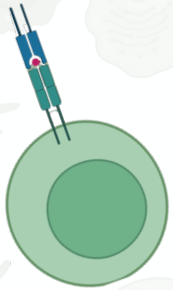
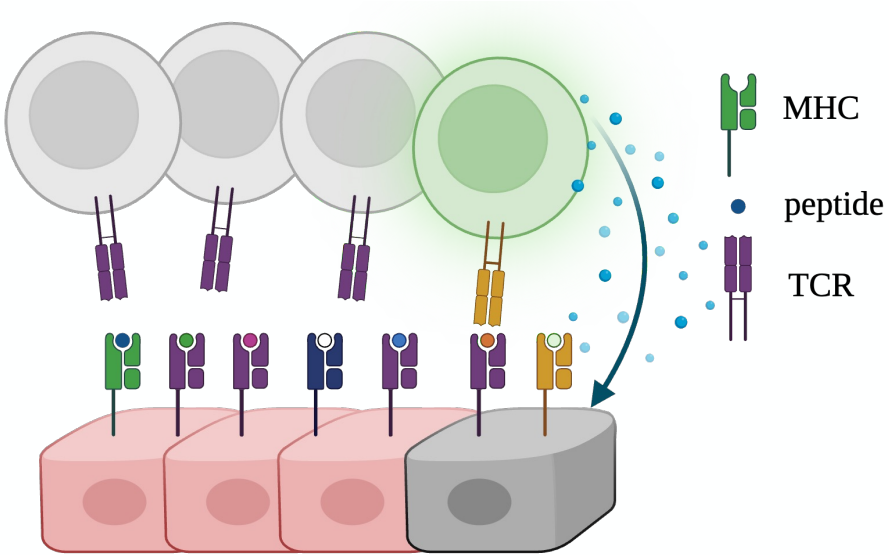


Machine learning models for TCR-epitope prediction

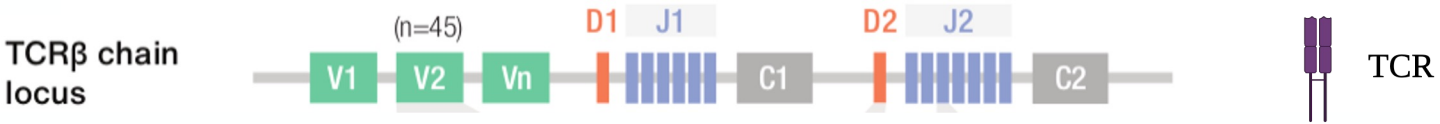
T-cell receptors recognize unique molecular surface structures



- Adaptive immune system
- Recognise diverse antigens
  - Peptides
  - Lipids
  - Small molecules

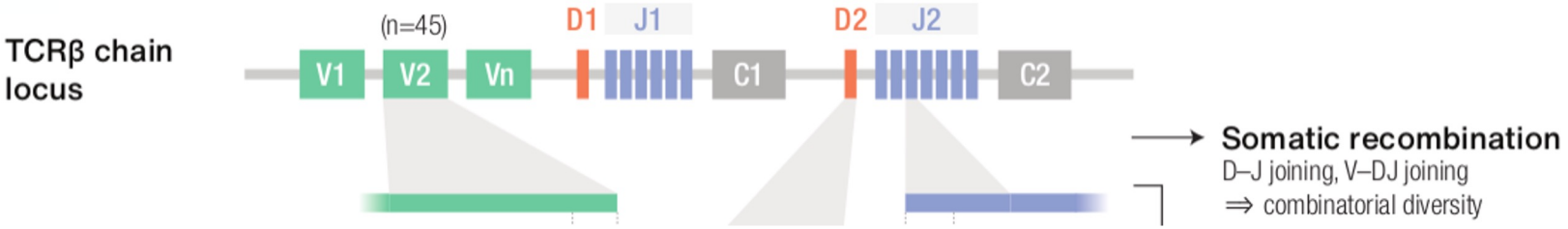
TCR recombination gives rise to highly diverse and unique repertiores

For each T-cell, a unique receptor is quasi-randomly generated during a process called V(D)J recombination



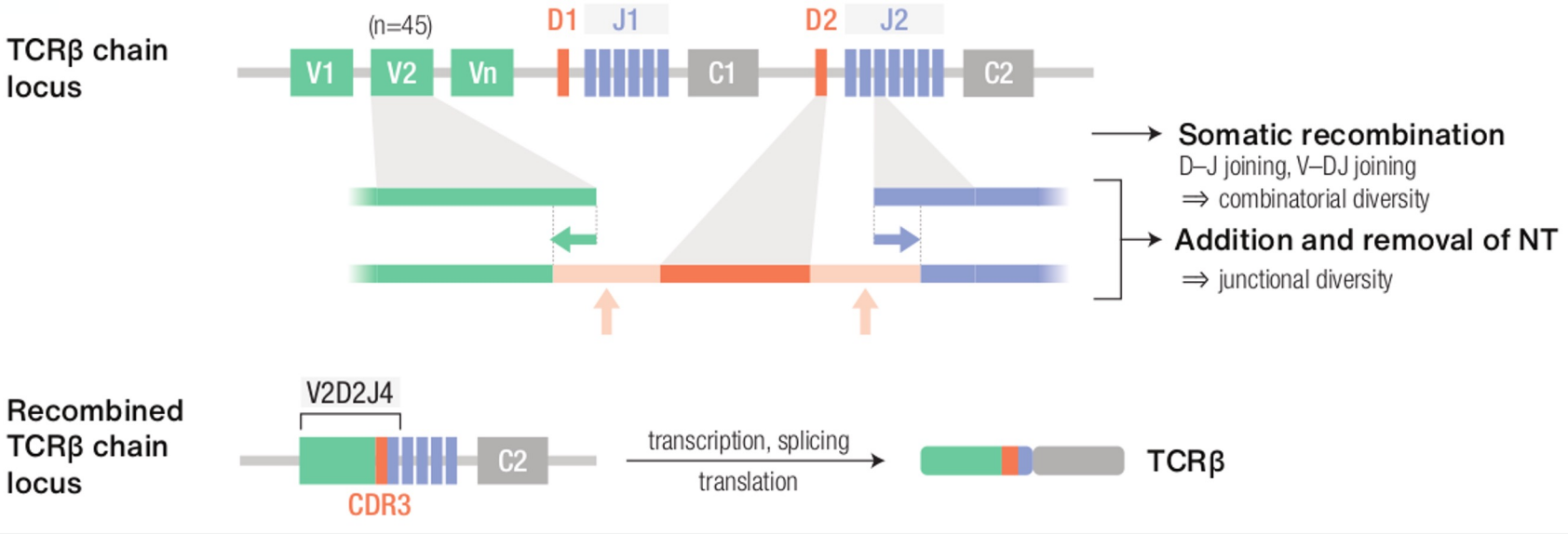
TCR recombination gives rise to highly diverse and unique repertiores

For each T-cell, a unique receptor is quasi-randomly generated during a process called V(D)J recombination



TCR recombination gives rise to highly diverse and unique repertiores

For each T-cell, a unique receptor is quasi-randomly generated during a process called V(D)J recombination



Necessity of studying the TCR repertoire

The TCR repertoire captures a fingerprint of:

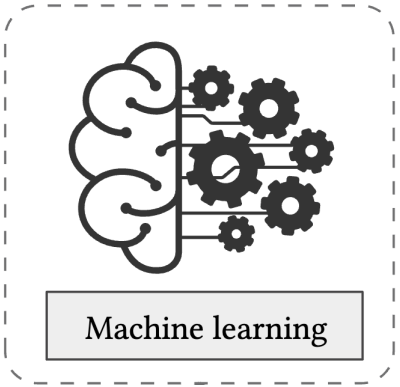
- Current immune responses
- Past immune exposures
- Future protection and infection outcomes

Current challenges of understanding the TCR repertoire

- Many TCR-epitope interaction are unknown
- Not experimentally feasible to test all interaction
- Prediction modelling of the TCR-epitope interactions
  - Necessary to identify which TCR-epitope combinations to experimentally validate

Overview

- Data driven prediction method for annotating TCR repertoire
  - Databases
  - Machine learning
  - Extending the prediction with scRNA-seq with TCR-seq

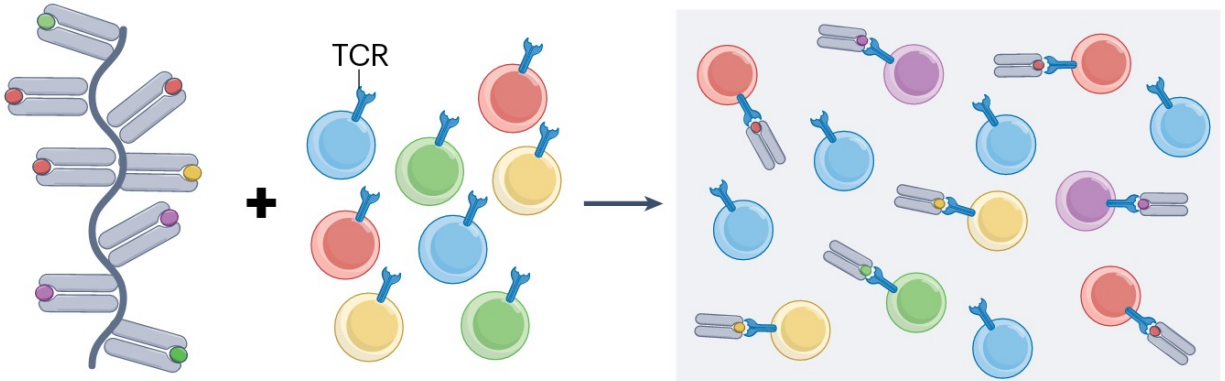


Overview

- Data driven prediction method for annotating TCR repertoire
  - Databases
  - Machine learning
  - Extending the prediction with

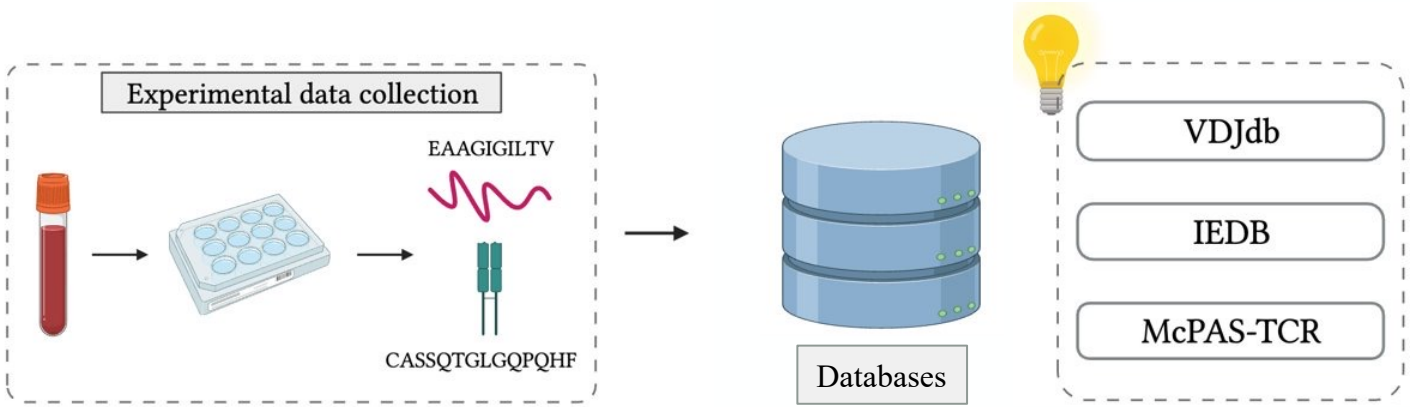
T-cell antigen discovery

Antigen-directed experimental methods for readout of TCR-pMHC interactions



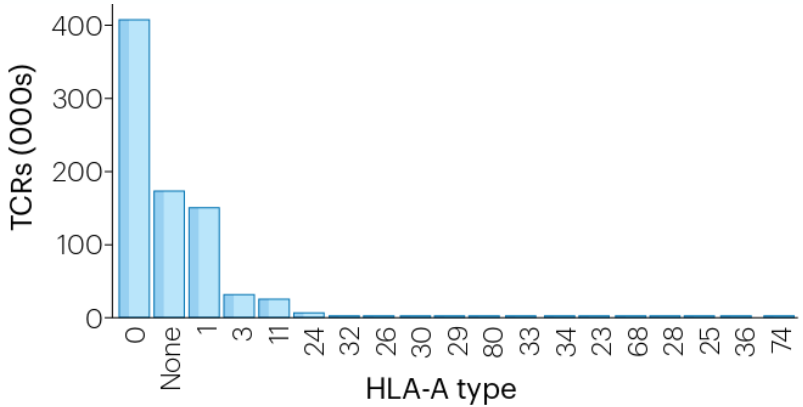
Joglekar and Li, 2021
Figure: Hudson et al., 2023

TCR-epitope pairs are compiled in curated databases



The current landscape of known TCR-epitope pairs

	TCRA	TCRB	Paired
VDJdb	29062	41427	23944
IEDB	24322	130979	27021
McPAS-TCR	8167	29049	36219
(MIRA)		135000+	



