Type 1 Diabetes project

Prediction of T1D status based on immunological features

Kseniia Lupyr

Artemiy Dakhnovets

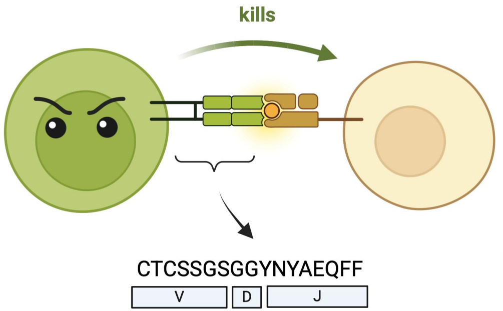
Valeriia Vladyina

March, 2025

Problem statement

Type I diabetes

The immune system of patients with T1D attacks their own body



Can we develop a therapy that targets crazy immune cells?

Dataset: tables of special immunological features. One table for a donor with sequences and their abundance

T1D

408

CASSGSN	10
CTSGLKAI	8
CAASKLCI	7
CARSAAN	5
CAYNPLSI	5
CTCSSSF	3
CTAYQEF	1

CASSGGK	70
CTSGKLSA	9
CAASQYFG	8
CARSLKQE	5
CAYLKERNF	5
CASSTCQE	3
CTAYRGNK	1

Healthy

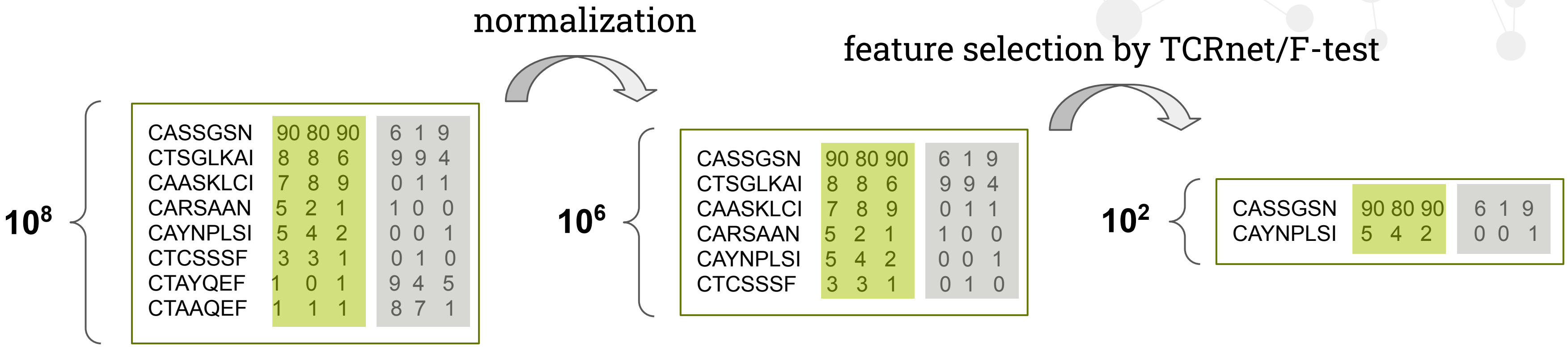
165

CASSGGY	8
CTSGYQE	4
CAASRRK	2
CARSLKF	1
CASSLWQ	1
CSVDSGD	1
CASSQGD	1

CSARERKLA	10
CSAPAGEDY	8
CASSSGNIQ	7
CASSPPGR	5
CASRTSGT	5
CASRTSGTY	3
CASSSGTR	1

Goal: prediction of T1D status based on immunological data and identification feature that have the greatest impact

Difficulty: an extreme number of features



10^8

CASSGSN	90	80	90	6	1	9
CTSGLKAI	8	8	6	9	9	4
CAASKLCI	7	8	9	0	1	1
CARSAAN	5	2	1	1	0	0
CAYNPLSI	5	4	2	0	0	1
CTCSSSF	3	3	1	0	1	0
CTAYQEF	1	0	1	9	4	5
CTAAQEF	1	1	1	8	7	1

