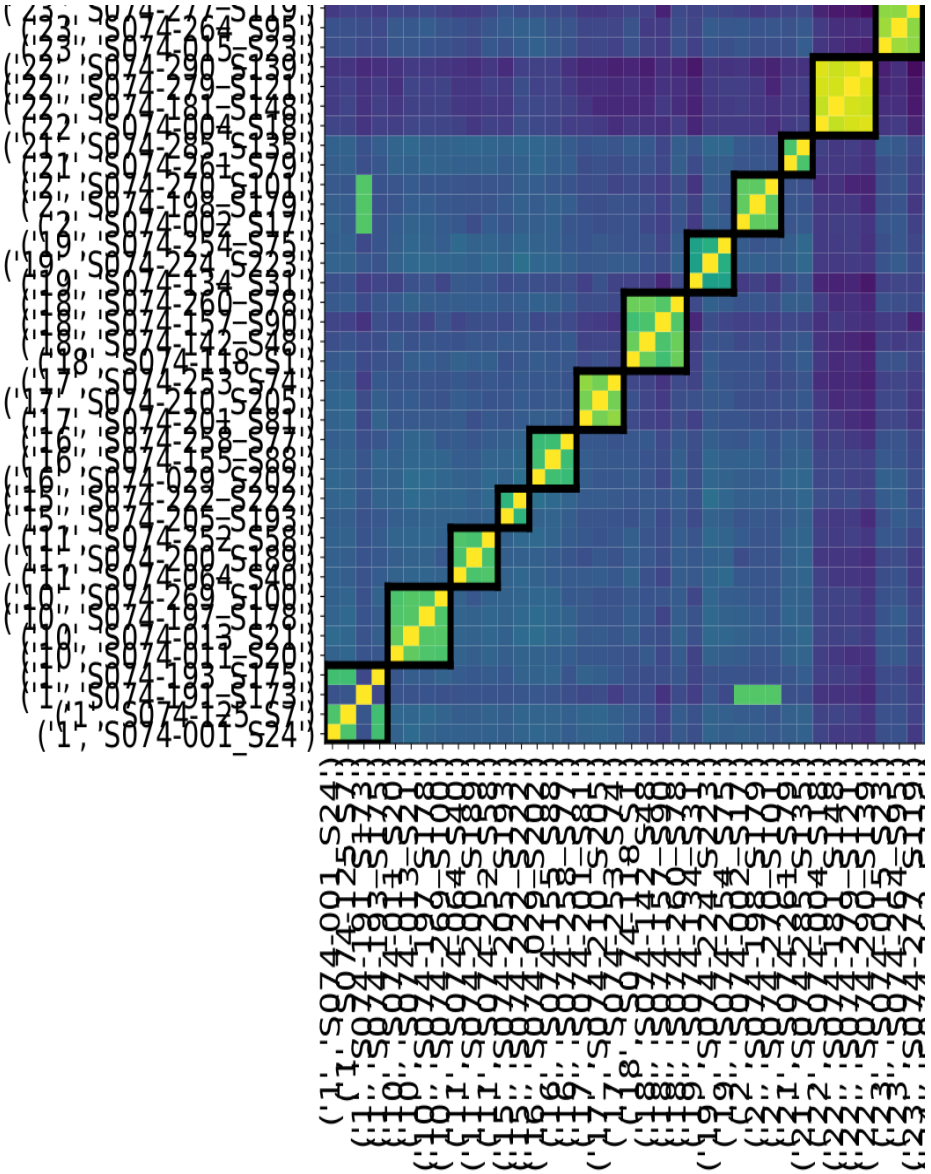
Quality control: Contamination (similarity)

- Pairwise comparison between all samples
- 0=no overlap at all (dark blue)
- 1=identical samples (yellow)

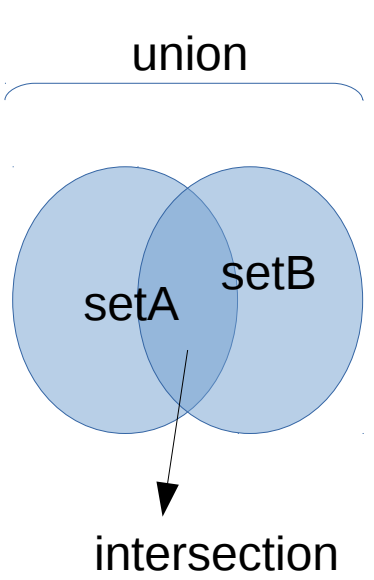


Quality control: Contamination (similarity)