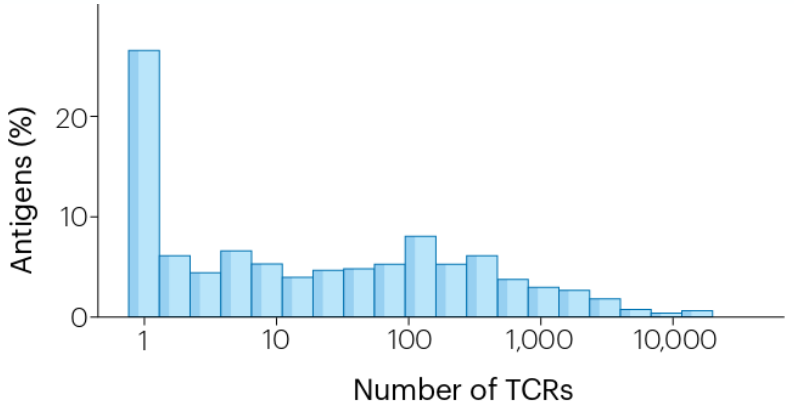
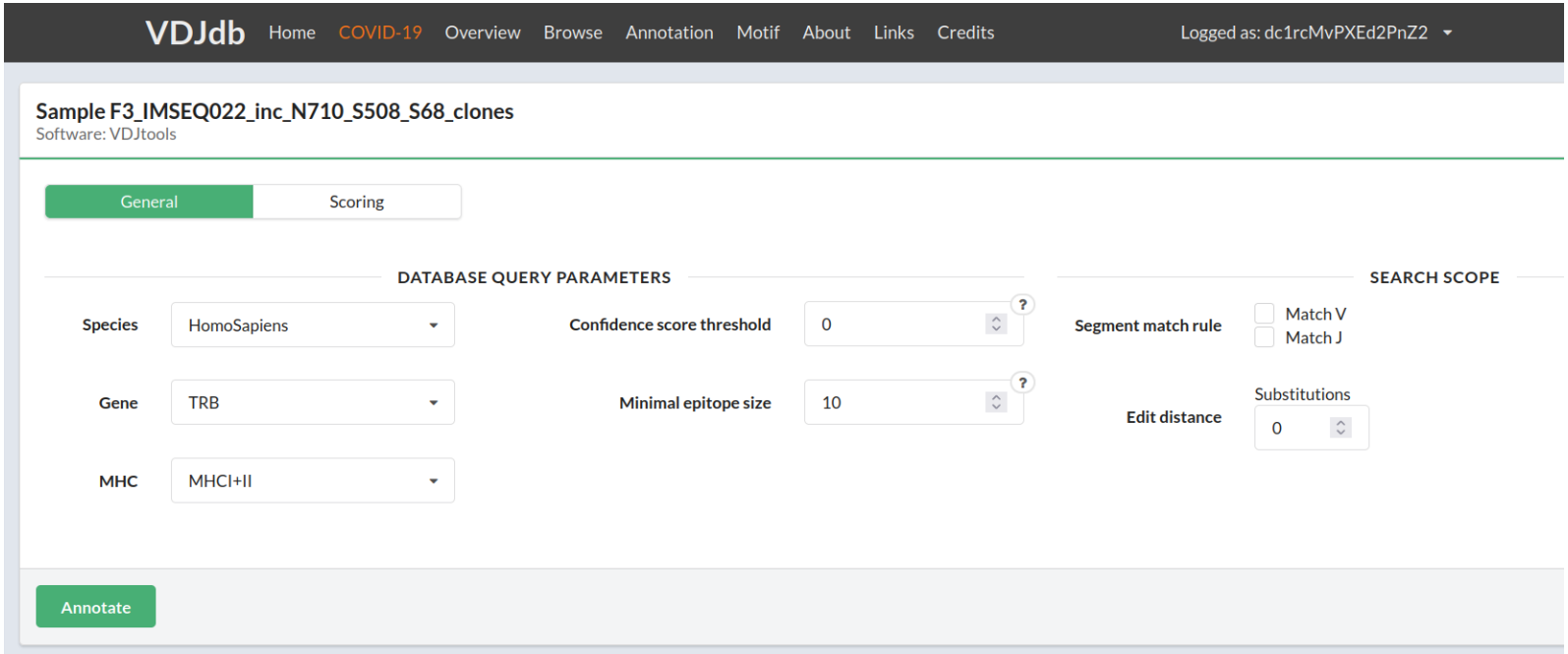
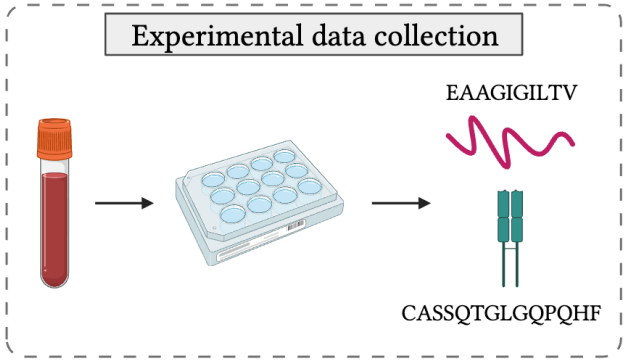


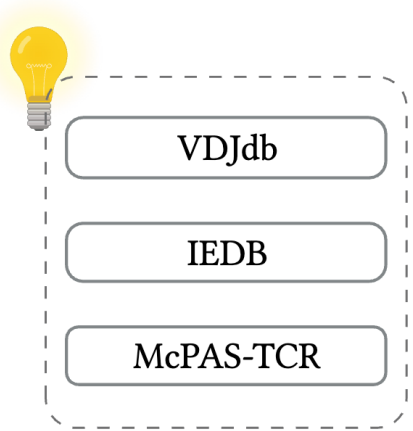
Figure: Hudson et al., 2023

Annotation approach 0: exact database matching



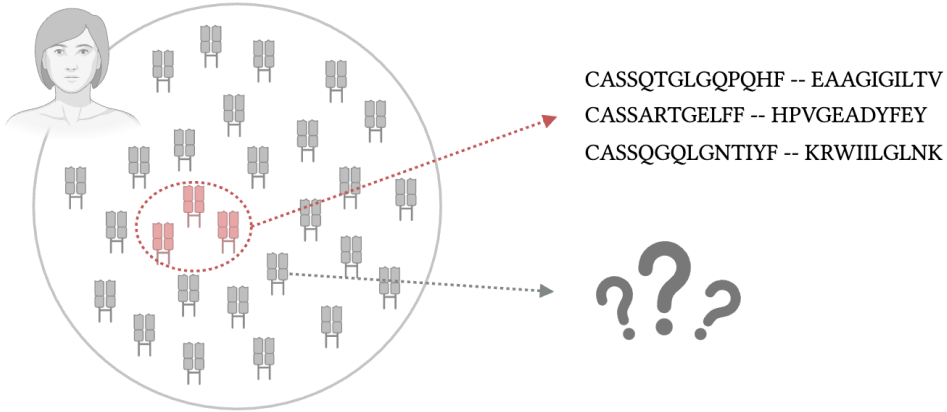
Exact annotations are sparse





BUT

Very limited number of TCR-epitope pairs identified



Distance-based TCR analysis

Similar TCR sequences often recognize the same epitope

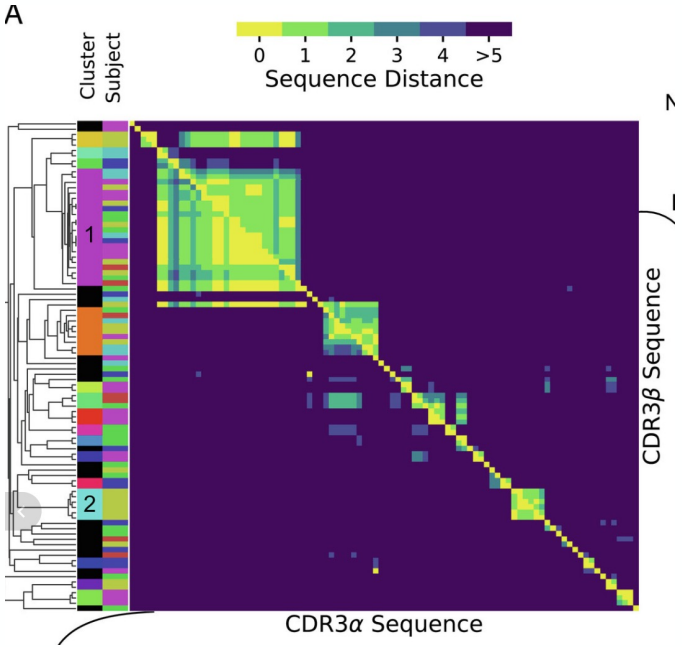


Figure: Mayer et al., 2023
Data: Dash et al., 2017

Distance-based TCR analysis

Similar TCR sequences often recognize the same epitope
=> approximate database matching

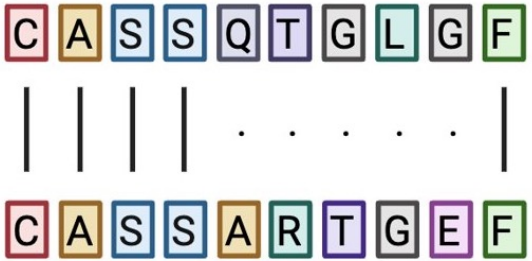
Two main approaches:

- Matching based on distance threshold (e.g. TCRMatch)
- Clustering-based (e.g. ClusTCR, GLIPH2)

One requirement: **distance metric**
"How do we define TCR similarity?"

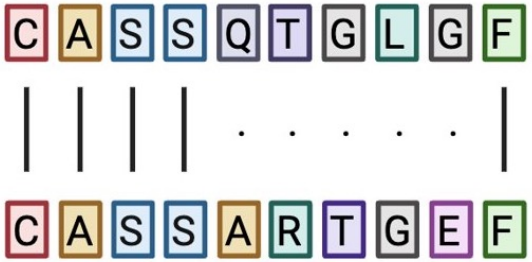
TCR similarity metrics

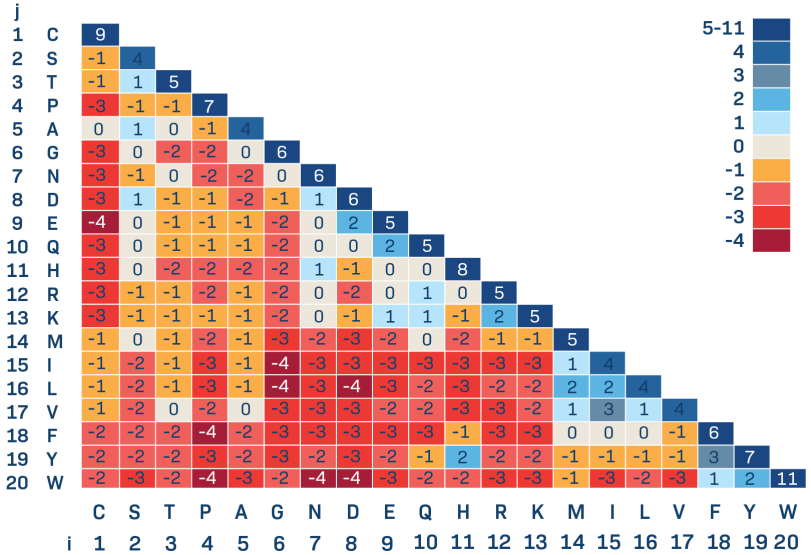
- **Hamming distance *(e.g. ClusTCR)***
- **Edit distance (LD) *(e.g. VDJdb search)***



TCR similarity metrics

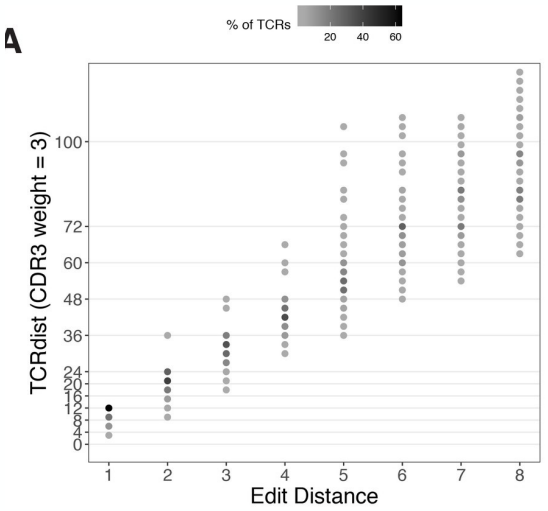
- **Hamming distance (*e.g. ClusTCR*)**
- **Edit distance (LD) (*e.g. VDJdb search*)**
- **Alignment+BLOSUM approaches (*e.g. TCRdist*)**

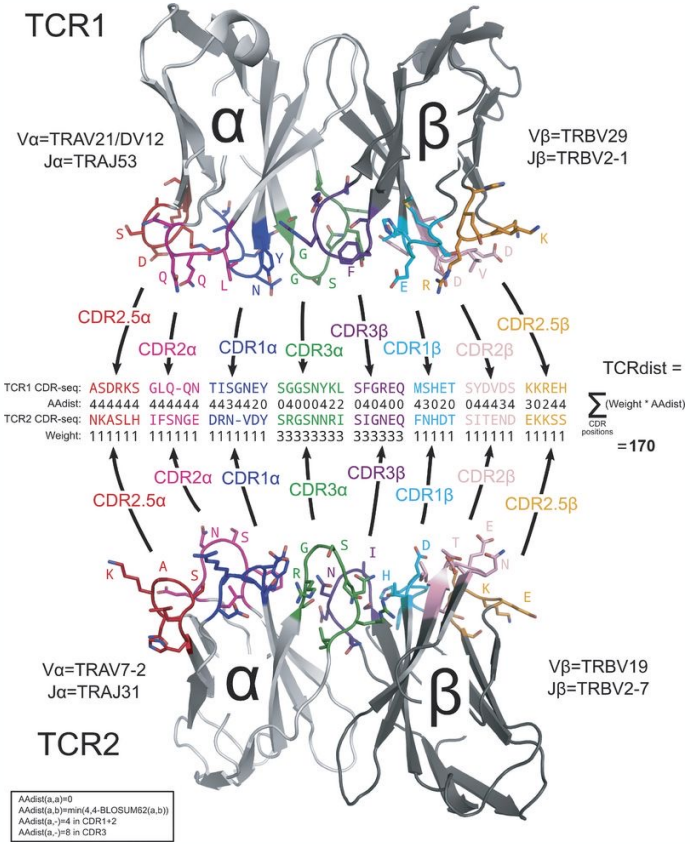




TCR similarity metrics

- **Hamming distance (*e.g. ClusTCR*)**
- **Edit distance (*e.g. VDJdb search*)**
- **Alignment+BLOSUM approaches (*e.g.TCRdist*)**