10^6

CASSGSN	90	80	90	6	1	9
CTSGLKAI	8	8	6	9	9	4
CAASKLCI	7	8	9	0	1	1
CARSAAN	5	2	1	1	0	0
CAYNPLSI	5	4	2	0	0	1
CTCSSSF	3	3	1	0	1	0

10^2

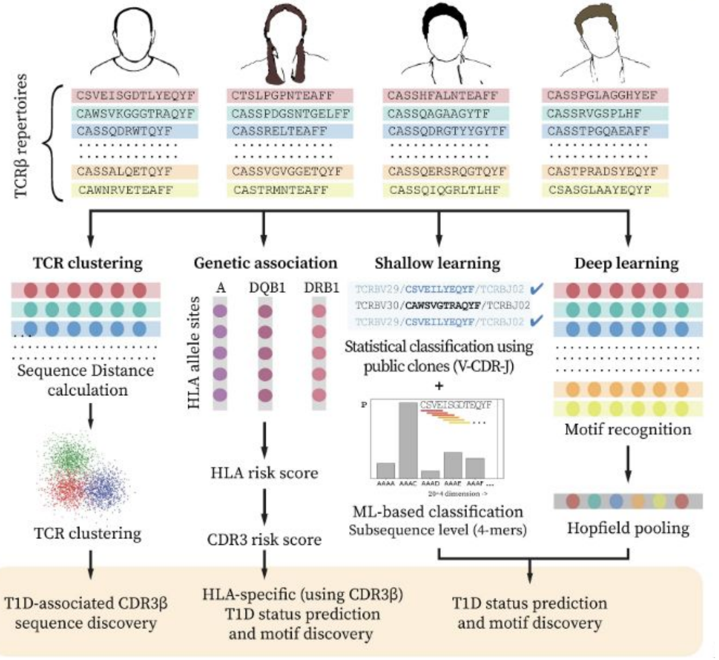
CASSGSN	90	80	90	6	1	9
CAYNPLSI	5	4	2	0	0	1

Test and training datasets were constructed on independent batches

dataset	status	batch	👥
train	T1D	T1D batch 2	230
	Healthy	rosati	66
test	T1D	T1D batch 1	153
	Healthy	aging	57

Current solutions

Preprint that is very similar to our work



medRxiv 2024.12.10.24318751

- No data available
- Machine learning analysis yielded AUROC of 0.77 on test cohort
- No immunological features that were shared between most of T1D patients