- Black squares: samples from same patient
- Similarity is expected within patient
- Similarity across patients indicate problems (contamination, labelling errors, etc)



Similarity indices



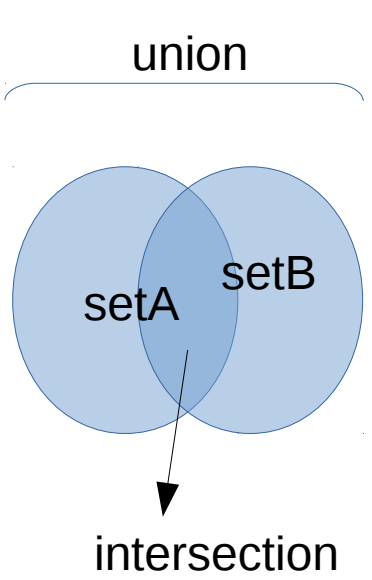
Jaccard index (occurrence of species)
J(A,B) = intersection_AB / (union_AB – intersection_AB)

Sorensen index (occurrence of species)
S(A,B) = 2 * intersection_AB / (set_A + set_B)

Bray-Curtis index (occurrence and frequency of species)
$BC_{ij} = 2 * C_{ij} / (S_i + S_j)$

C_{ij} is sum of values for species in common between both sites
S_i is total number of species at site i
S_j is total number of species at site j