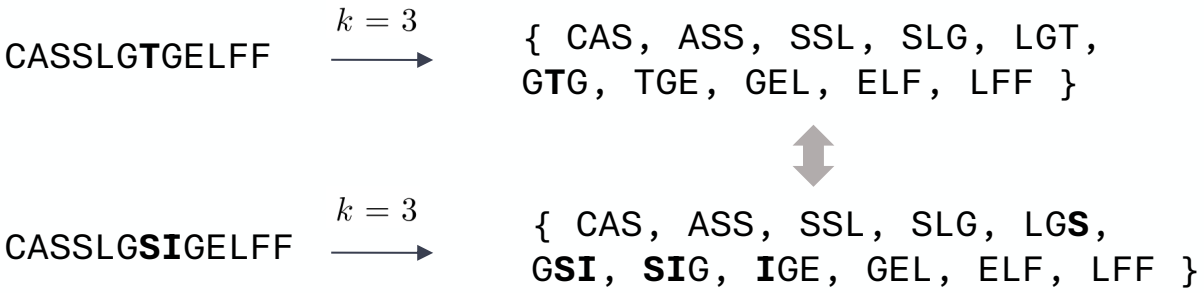




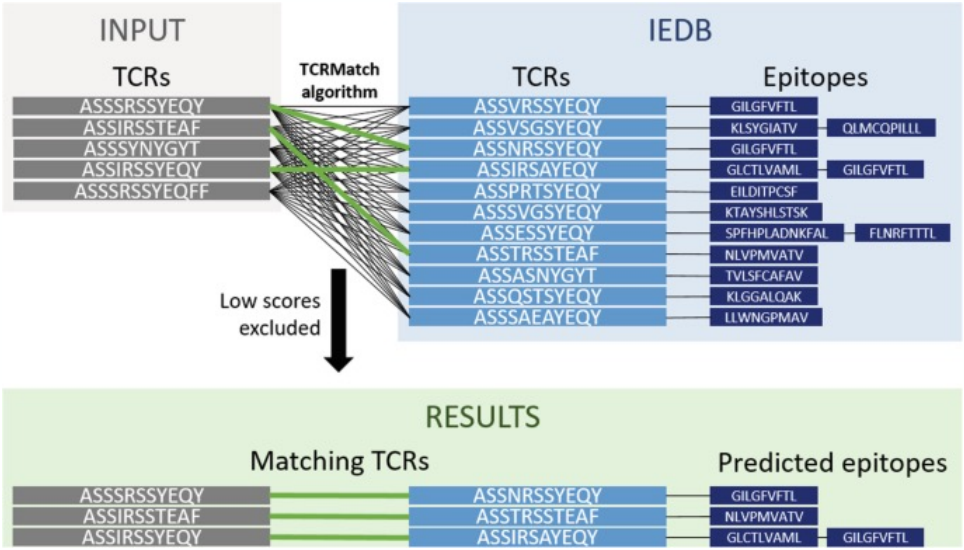
Mayer-Blackwell et al. 2021, Dash et al. 2017

TCR similarity metrics

- **Hamming distance (*e.g. ClusTCR*)**
- **Edit distance (LD)**
- **Alignment+BLOSUM approaches (*e.g. TCRdist*)**
- **K-mer approaches (*e.g. GLIPH2, TCRMatch*)**

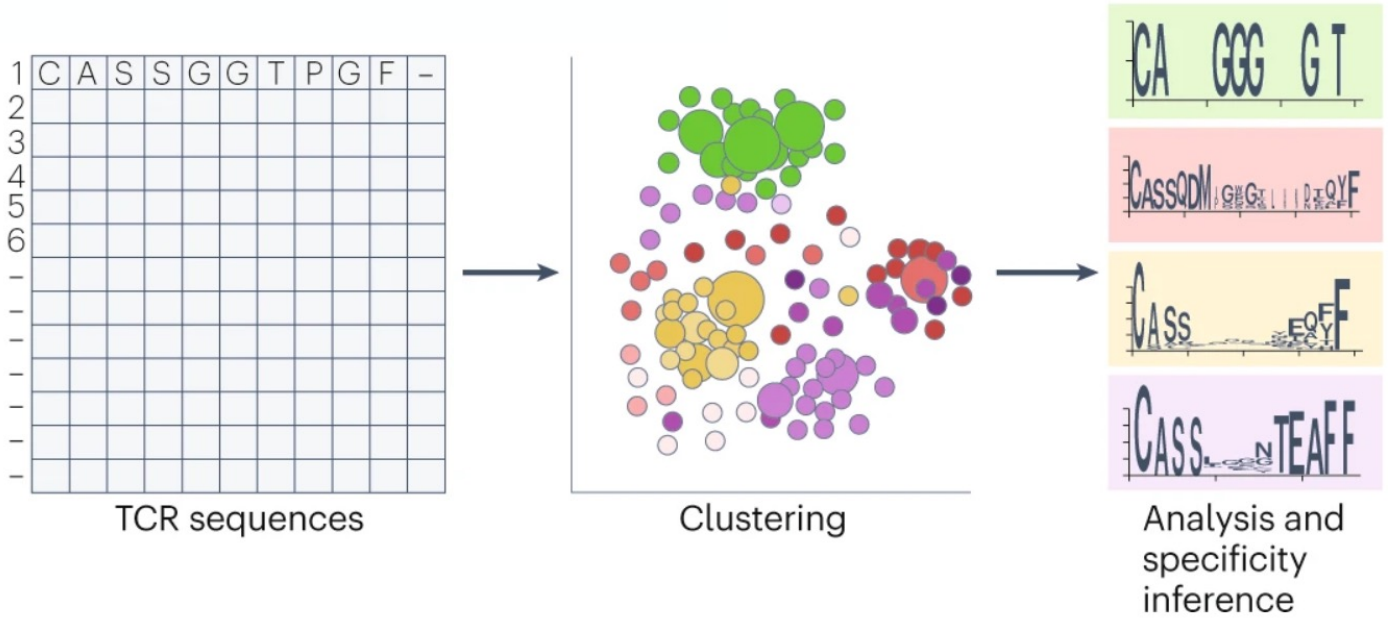


Approach 1: distance-based annotation



Chronister et al., 2021
http://tools.iedb.org/tcrmatch/

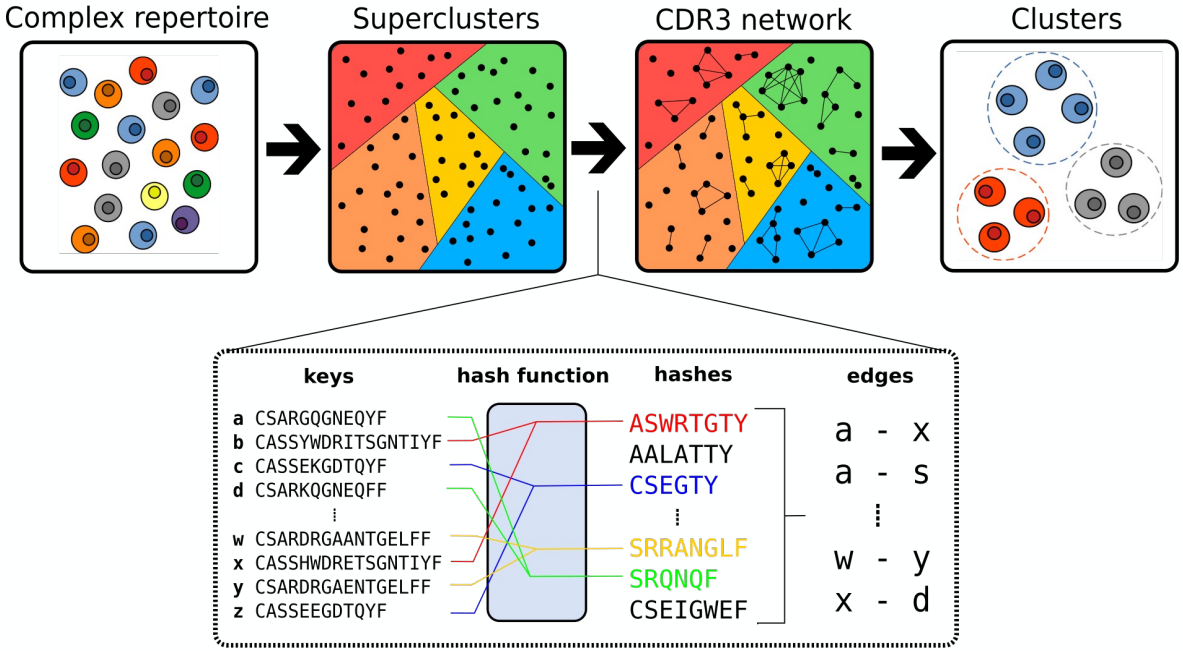
Approach 2: clustering-based annotation



Meysman et al., 2019, Valkiers et al., 2021
Figure: Hudson et al., 2023

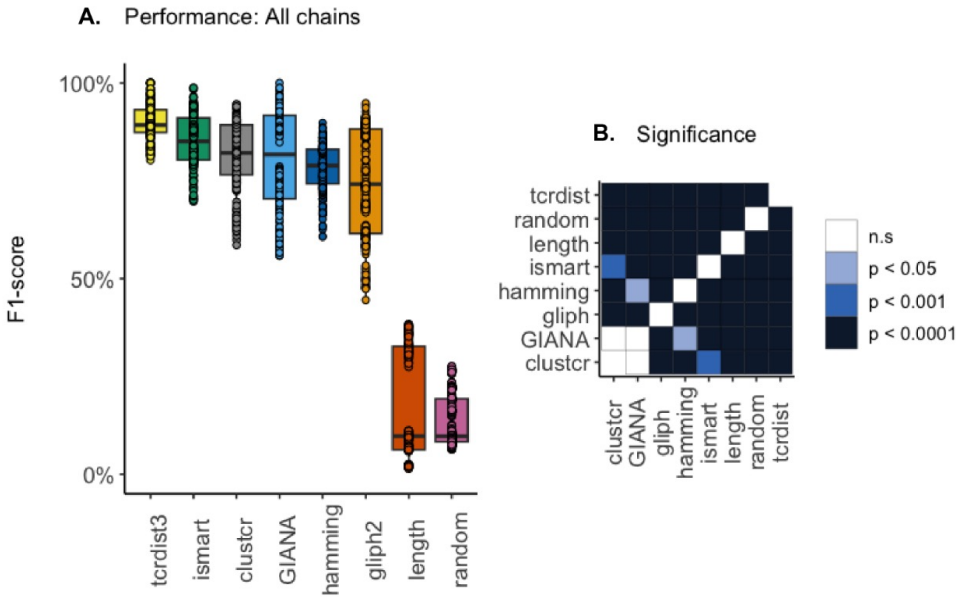
Approach 2: clustering-based annotation

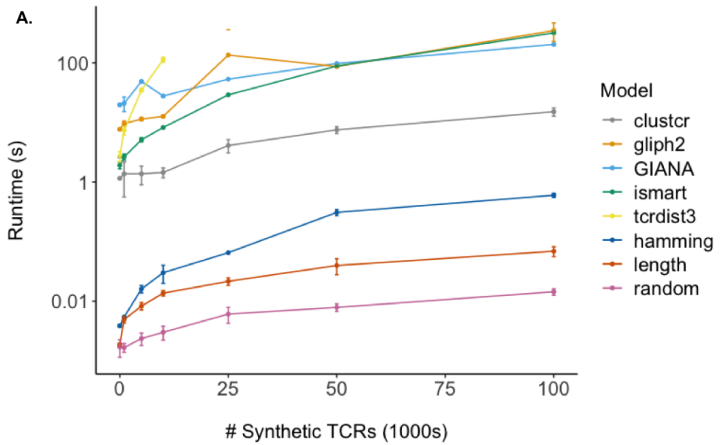
- ***Example: ClusTCR***



Valkiers et al., 2021

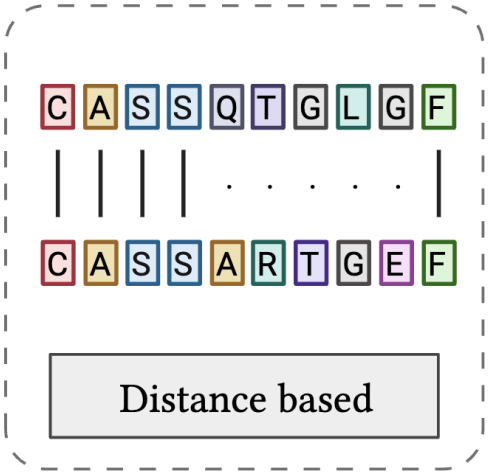
Which clustering approach to use?

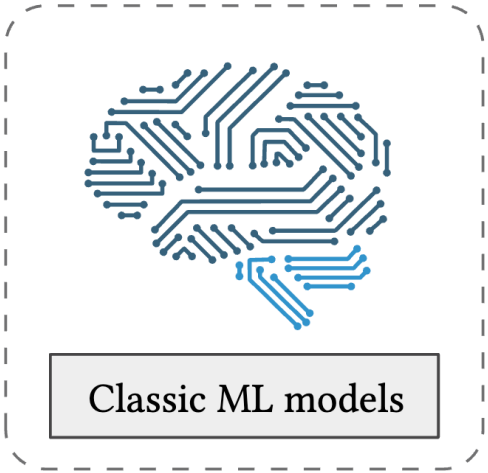




Hudson et al., 2023 (preprint)
https://doi.org/10.1101/2023.08.04.551940

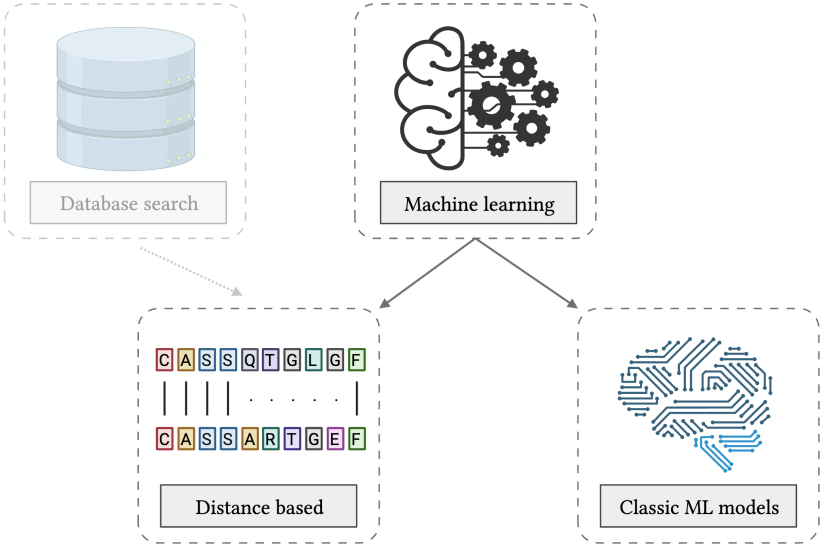
Why use machine learning models?





