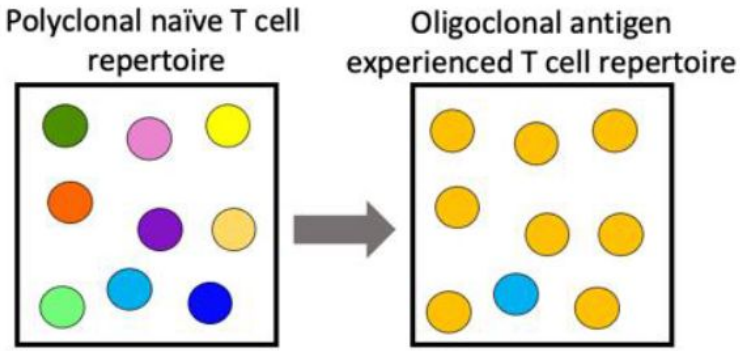
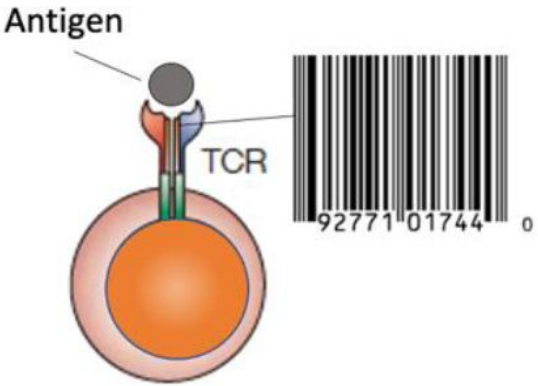
Фонд интеллект

# Машинное обучение в иммунологии и медицине

Данные в иммунологии и медицине. Работа с признаками

# Репертуары адаптивных иммунных рецепторов



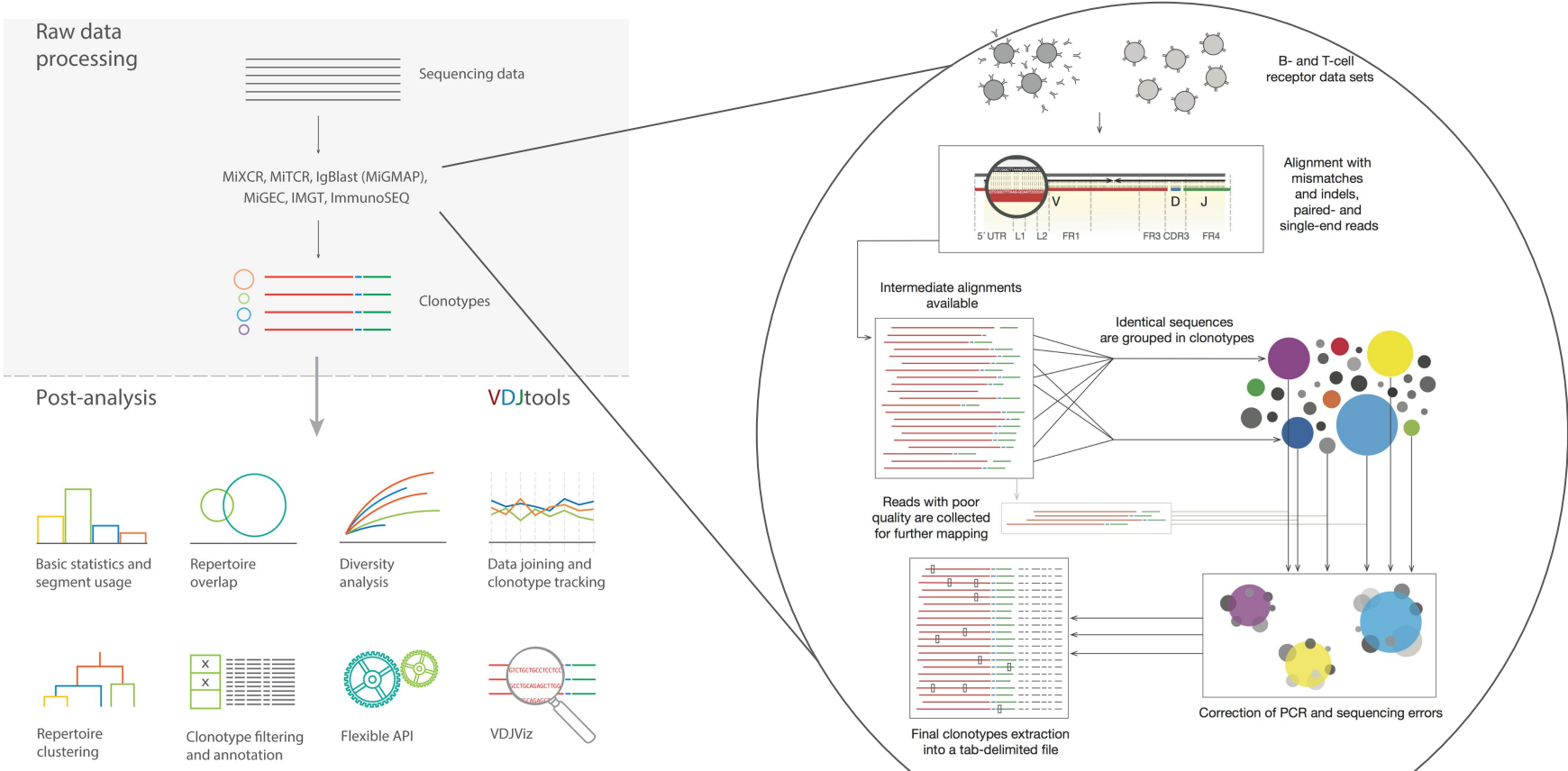

TCR представляет собой уникальный идентификатор клонов Т-клеток из-за низкой вероятности его сборки во второй Т-клетке у того же человека.