Look at shared clones

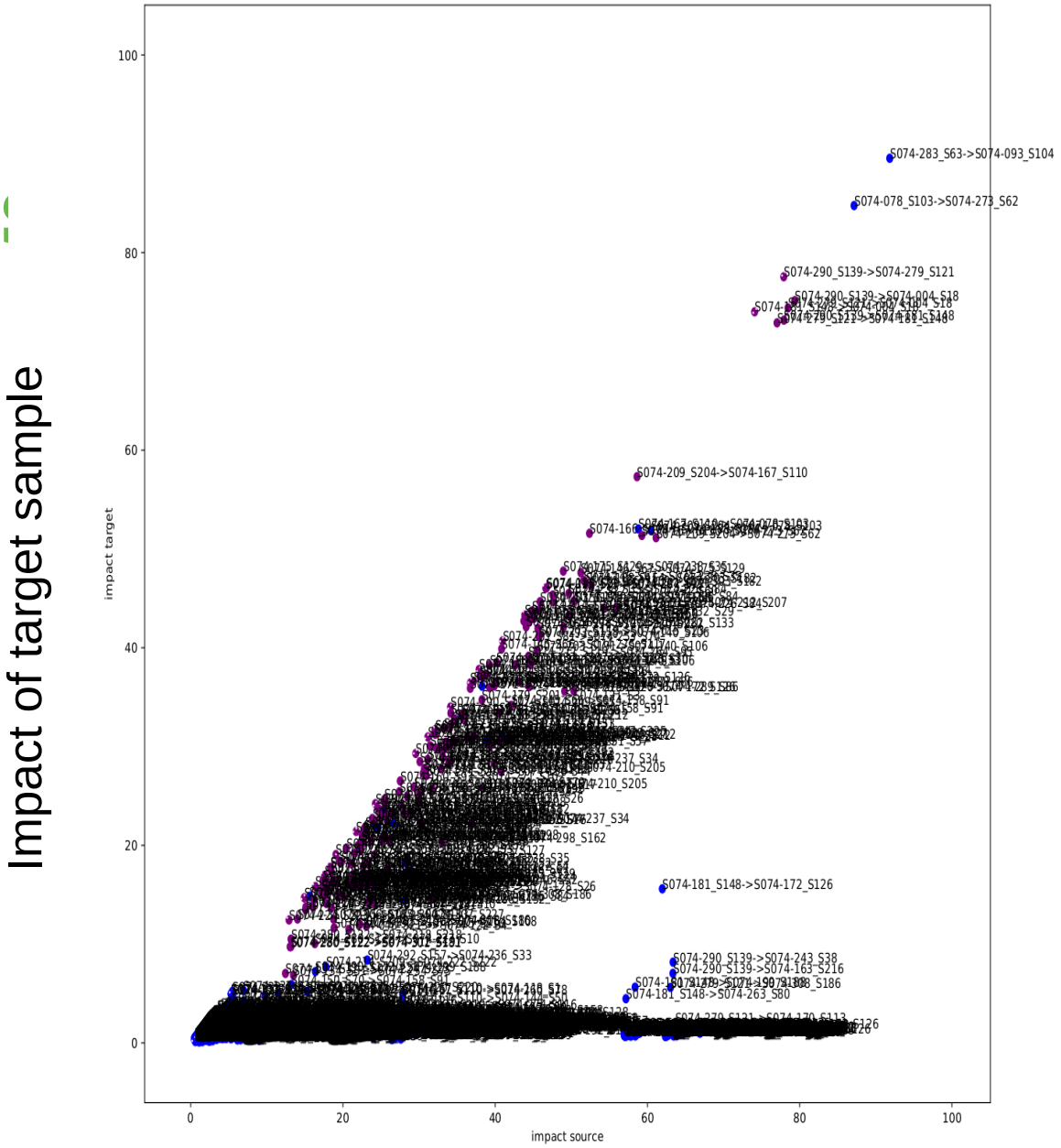


Compare all samples pairwise
Look for shared clones (same CDR3 sequences) in both samples
Calculate impact of the shared clones in both samples:
    Impact = sum(frequency) / total frequency in sample

Assumption:
    Contamination direction is from sample with highest impact to lowest impact