Overview

- Data driven prediction method for annotating TCR repertoire
  - Databases
  - Machine learning
  - Extending the prediction



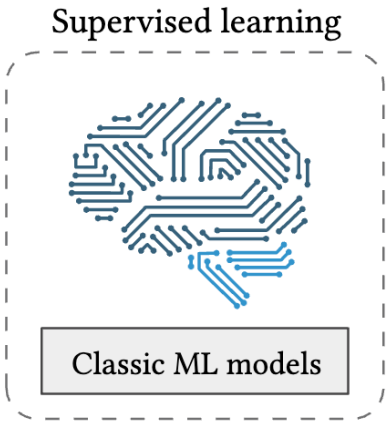
Feature-based machine learning models

- Mechanism:
  - Discover common patterns in the sequences
  - LEARN which features are important
  - Predict binding for unseen sequences

TCR sequences	Epitope specificity
CASSQTGLGQPQHF	✓
CASSARTGELFF	✓
CASSQGQLGNTIYF	✗
CASSLGLNTEAFF	✓
CSARDSYEQYF	✗
CASSGGSSYEQYF	✗

Input

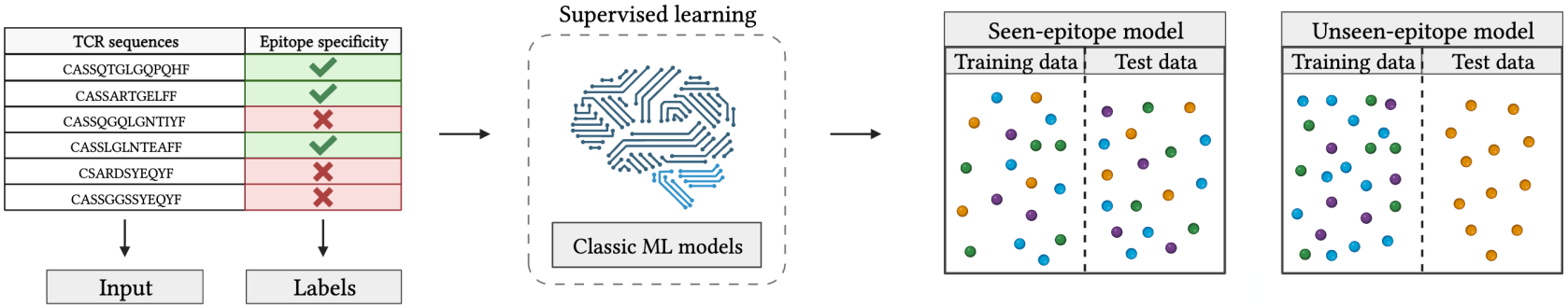
Labels



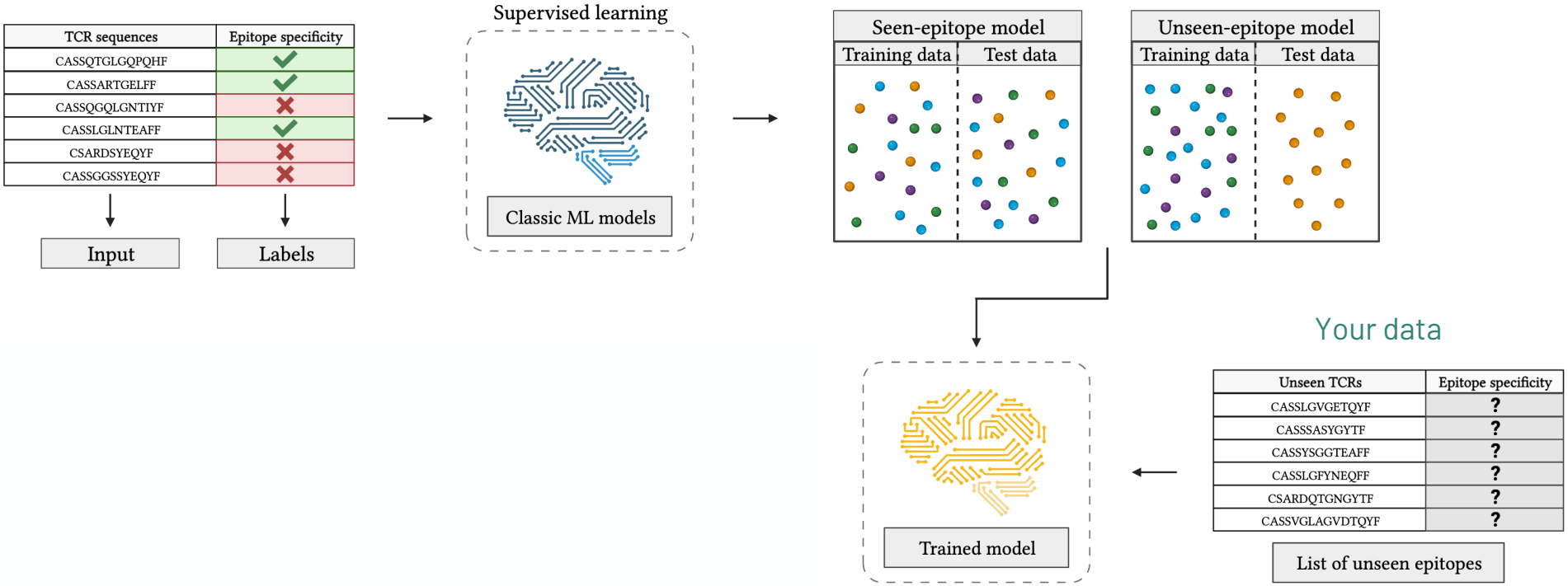
Examples:

- DiffRBM
- ImRex
- NetTCR
- SONIA
- TCR-BERT
- TCRex
- TITAN

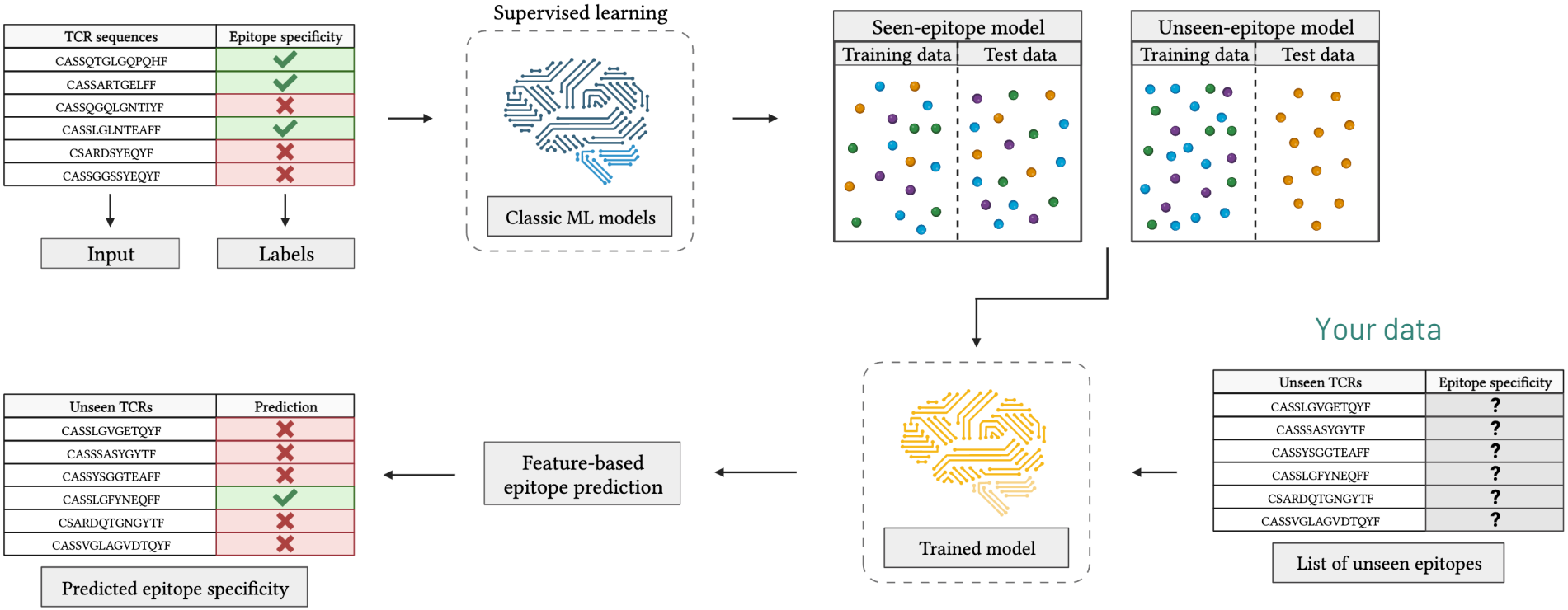
Feature-based machine learning models



Feature-based machine learning models



Feature-based machine learning models



Benchmarking public TCR-epitope prediction

- IMMREP22 workshop:
  - Public TCR-epitope prediction benchmark
  - Evaluate current tools & evaluation strategies

- In total 23 different models:
  - Most are feature-based models
  - Some are distance-based models

ImmunoInformatics 9 (2023) 100024

Contents lists available at ScienceDirect

ImmunoInformatics

journal homepage: www.journals.elsevier.com/immunoinformatics

Benchmarking solutions to the T-cell receptor epitope prediction problem: IMMREP22 workshop report

Pieter Meysman[a,b,*], Justin Barton[c,1], Barbara Bravi[d,1], Liel Cohen-Lavi[e,f,1], Vadim Karnaukhov[h,i,1], Elias Lilleskov[j,1], Alessandro Montemurro[k,1], Morten Nielsen[k,1], Thierry Mora[i,1], Paul Pereira[i,l,1], Anna Postovskaya[a,b,m,1], María Rodríguez Martínez[n,1], Jorge Fernandez-de-Cossio-Diaz[i,1], Alexandra Vujkovic[a,b,m,1], Aleksandra M. Walczak[i,1], Anna Weber[n,1], Rose Yin[o,1], Anne Eugster[g,2,**], Virag Sharma[p,2,**]

Benchmarking public TCR-epitope prediction

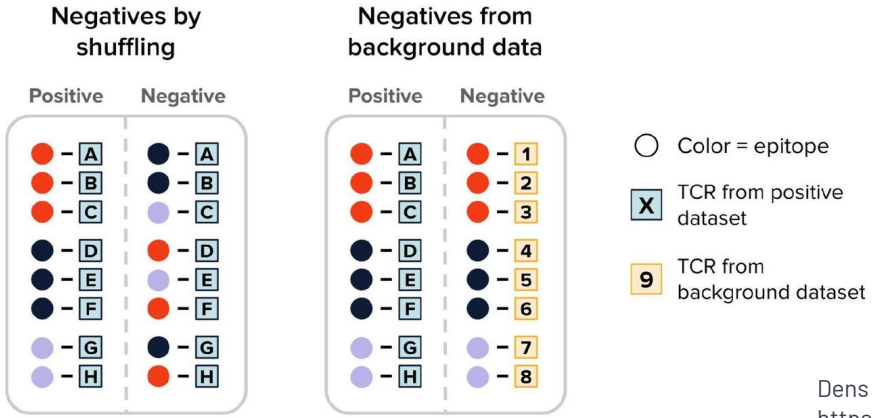
- IMMREP22 workshop:
  - Public TCR-epitope prediction benchmark
  - Evaluate current tools & evaluation strategies
- In total 23 different models:
  - Most are feature-based models
  - Some are distance-based models
- Only paired data was collected
- Provided **positive** and **negative** data
  - Positive data = known TCR-epitope pairs
  - Negative data = random sampling & negative control set
  - Training – Test : 80/20 ratio
- Compare all tools

Epitope	# Training samples
LTDEMIAQY	200
GILGFVFTL	1088
TTDPSFLGRY	386
NQKLIANQF	112
HPVTKYIM	96
GPRLGVRAT	80
KSKRTPMGF	170
CINGVCWTV	366
TPRVTGGGAM	90
SPRWYFYYL	184
LLWNGPMAV	376
GLCTLVAML	292
YLQPRTFLL	534
ATDALMTGF	208
NLVPMVATV	548
RAQAPPPSW	72
NYNYLYRLF	88

The pitfalls of negative data bias

Negative data

- There is no ‘ground-truth’ due to lack in high-quality negative data
- Artificially generate the negative data:
  - Shuffling of positive data
  - Use background TCR data set
- Model might learn to differentiate based on bias instead of TCR features



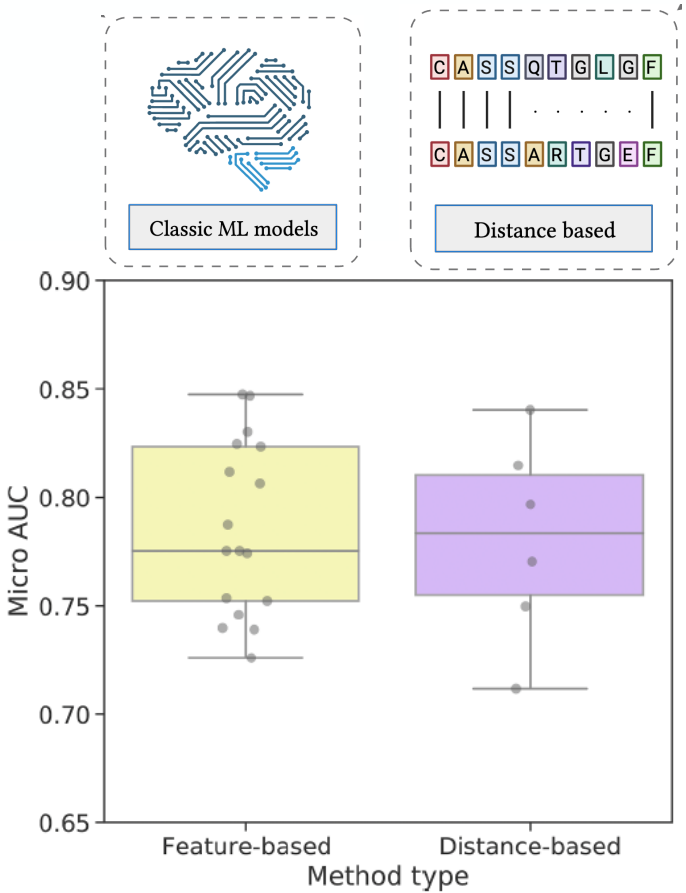
Dens et al., 2023
https://doi.org/10.1038/s42256-023-00727-0

Insights into TCR-epitope prediction

- General conclusions:

➢ Predictions work better for TCRs similar to the training data

➢ The ‘simpler’ distance-based models also perform very well



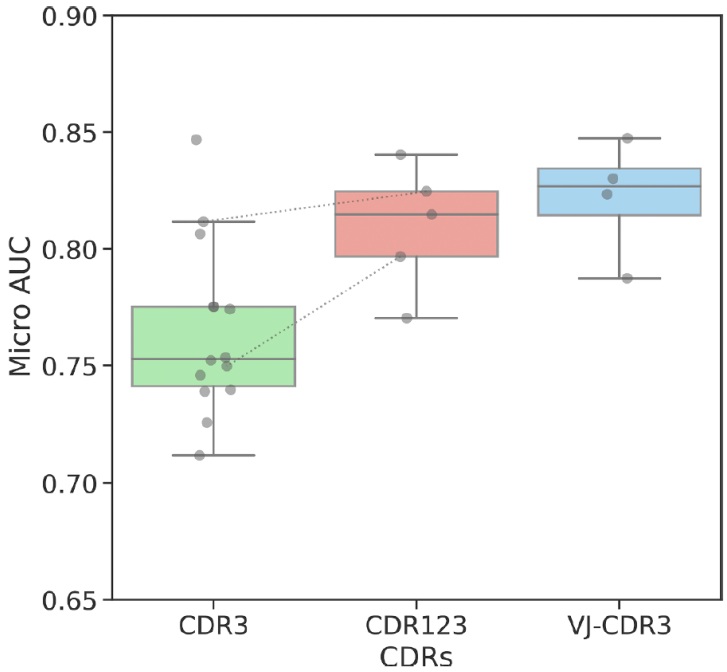
Insights into TCR-epitope prediction

- General conclusions:

➢ Predictions work better for TCRs similar to the training data

➢ The ‘simpler’ distance-based models also perform very well

➢ Prediction is better when also including the V- and J-genes



Insights into TCR-epitope prediction

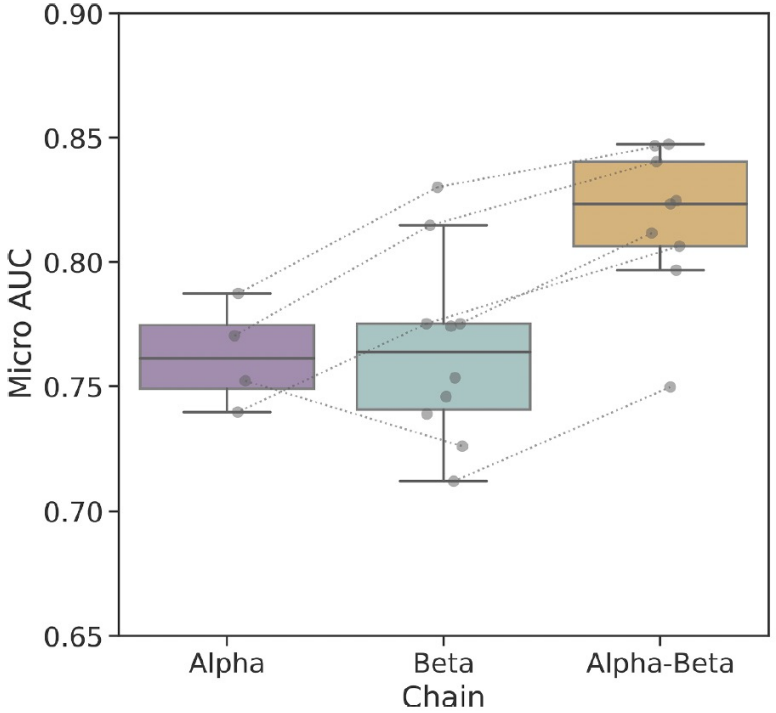
- General conclusions:

➢ Predictions work better for TCRs similar to the training data

➢ The ‘simpler’ distance-based models also perform very well

➢ Prediction is better when also including the V- and J-genes

➢ Prediction is better when using both alpha and beta chains



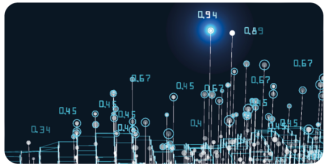
Stay tuned!

JUSTIN BARTON · COMMUNITY PREDICTION COMPETITION · 5 DAYS TO GO

Join Competition

IMMREP23: TCR Specificity Prediction Challenge

Competitors will make predictions on previously unpublished TCR-epitope binding data in order to benchmark prediction methods.



Overview Data Code Models Discussion Leaderboard Rules

Overview

Start
a month ago

Close
5 days to go

Description

IMMREP23, the second annual IMMREP benchmark on TCR-epitope specificity prediction will run from November 1, 2023 to December 11, 2023. Together with several experimental groups, we have compiled a data set of paired TCR data with annotated specificity to 21 pHLA (covering 6 distinct HLA molecules).

This challenge models TCR epitope recognition as a binary classification task. For a given test set of TCR-epitope pairs, the task of the model is to identify which pairs will bind and which will not bind.

Competition Host
Justin Barton

Prizes & Awards
Kudos
Does not award Points or Medals

Participation
42 Competitors
42 Teams
262 Entries

Tags
Biology Biotechnology Binary Classification

What is TCRex?

TCR-epitope prediction tool

Train your own model

TCRex

Online user-friendly interface

Prediction for 100 epitopes

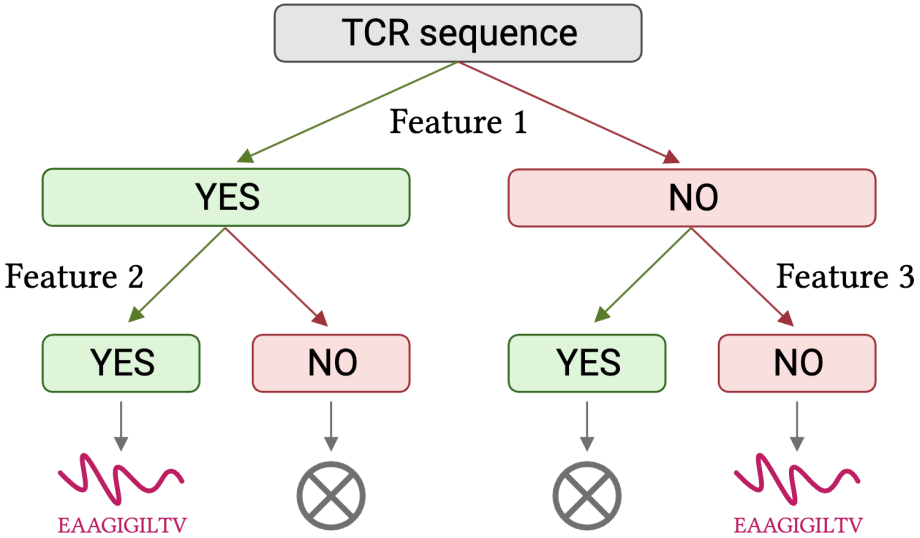
Random forest models

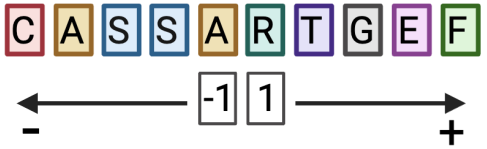
Manually curated data

https://tcrex.biodatamining.be

The secret behind TCRex

TCRex

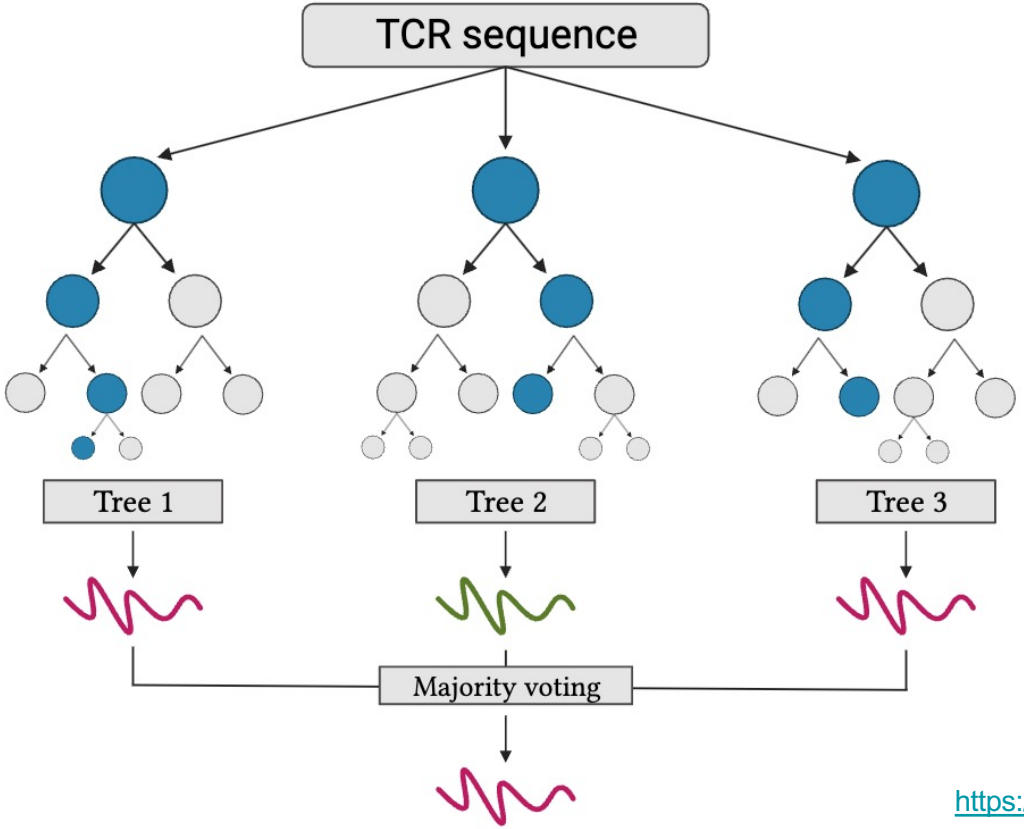




- V- & J-gene
- Sequence length
- Amino acid count
- Average physico-chemical properties
- Sequence pI
- Peptide mass
- Position-specific amino acid presence
- Position-specific physico-chemical properties

The secret behind TCRex

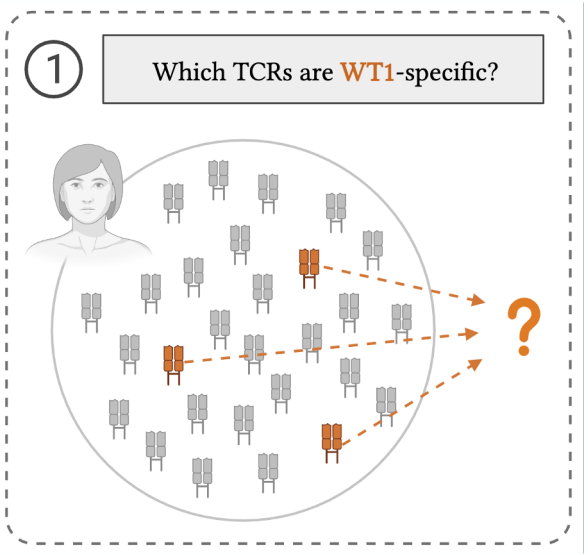
TCRex



A variety of epitopes in TCRex

TCRex

- Random forest models for 100 epitopes
  - 93 viral epitope
  - 7 cancer epitopes (including WT1 epitopes)



Viral

CMV
IPSINVHHY NLVPMVATV QIKVRVKMV
QYDPVAALF TPRVTGGGAM VTEHDTLLY
YSEHPTFTSQY

DENV1
GTSGSPIVNR

DENV2
GTSGSPIIDK

DENV3/4
GTSGSPIINR

EBV
EPLPQGQLTAY GLCTLVAML HPVGEADYFEY
IVTDFSVIK RAKFKQLL YVLDHLIVV

HCV
ARMILMTHF ATDALMTGY CINGVCWTV
HSKKKCDEL KLVALGINAV

Cancer

Melanoma
AMFWSVPTV EAAGIGILTV ELAGIGILTV
FLYNLLTRV

Multiple Myeloma
LLLGIGILV

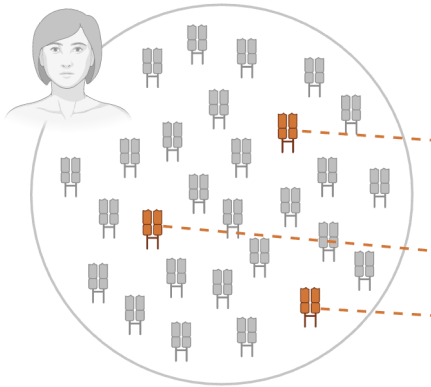
Tumor associated antigen (WT1)
RMFPNAPYL VLDFAPPGA

https://tcrex.biodatamining.be

Identifying WT1-specific epitopes in melanoma

TCRex

① Which TCRs are WT1-specific?



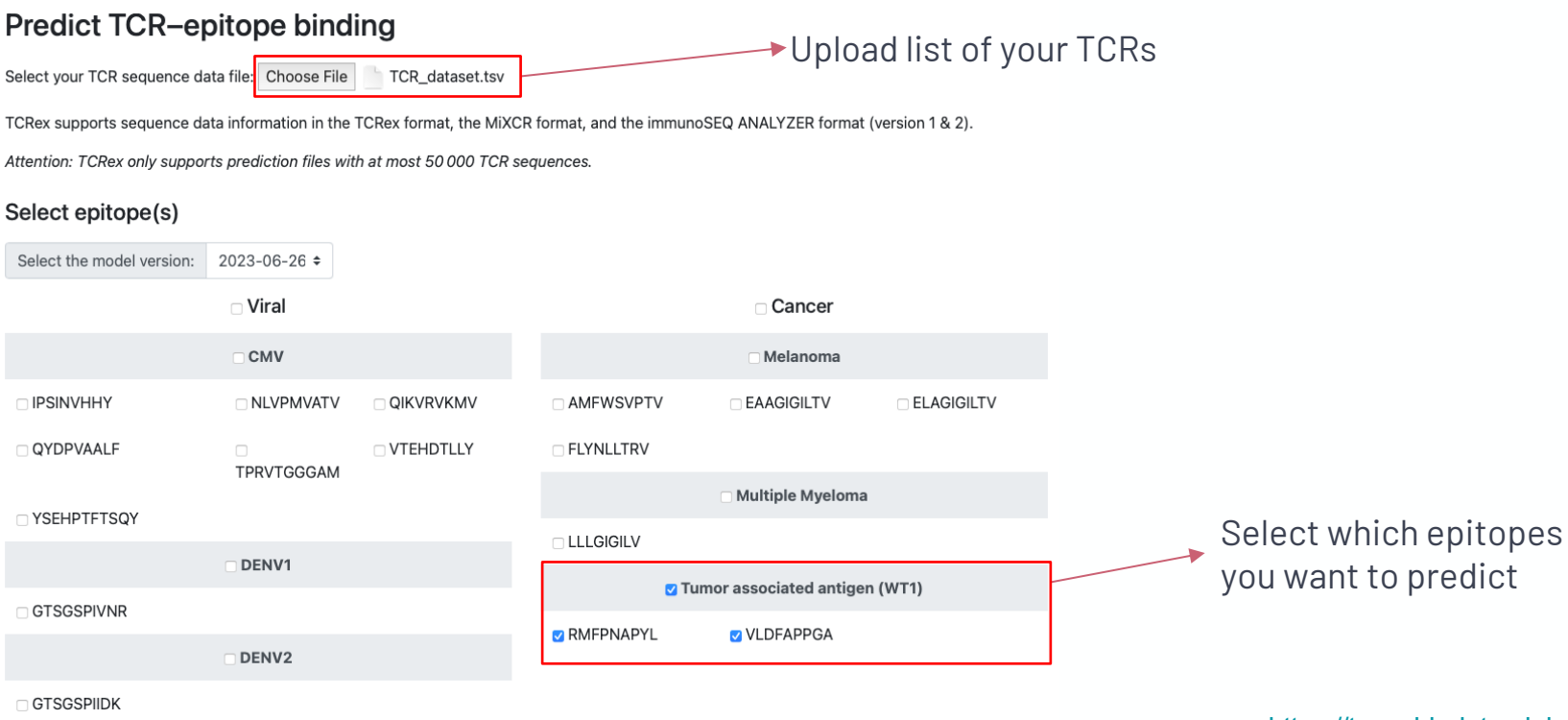
② List all TCRs in your sample(s)