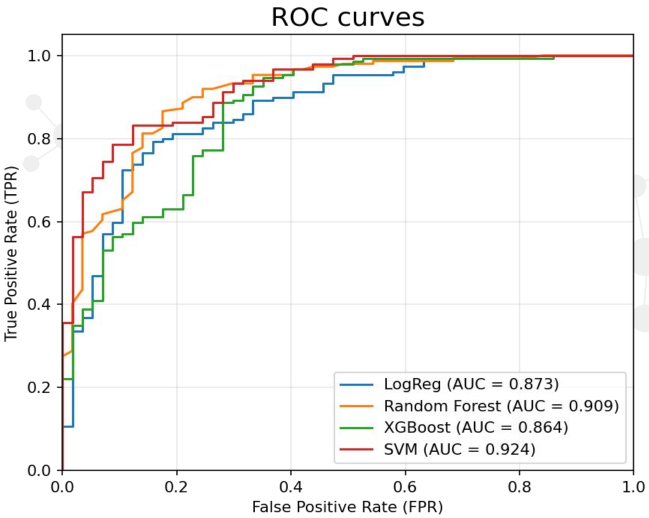
Novelty

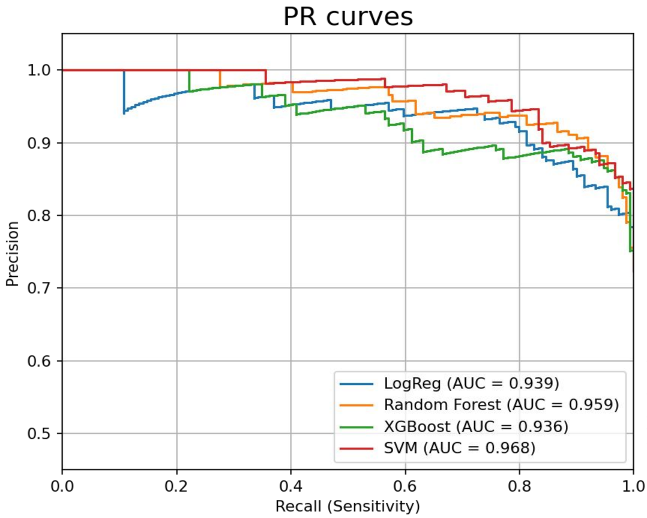
- We implemented reasonable methods from this work on **our data**
- Additional methods were employed
- Got better performance
- Identified immunological features that are shared between half of T1D patients

Models used with TCRnet and F-test selected features

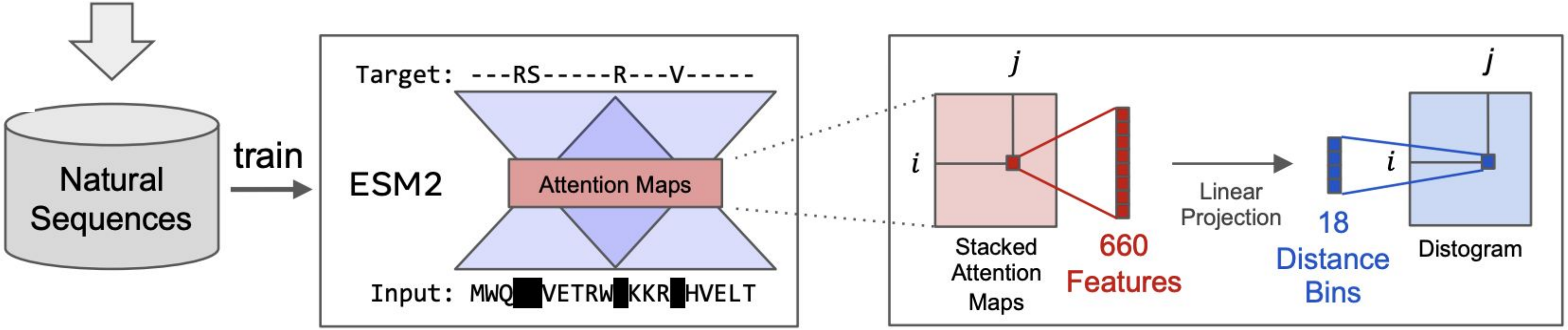
Feature Selection Method	F1	AUROC
TCRnet	0.85	0.68
F-test	**0.89**	**0.87**

Model	F1-score	AUROC
LogReg+ElasticNet	0.84	0.89
LogReg+L2	0.89	0.87
Random Forest	**0.91**	**0.91**
XGBoost	0.90	0.86
SVM	**0.90**	**0.92**


ROC curves


PR curves

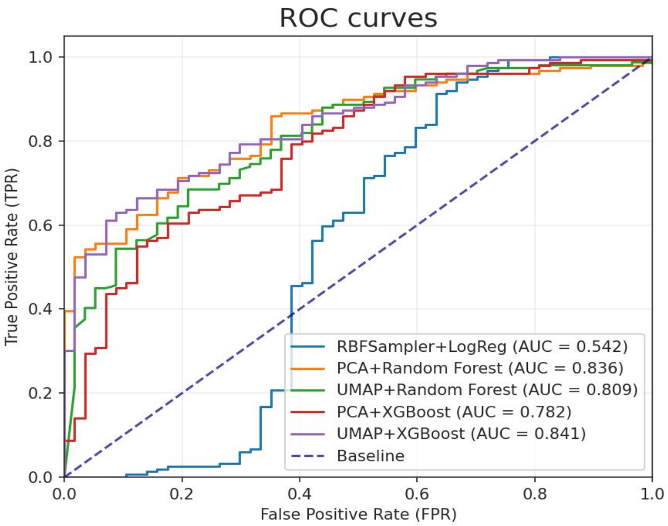
ESM-2 protein language model to construct sequence embeddings