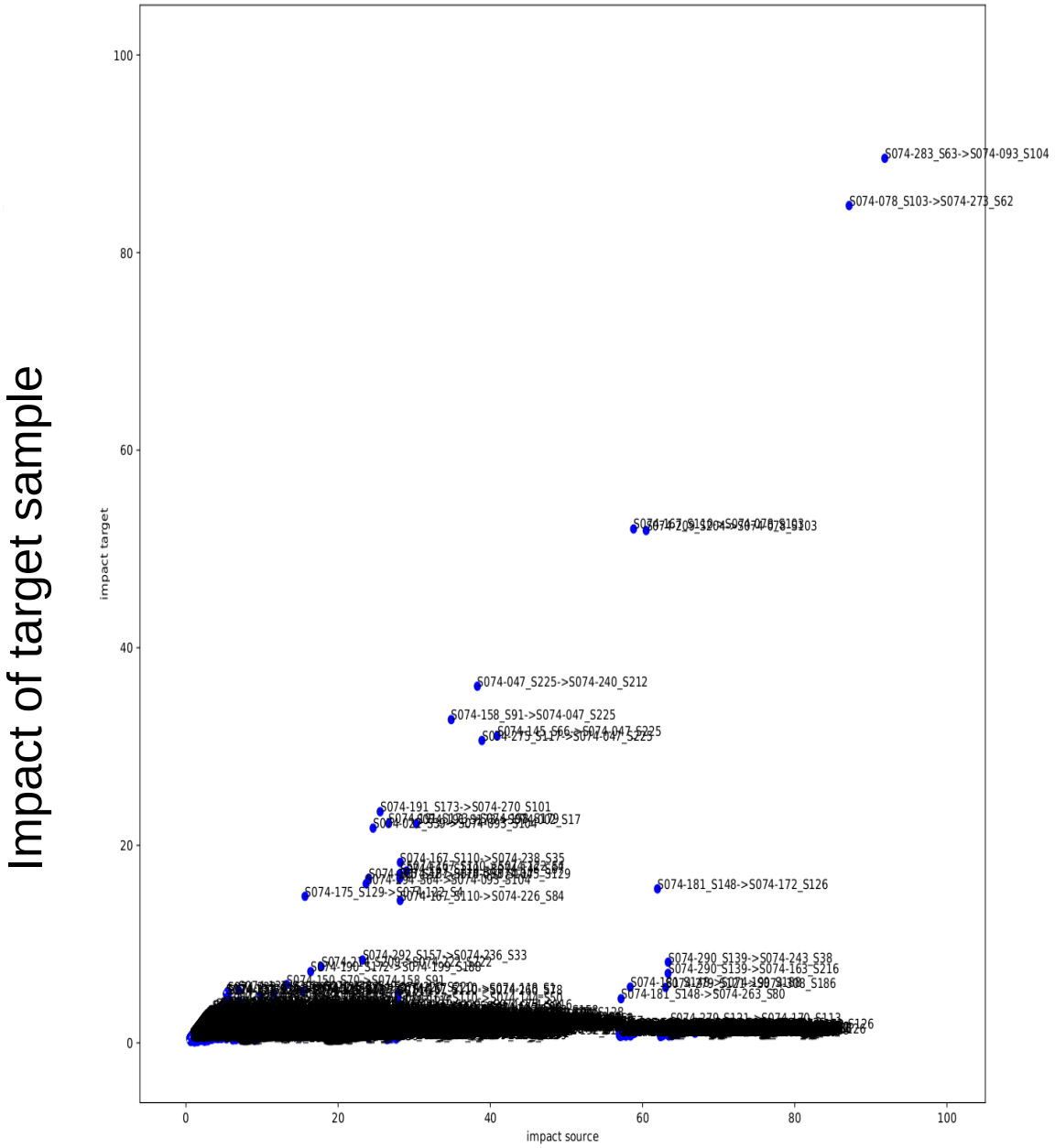
Shared clones

Purple = same patient
Blue = different patient

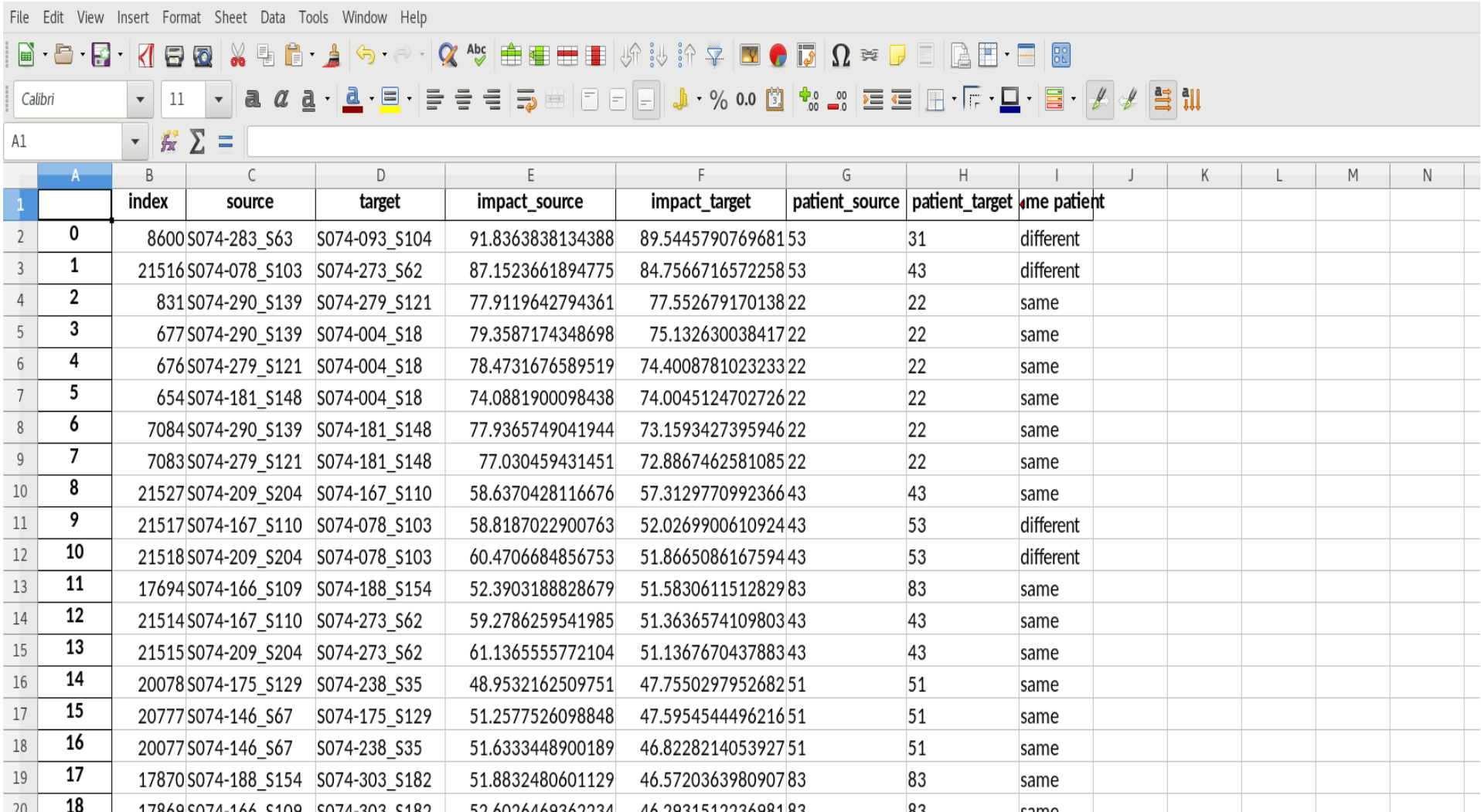


Shared clone

Only show
different patient pair



Shared clones



File Edit View Insert Format Sheet Data Tools Window Help

Calibri 11 A1

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1		index	source	target	impact_source	impact_target	patient_source	patient_target	me patient					
2	0	8600	S074-283_S63	S074-093_S104	91.8363838134388	89.54457907696	81 53	31	different					
3	1	21516	S074-078_S103	S074-273_S62	87.1523661894775	84.7566716572258	53	43	different					
4	2	831	S074-290_S139	S074-279_S121	77.9119642794361	77.5526791701382	2	22	same					
5	3	677	S074-290_S139	S074-004_S18	79.3587174348698	75.1326300384172	2	22	same					
6	4	676	S074-279_S121	S074-004_S18	78.4731676589519	74.4008781023233	2	22	same					
7	5	654	S074-181_S148	S074-004_S18	74.0881900098438	74.0045124702726	2	22	same					
8	6	7084	S074-290_S139	S074-181_S148	77.9365749041944	73.1593427395946	2	22	same					
9	7	7083	S074-279_S121	S074-181_S148	77.030459431451	72.8867462581085	2	22	same					
10	8	21527	S074-209_S204	S074-167_S110	58.6370428116676	57.3129770992366	43	43	same					
11	9	21517	S074-167_S110	S074-078_S103	58.8187022900763	52.0269900610924	43	53	different					
12	10	21518	S074-209_S204	S074-078_S103	60.4706684856753	51.8665086167594	43	53	different					
13	11	17694	S074-166_S109	S074-188_S154	52.3903188828679	51.5830611512829	83	83	same					