TRBV_gene	CDR3_beta	TRBJ_gene
TRBV07-09	CASSGTQYGYTF	TRBJ01-02
TRBV14	CASSQTGLGQPQHF	TRBJ01-05
TRBV09	CASSARTGELFF	TRBJ02-02
TRBV04-03	CASSQGQLGNTIYF	TRBJ01-03
TRBV05-04	CASSLGTDLAKNIQYF	TRBJ02-04
TRBV12-03	CASSFGGGELFF	TRBJ02-02
TRBV07-09	CASSLIGVSSYNEQFF	TRBJ02-01
TRBV09	CASSVGQGAYEQYF	TRBJ02-07
TRBV09	CASSVRDWPYEQYF	TRBJ02-07
TRBV09	CASSAGTVNTGELFF	TRBJ02-02
TRBV07-08	CASSLGQAYEQYF	TRBJ02-07
TRBV14	CASSLGLNTEAFF	TRBJ01-01

③ Run this list through TCRex

Identifying WT1-specific epitopes in melanoma

TCRex



Identifying WT1-specific epitopes in melanoma

TCRex

Enrichment results

Be cautious when interpreting these enrichment results: they are valid for the used background dataset which might not provide the best background for your dataset.

Show 10 entries Search:

Epitope	Pathology	P value	FDR-corrected P value
RMFPNAPYL	Tumor associated antigen (WT1)	4.32e-03	8.65e-03
VLDFAPPGA	Tumor associated antigen (WT1)	1.37e-01	1.37e-01

Showing 1 to 2 of 2 entries Previous 1 Next

Prediction results

BPR threshold 0.01 %

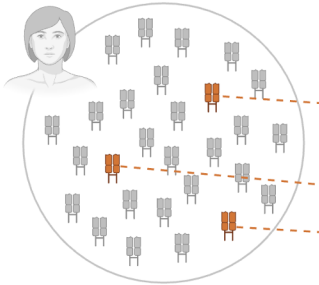
Show 10 entries Search:

TRBV gene	CDR3 sequence	TRBJ gene	Epitope	Pathology	Score	BPR (%)
TRBV05-04	CASSLGTDLAKNIQYF	TRBJ02-04	RMFPNAPYL	Tumor associated antigen (WT1)	0.62	0.0000
TRBV04-01	CASSLLAGEQETQYF	TRBJ02-05	RMFPNAPYL	Tumor associated antigen (WT1)	0.57	0.0010
TRBV19	CASSNLAGVRDTQYF	TRBJ02-03	RMFPNAPYL	Tumor associated antigen (WT1)	0.57	0.0010
TRBV07-02	CASSWGGQGSDTQYF	TRBJ02-03	RMFPNAPYL	Tumor associated antigen (WT1)	0.54	0.0030
TRBV09	CASSVLAGDQETQYF	TRBJ02-05	RMFPNAPYL	Tumor associated antigen (WT1)	0.54	0.0030
TRBV30	CAWSRLAGGSDTQYF	TRBJ02-03	RMFPNAPYL	Tumor associated antigen (WT1)	0.54	0.0030
TRBV05-04	CASSTLAGPQETQYF	TRBJ02-05	RMFPNAPYL	Tumor associated antigen (WT1)	0.54	0.0030

Identifying WT1-specific epitopes in melanoma

TCRex

1 Which TCRs are WT1-specific?



2 List all TCRs in your sample(s)

TRBV_gene	CDR3_beta	TRBJ_gene
TRBV07-09	CASSGTQYGYTF	TRBJ01-02
TRBV14	CASSQTGLGQPQHF	TRBJ01-05
TRBV09	CASSARTGELFF	TRBJ02-02
TRBV04-03	CASSQGQLGNTIYF	TRBJ01-03
TRBV05-04	CASSLGTDLAKNIQYF	TRBJ02-04
TRBV12-03	CASSFGGGELFF	TRBJ02-02
TRBV07-09	CASSLIGVSSYNEQFF	TRBJ02-01
TRBV09	CASSVGQGAYEQYF	TRBJ02-07
TRBV09	CASSVRDWPYEQYF	TRBJ02-07
TRBV09	CASSAGTVNTGELFF	TRBJ02-02
TRBV07-08	CASSLGQAYEQYF	TRBJ02-07
TRBV14	CASSLGLNTEAFF	TRBJ01-01

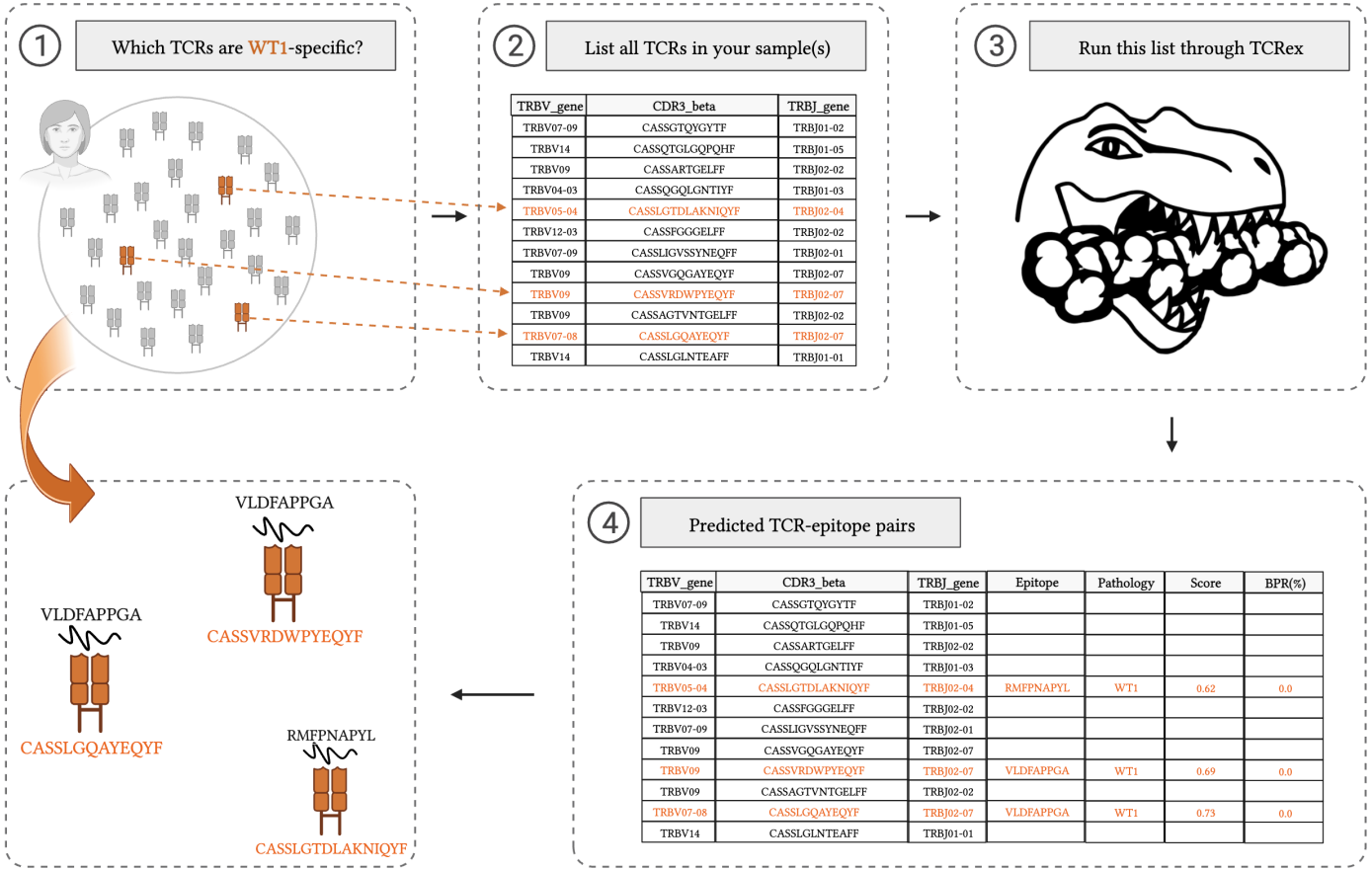
3 Run this list through TCRex

4 Predicted TCR-epitope pairs

TRBV_gene	CDR3_beta	TRBJ_gene	Epitope	Pathology	Score	BPR(%)
TRBV07-09	CASSGTQYGYTF	TRBJ01-02				
TRBV14	CASSQTGLGQPQHF	TRBJ01-05				
TRBV09	CASSARTGELFF	TRBJ02-02				
TRBV04-03	CASSQGQLGNTIYF	TRBJ01-03				
TRBV05-04	CASSLGTDLAKNIQYF	TRBJ02-04	RMFPNAPYL	WT1	0.62	0.0
TRBV12-03	CASSFGGGELFF	TRBJ02-02				
TRBV07-09	CASSLIGVSSYNEQFF	TRBJ02-01				
TRBV09	CASSVGQGAYEQYF	TRBJ02-07				
TRBV09	CASSVRDWPYEQYF	TRBJ02-07	VLDFAPPGA	WT1	0.69	0.0
TRBV09	CASSAGTVNTGELFF	TRBJ02-02				
TRBV07-08	CASSLGQAYEQYF	TRBJ02-07	VLDFAPPGA	WT1	0.73	0.0
TRBV14	CASSLGLNTEAFF	TRBJ01-01				

Gain additional insights using ML models

TCRex

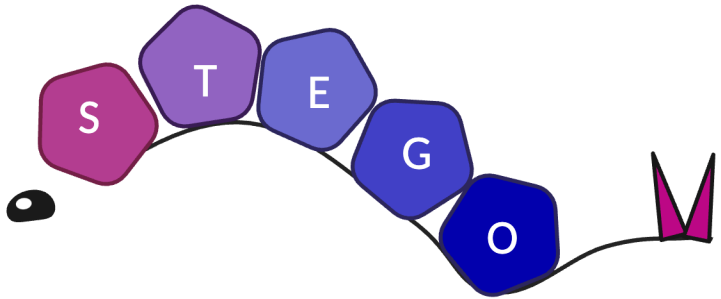


TRBV_gene	CDR3_beta	TRBJ_gene
TRBV07-09	CASSGTQYGYTF	TRBJ01-02
TRBV14	CASSQTGLGQPQHF	TRBJ01-05
TRBV09	CASSARTGELFF	TRBJ02-02
TRBV04-03	CASSQGQLGNTIYF	TRBJ01-03
TRBV05-04	CASSLGTDLAKNIQYF	TRBJ02-04
TRBV12-03	CASSFGGGELFF	TRBJ02-02
TRBV07-09	CASSLIGVSSYNEQFF	TRBJ02-01
TRBV09	CASSVGQGAYEQYF	TRBJ02-07
TRBV09	CASSVRDWPYEQYF	TRBJ02-07
TRBV09	CASSAGTVNTGELFF	TRBJ02-02
TRBV07-08	CASSLGQAYEQYF	TRBJ02-07
TRBV14	CASSLGLNTEAFF	TRBJ01-01

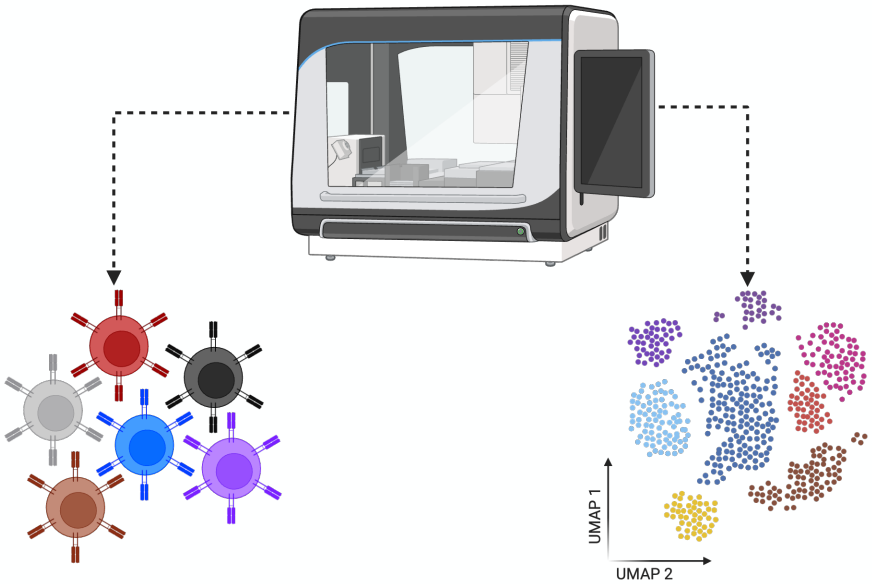
TRBV_gene	CDR3_beta	TRBJ_gene	Epitope	Pathology	Score	BPR(%)
TRBV07-09	CASSGTQYGYTF	TRBJ01-02				
TRBV14	CASSQTGLGQPQHF	TRBJ01-05				
TRBV09	CASSARTGELFF	TRBJ02-02				
TRBV04-03	CASSQGQLGNTIYF	TRBJ01-03				
TRBV05-04	CASSLGTDLAKNIQYF	TRBJ02-04	RMFPNAPYL	WT1	0.62	0.0
TRBV12-03	CASSFGGGELFF	TRBJ02-02				
TRBV07-09	CASSLIGVSSYNEQFF	TRBJ02-01				
TRBV09	CASSVGQGAYEQYF	TRBJ02-07				
TRBV09	CASSVRDWPYEQYF	TRBJ02-07	VLDFAPPGA	WT1	0.69	0.0
TRBV09	CASSAGTVNTGELFF	TRBJ02-02				
TRBV07-08	CASSLGQAYEQYF	TRBJ02-07	VLDFAPPGA	WT1	0.73	0.0
TRBV14	CASSLGLNTEAFF	TRBJ01-01				