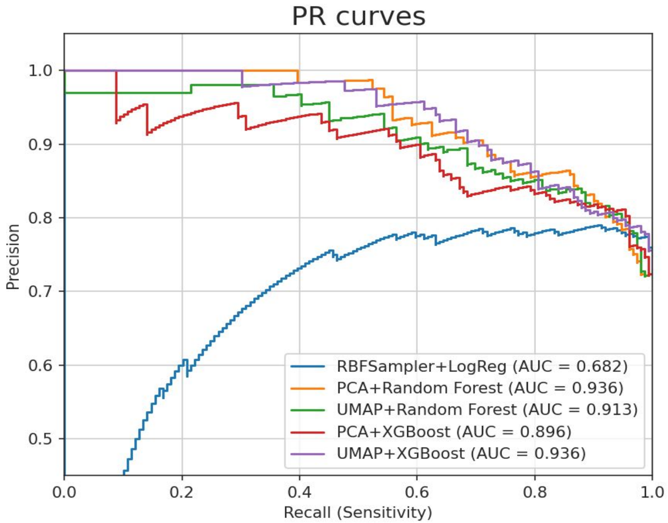


- We used pre-trained ESM-2 T33 UR50D model with 650 million parameters and 33 layers.
- First, per-sequence embeddings are mean-pooled across all tokens.
- Second, per-patient embeddings as weighted averaged per-sequence embeddings. Weights are immuno sequence abundances in a patient.
- Additional feature - entropy of immuno-sequence abundances to capture immunological sequence diversity.
- In total: 1280 embedding dims + 1 entropy = 1281 features.

Model evaluation on ESM-derived embeddings

Model	Balanced Acc	F1-score	AUROC
RBFSampler+LogReg	0.624	0.862	0.542
PCA+Random forest	0.562	0.845	0.836
UMAP+Random forest	0.516	0.852	0.782
PCA+XGBoost	0.588	0.852	0.782
UMAP+XGBoost	0.588	0.852	0.841





Conclusions

1. We implemented ML models following:
   a. statistical feature selection approach
   b. feature selection using immunological software
   c. deep learning scheme for feature engineering
2. Classical classification framework demonstrated higher performance compared to ESM-2 embedding approach
3. Random forest and SVM classifiers displayed the strongest performance with AUROC 0.92 and 0.91 respectively
4. We analyzed feature importance and identify immune features shared between half of the patients

Acknowledgments

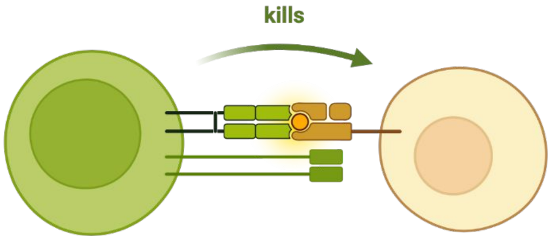
Georgy Sharonov
Irina Shagina
Mikhail Pogorelyy
Vladimir Zagainov
Dmitry Chudakov
Olga Britanova

пиму

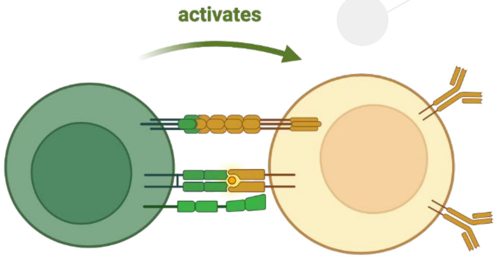
ИБХ

Supplementary

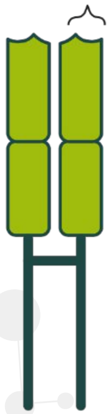
Introduction

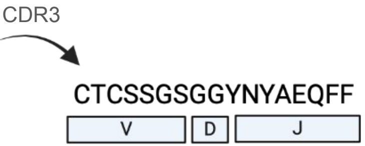


CD8+ cells kill infected and **cancer** cells

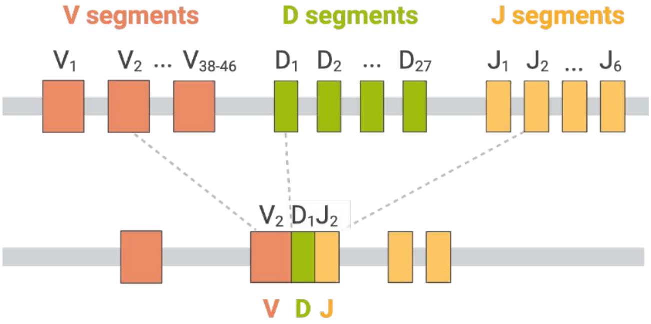


CD4+ cells regulate immune response by activating (**Th**) and inhibiting (**Treg**)