14	12	21514	S074-167_S110	S074-273_S62	59.2786259541985	51.3636574109803	43	43	same					
15	13	21515	S074-209_S204	S074-273_S62	61.1365555772104	51.1367670437883	43	43	same					
16	14	20078	S074-175_S129	S074-238_S35	48.9532162509751	47.7550297952682	51	51	same					
17	15	20777	S074-146_S67	S074-175_S129	51.2577526098848	47.5954544496216	51	51	same					
18	16	20077	S074-146_S67	S074-238_S35	51.6333448900189	46.8228214053927	51	51	same					
19	17	17870	S074-188_S154	S074-303_S182	51.8832480601129	46.5720363980907	83	83	same					
20	18	17869	S074-166_S109	S074-303_S182	52.6026469362234	46.2831512236981	83	83	same					

Computer clusters, cloud, linux

Handling sequencing data

PC

Small cluster

Option: buy a computer cluster

Option: cloud computing

Often you do not need computing power all the time
Compute by demand
Compute resources are shared

Computer exercises

- Data analysis on the SurfSara cloud
- Analyze a public AIRRseq dataset
  - Pairwise assembly
  - Identify V, J and CDR3
  - Group related sequences
  - Count dominant clones

https://www.surfsara.nl/

Overview

Introduction AIRRseq

* T- and B-cell receptors
* VDJ recombination
* Clonal expansion and somatic hypermutation
* Class switching
* Repertoire sequencing

Data analysis

* Pre-processing
* Identify V, (D), J and C regions
* Identify CDR3 region
* Determine clones and their frequency

Follow-up analysis

* Dominant clones
* Compare samples
* Diversity of samples
* Shared clones between samples

Computer systems for data analysis

* Clusters, cloud, linux
* Overview of computer lab