Overview

- Data driven prediction method for annotating TCR repertoire
  - Databases
  - Machine learning
  - Extending the prediction to scRNA-seq with TCR-seq



Functional annotating the predictions with single cell RNA-seq

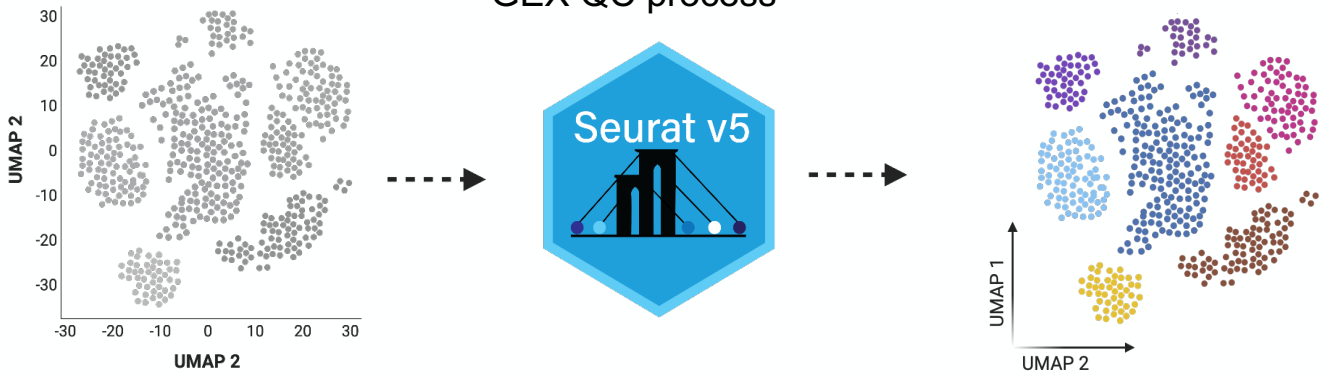


TCR repertoire

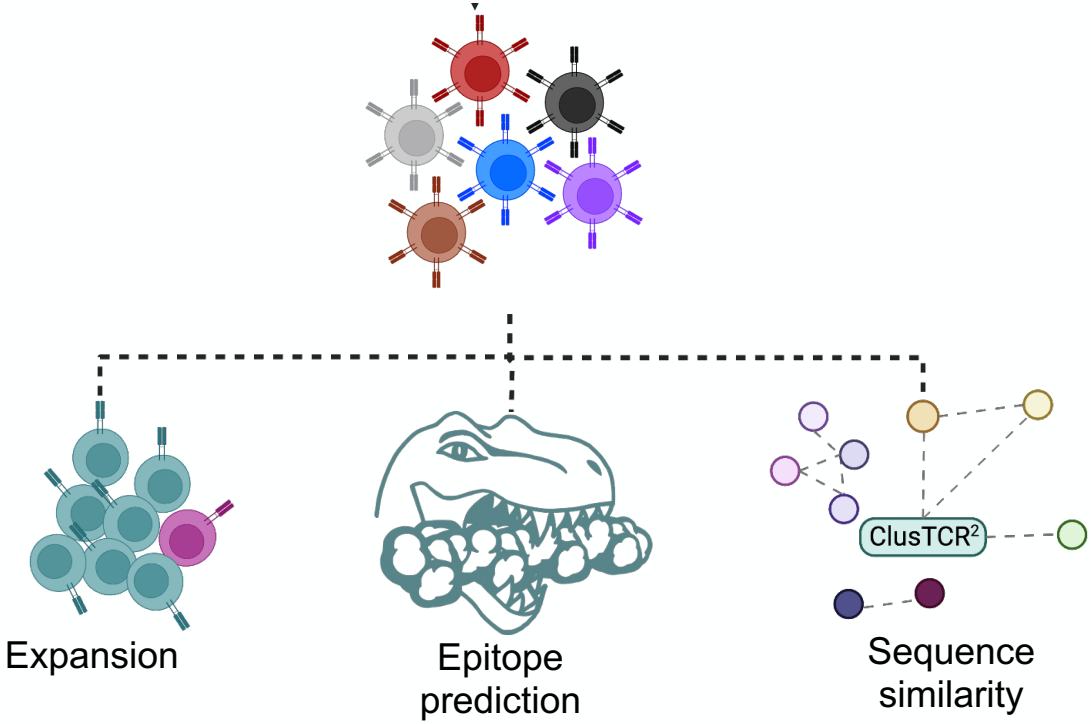
GEx

- scRNA-seq with TCR-seq
  - Paired clonotype
  - Expression layer as well

Pre-processing in STEGO.R (Shiny R package)



Pre-processing in STEGO.R (Shiny R package)

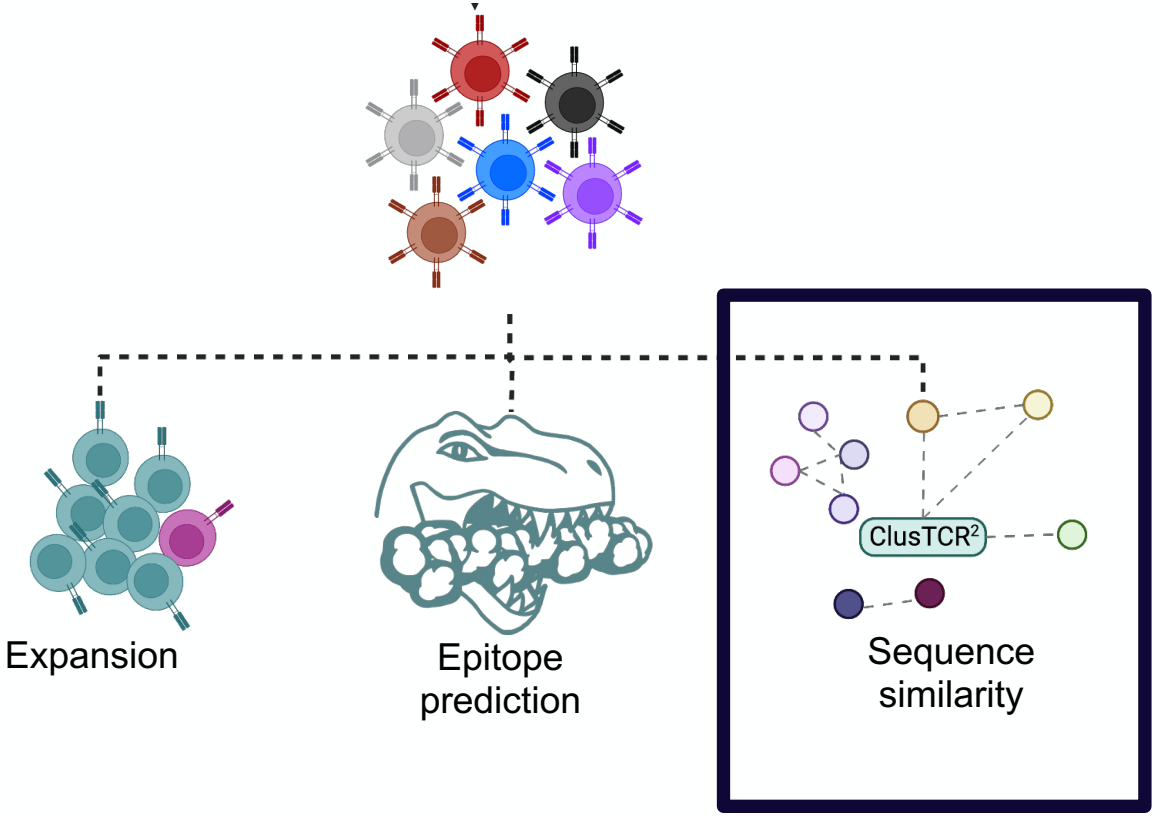


Example dataset

- **Dataset: colitis complication to melanoma treatment (GSE144469)**
  - Colitis (melanoma)
  - No colitis (melanoma)
  - Healthy controls

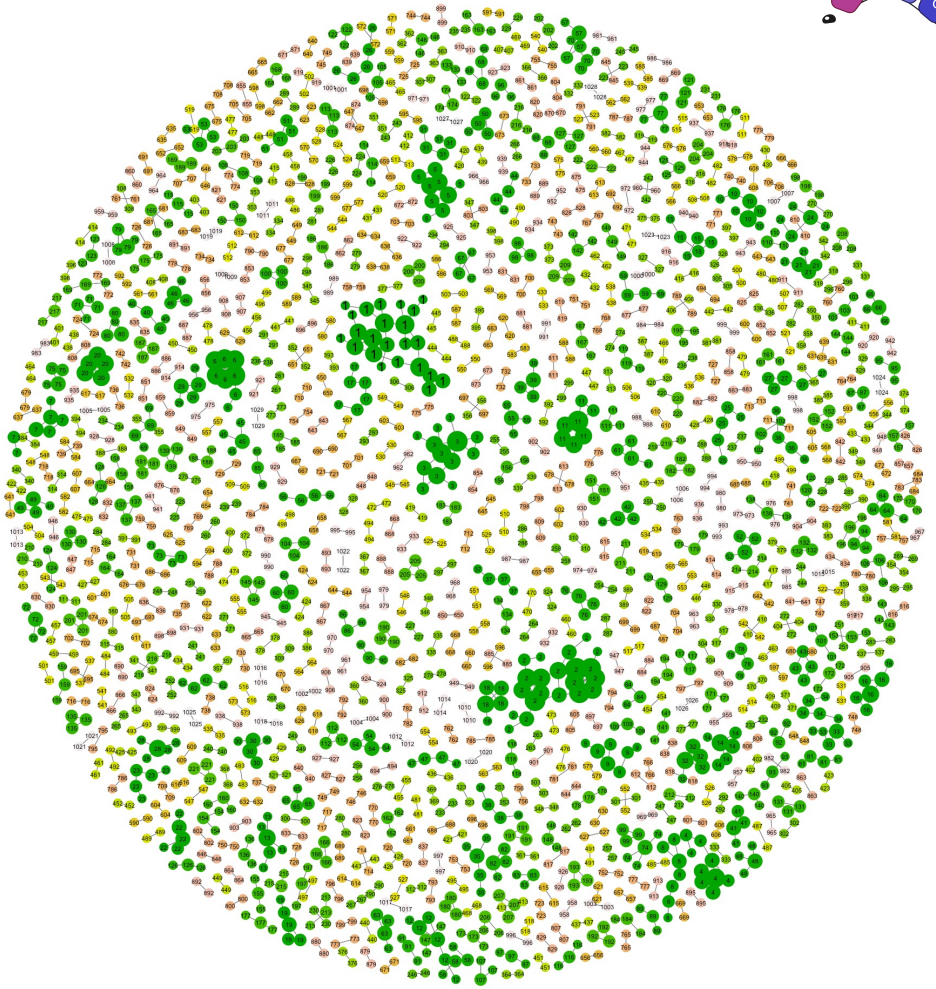
- **Extended analysis goals**
  - Clustering
  - WT1 predictions

Pre-processing in STEGO.R (Shiny R package)



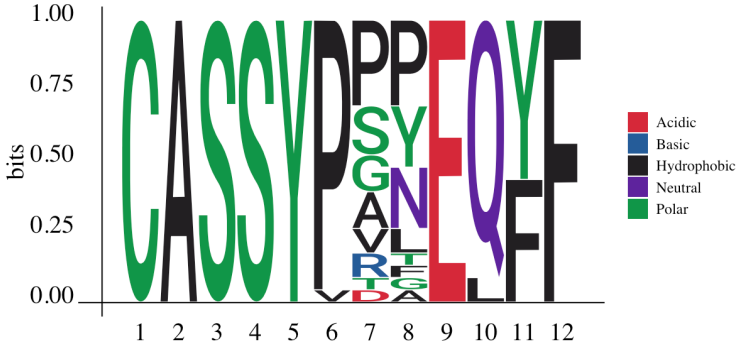
TCR-beta cluster

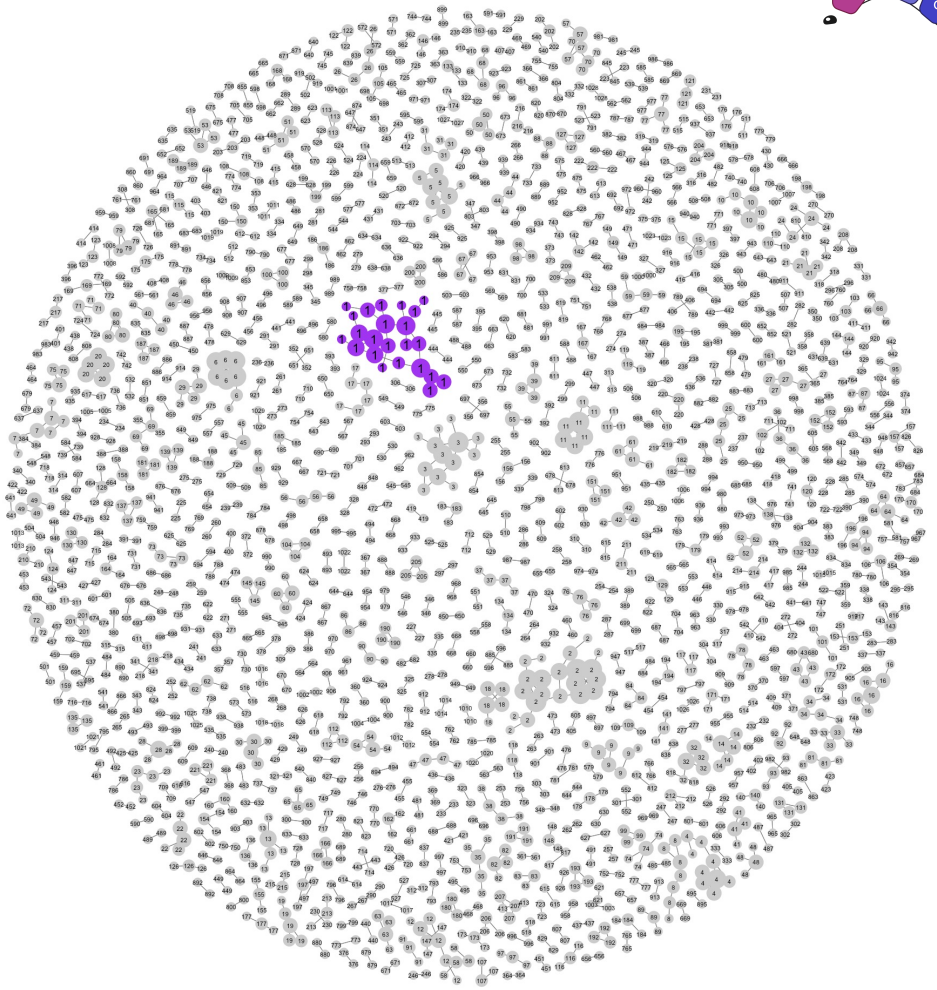
- **Network 1029 clusters**
  - 2 or more connections