- T-cell specificity is defined by T-cell receptor (TCR)
- TCR is formed by quasi-stochastic process – **V(D)J recombination**
- Some TCRs have a greater generation probability than others



$$P_{gen}(\sigma) = P_V \times P_{DJ} \times P_{del} \times P_{ins}$$

Figures are created in BioRender

Skoltech

Type I diabetes

an autoimmune disease in which insulin-producing β-cells are destroyed by the immune system

8.7 million
people are living with T1D diabetes around the world

32 years
of healthy life lost on average per person

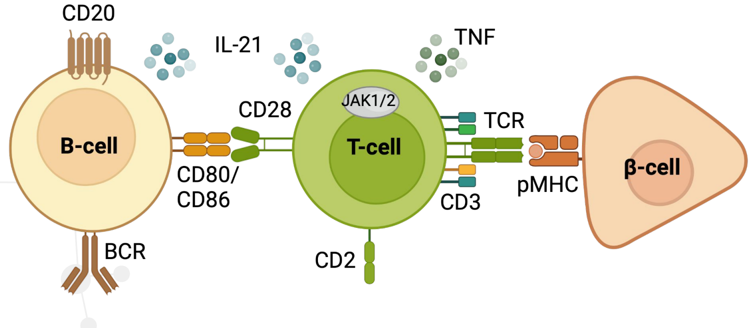
According to [3]

- **Insulin therapy** is the only one generally accepted method of treating T1D
- Insulin therapy does not prevent the development of severe chronic complications

Is T-cell targeted treatment possible for Type I Diabetes?

- T1D associated HLA haplotypes **DR3-DQ2** and **DR4-DQ8** are present in up to 90% of individuals with T1D [4,5]
- Genetic variations that are associated with a high expression of proinsulin in the thymus causes a T1D protective effect by enhancing T cell tolerance [6-8]
- T1D-associated gene variants are particularly enriched in the open chromatin of stimulated **CD4+ effector T cells** [9]

Figures are created in BioRender

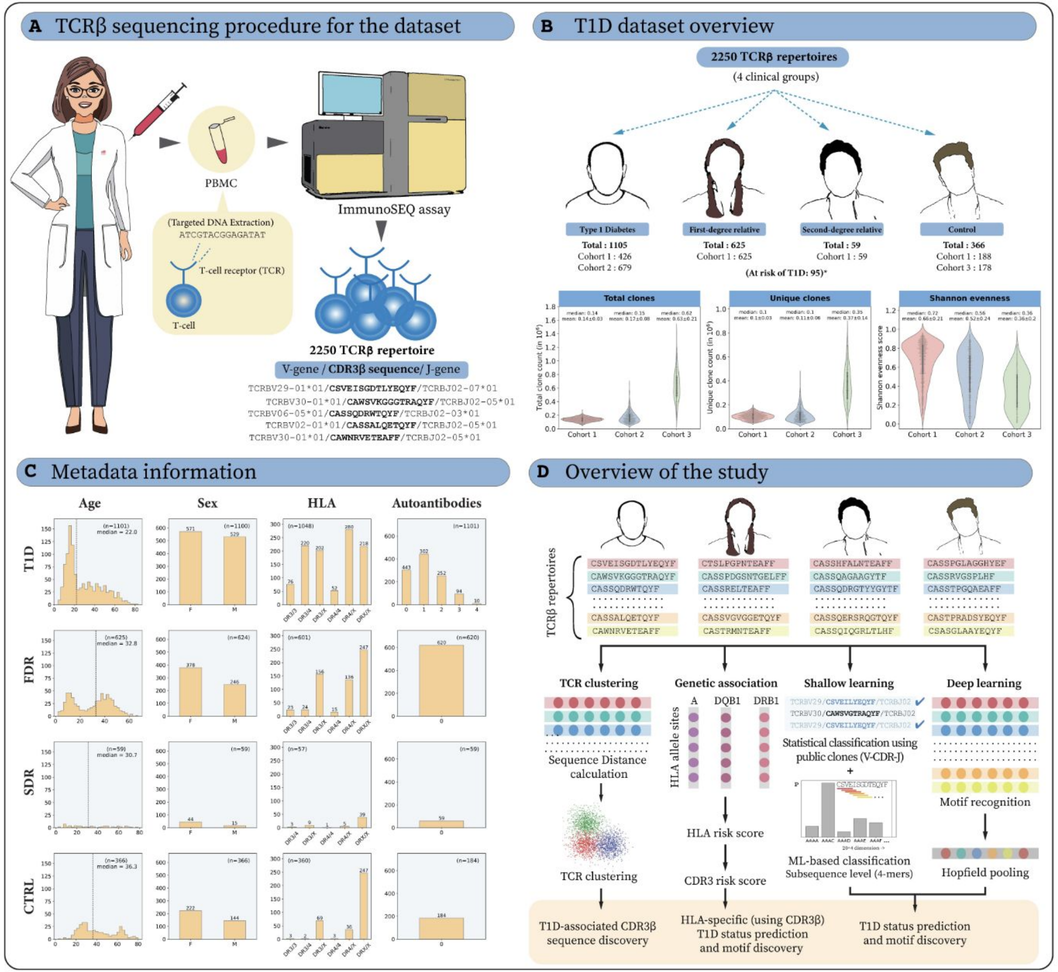


Drugs and mechanisms that have shown efficacy in TD1

- Anti-CD20 mAb
- blocking of CD28 costimulatory signals
- anti-thymocyte globulin
- anti-CD3 mAb
- blocking of CD2 costimulatory receptor
- anti-TNF mAb
- JAK1/JAK2 inhibition
- tyrosine kinase inhibitor

[10-19]

	Cohort 1
Model	**AUROC**
HLA risk score	0,7279
CDR3 risk scores	0,7533
Average CDR3 risk score	0,7146
pHLA-motif	0,6804
nHLA-motif	0,5869
Logistic regression (LR)	1
DeepRC	0,7603
Ensemble DeepRC (LR)	0,7894
DeepRC-motif	0,7054
Consensus-motif*	0,6844