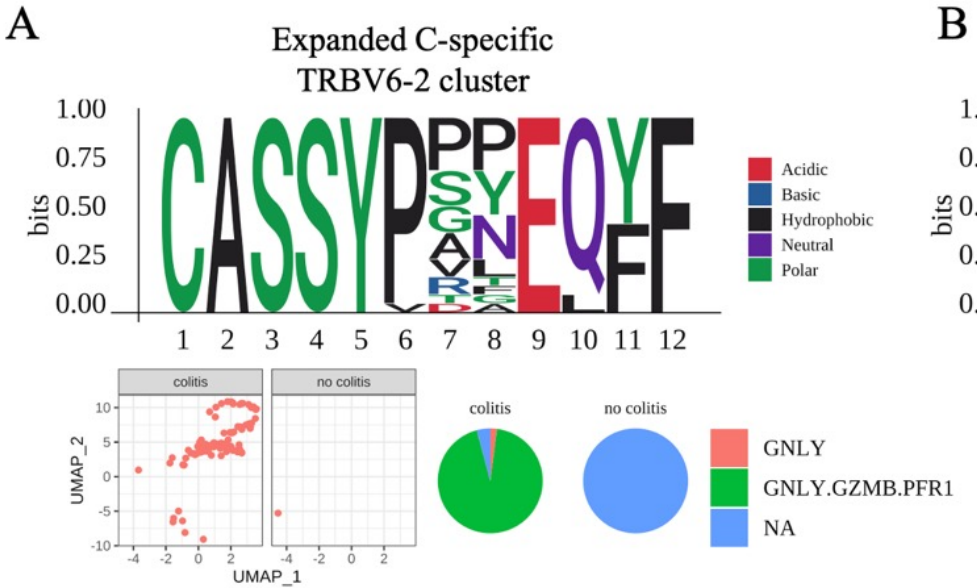
Top cluster

23 unique clones from the TRBV6-2

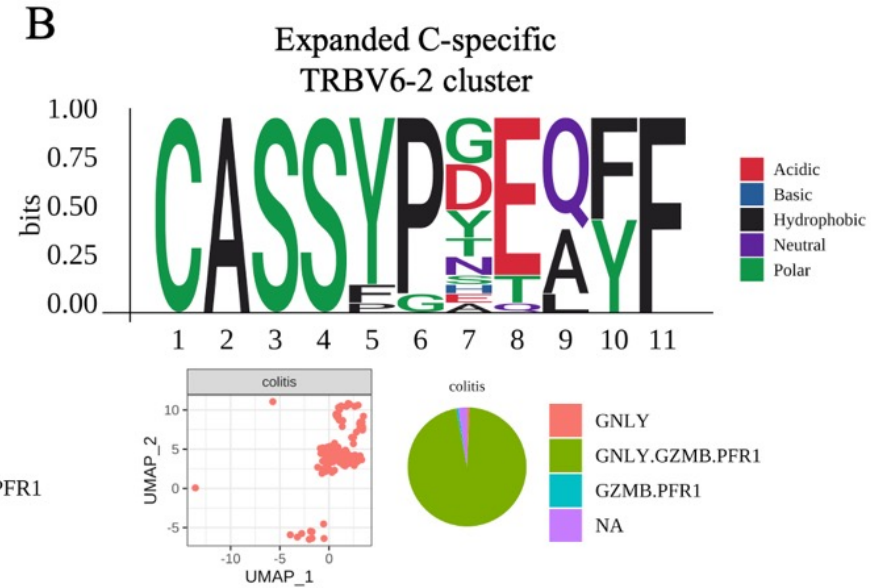




Hamming clustering identified colitis-specific TCR's



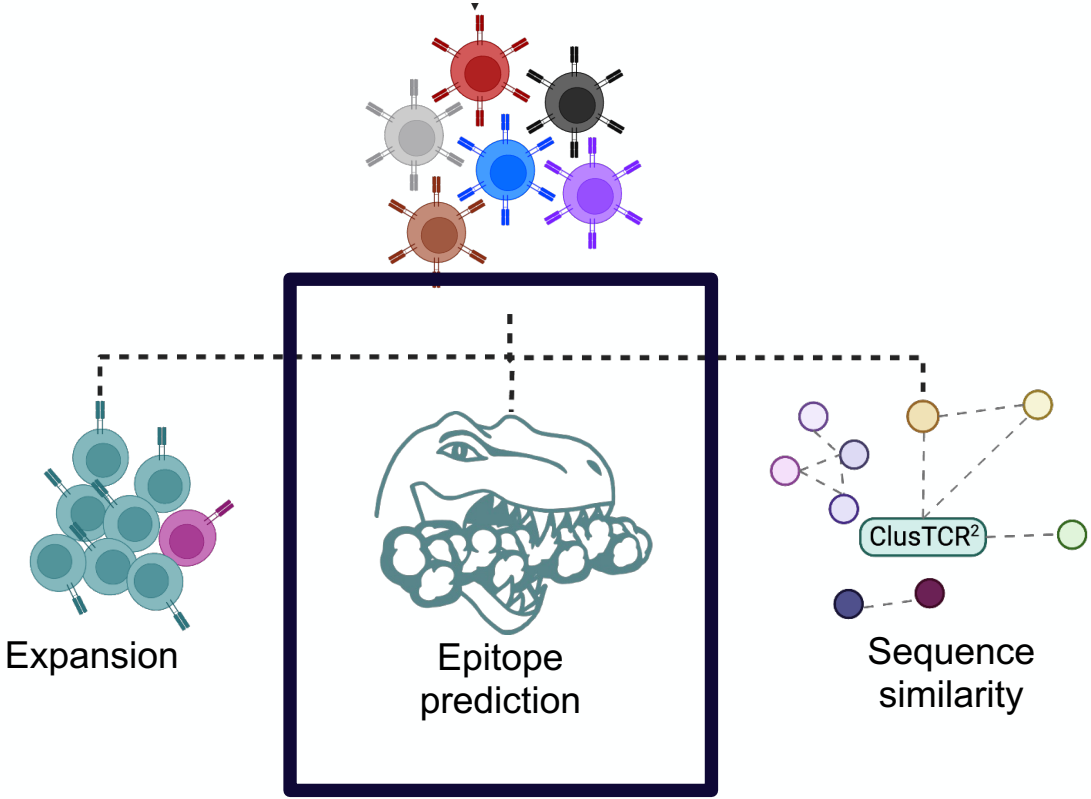
Cytotoxic profile



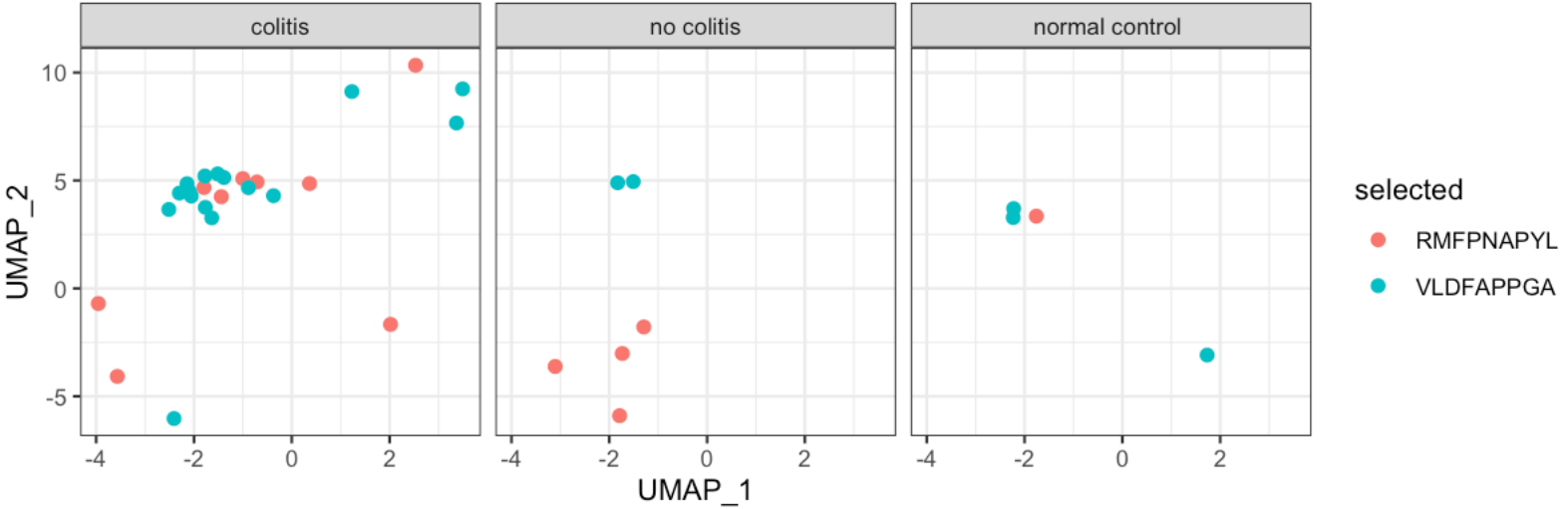
Cytotoxic profile

Mullan et al. bioRxiv (2023) **doi:** https://doi.org/10.1101/2023.09.27.559702

Pre-processing in STEGO.R (Shiny R package)

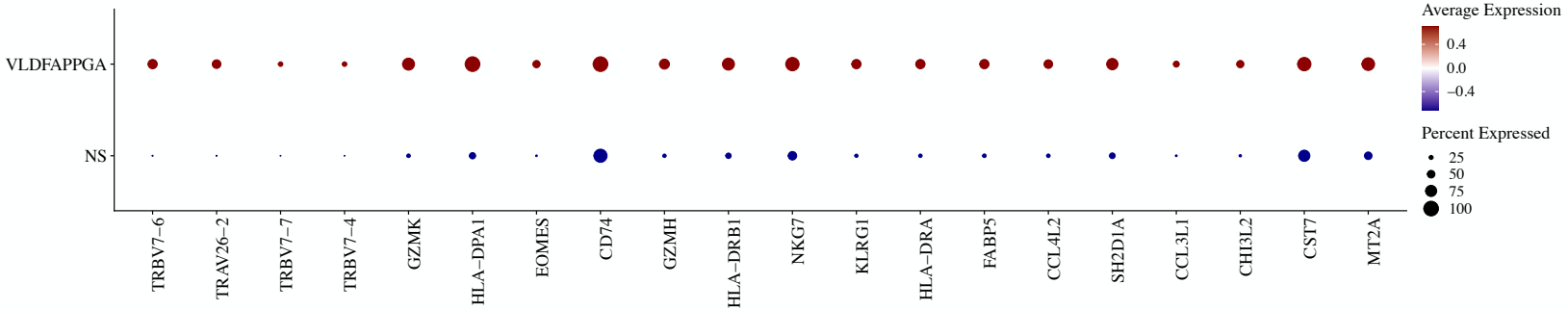


Predicted WT1-specific epitopes



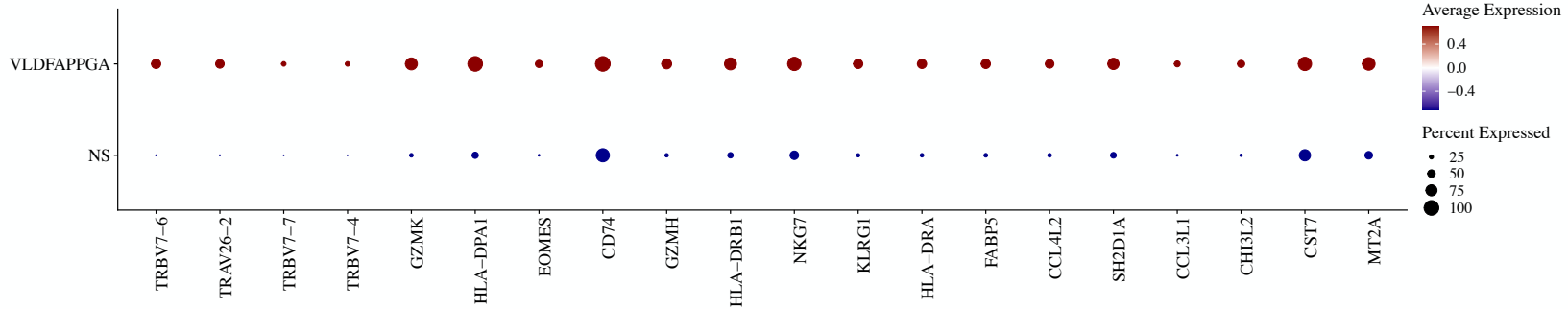
Possible WT1 epitope is associated with driving colitis ADR

VLDFAPPGA: Activation memory phenotype (GZMK, NKG7, KLRG1)



Possible WT1 epitope is associated with driving colitis ADR

VLDFAPPGA: Activation memory phenotype (GZMK, NKG7, KLRG1)



This was from an expanded cluster from the C5 colitis case (n=11)

TRAV26-2.TRAJ52_CILPLAGGTSYGKLTF & TRBV7-8.TRBJ2-7_CASSLGQAYEQYF

Focus on testing one peptide for this clone by HLA-A*02:01

Future directions

- Refining STEGO to include predicting the type of possible epitope
  - Peptides, lipids or small molecules

- Prediction unseen epitopes while incredibly difficult holds great promise for actionable insights

- Additional experimental data needed for:
  - Covering more HLA’s
  - Negative data for improving modelling
  - Validating the predictions

Take away messages

- Unannotated TCRs can be transparently matched to those in curated databases with simple TCR distance-based approaches.
- Machine learning tools allow us to use the limited data currently available to its fullest and identify additional TCR-epitope interactions that are otherwise impossible to find.
- Using either the distance or predictions with single cell GEX & TCR-seq can further narrow down which clones to experimentally validate.