https://www.medrxiv.org/content/10.1101/2024.12.10.24318751v1?s=35

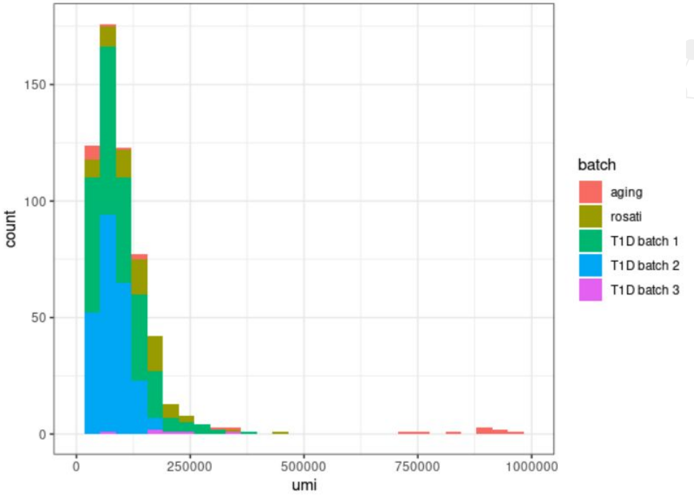
Bulk dataset description

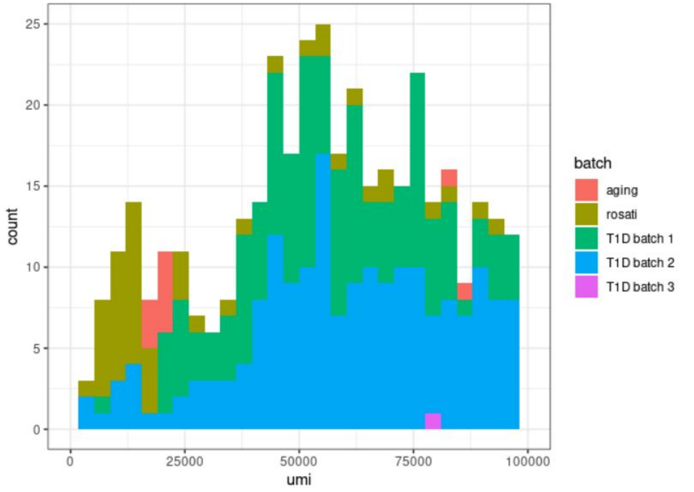
TCR repertoires of patients with T1D:
batch 1 – 158 TCR repertoires
batch 2 – **250** TCR repertoires
batch 3 – **3** patients, 2 TCR repertoires per patient

TCR repertoires of healthy patients:
Aging – **65** TCR repertoires
Rosati – **100** TCR repertoires

414 TCR repertoires of patients with T1D
165 TCR repertoire of healthy individuals

TCR repertoires were normalized to 30k UMIs per sample





Bulk dataset description

TCR repertoires of patients with T1D:
batch 1 – 158 TCR repertoires
batch 2 – **250** TCR repertoires
batch 3 – **3** patients, 2 TCR repertoires per patient

TCR repertoires of healthy patients:
Aging – **65** TCR repertoires
Rosati – **100** TCR repertoires

414 TCR repertoires of patients with T1D
165 TCR repertoires of healthy individuals

normalization

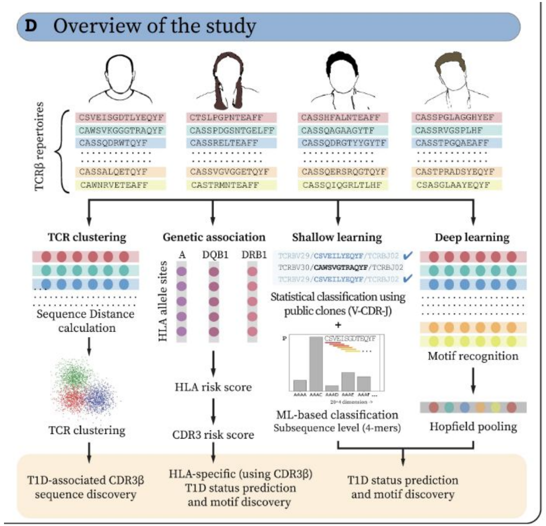
TCR repertoires of patients with T1D:
batch 1 – 153 TCR repertoires
batch 2 – **230** TCR repertoires
batch 3 – **3** patients, 2 TCR repertoires per patient

TCR repertoires of healthy patients:
Aging – **57** TCR repertoires
Rosati – **66** TCR repertoires

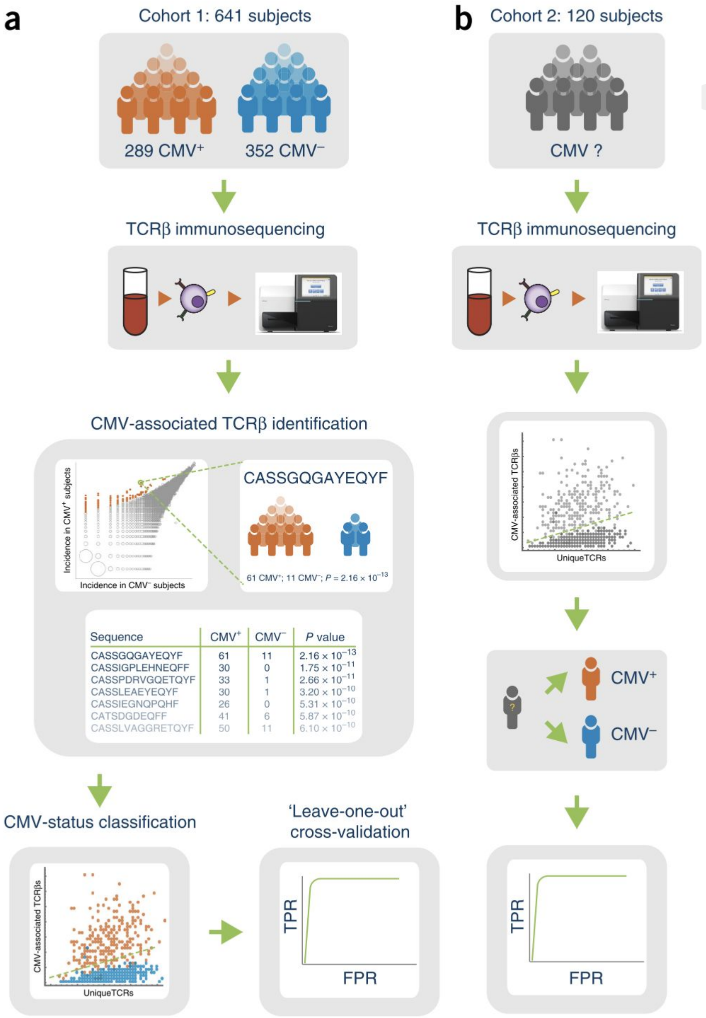
389 TCR repertoires of patients with T1D
123 TCR repertoires of healthy individuals

Classification of TCR repertoires by T1D status

Identification of T1D-associated TCRs as feature selection problem



medRxiv 2024.12.10.24318751



Sequence	CMV$^+$	CMV$^-$	*P* value
CASSGQGAYEQYF	61	11	2.16×10^{-13}
CASSIGPLEHNEQFF	30	0	1.75×10^{-11}
CASSPDRVGQETQYF	33	1	2.66×10^{-11}
CASSLEAEYEQYF	30	1	3.20×10^{-10}
CASSIEGNQPQHF	26	0	5.31×10^{-10}
CATSDGDEQFF	41	6	5.87×10^{-10}
CASSLVAGGRETQYF	50	11	6.10×10^{-10}

Emerson et al. 2017

Autoimmunity

Three levels of defence

1. Central tolerance
2. Peripheral tolerance
3. Low levels of self-peptides presentation by APCs

Target treatment of autoimmunity

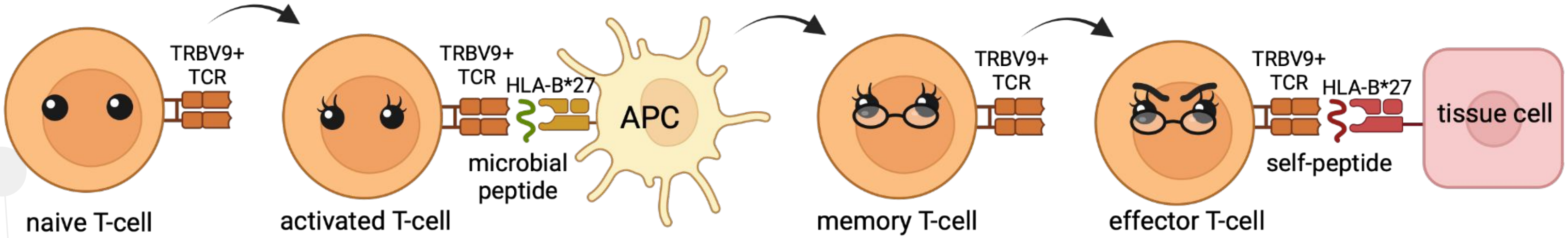
- Treg therapy
- Treg inducing-vaccines
- Depletion of autoimmune clonotypes

Therapeutic antibody for TRBV9+ T-cells depletion in patients with AS was registered in Russia in April

Figures are created in BioRender

Breaking self-tolerance

Ankylosing Spondylitis (AS) example



TRBV9+ TCR
naive T-cell
activated T-cell
TRBV9+ TCR
HLA-B*27
microbial peptide
APC
memory T-cell
TRBV9+ TCR
effector T-cell
TRBV9+ TCR
HLA-B*27
self-peptide
tissue cell