Увеличение частоты специфических TCR может служить показателем антигенспецифических иммунных реакция и соответствующей клональной экспансии Т-клеток с таким TCR.

# Пайплайн обработки данных

VDJtools

MiXCR

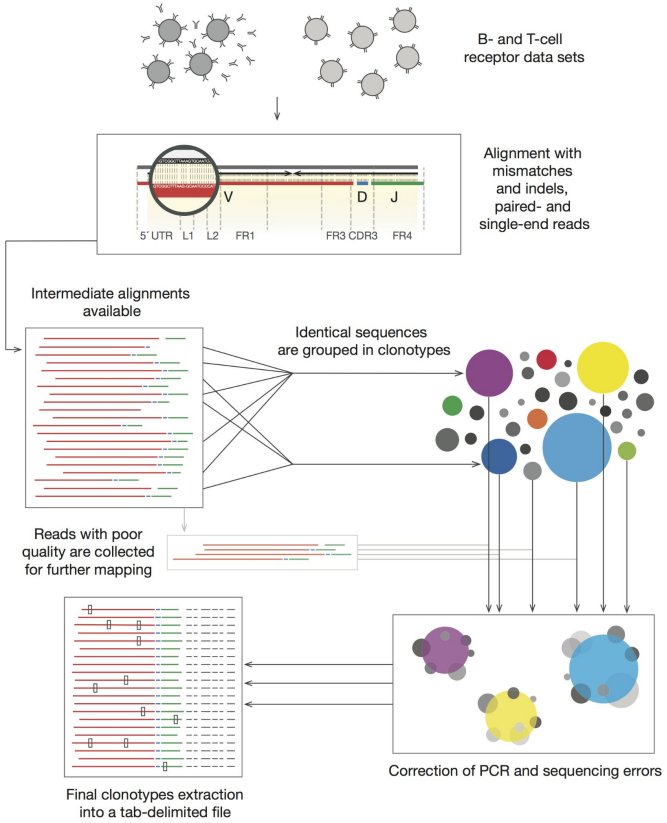


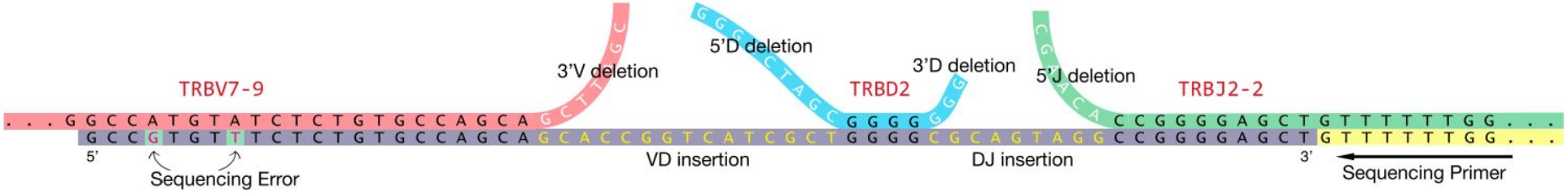
# MiXCR

Сырые прочтения выравниваются с референсным набором последовательностей для генов V, D и J сегментов

Идентичные последовательности группируются в **клонотипы**

Исправляются ошибки ПЦР и секвенирования, распределяются прочтения с плохим качеством

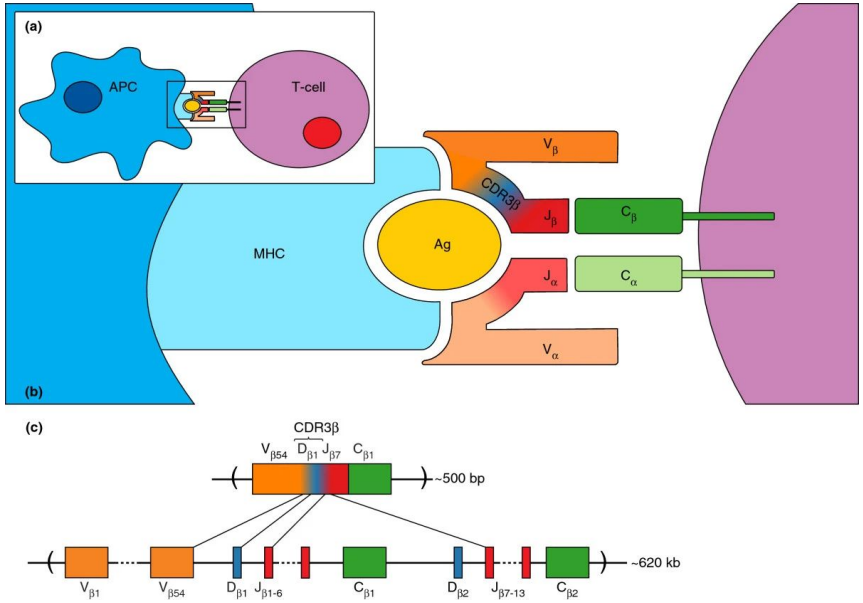


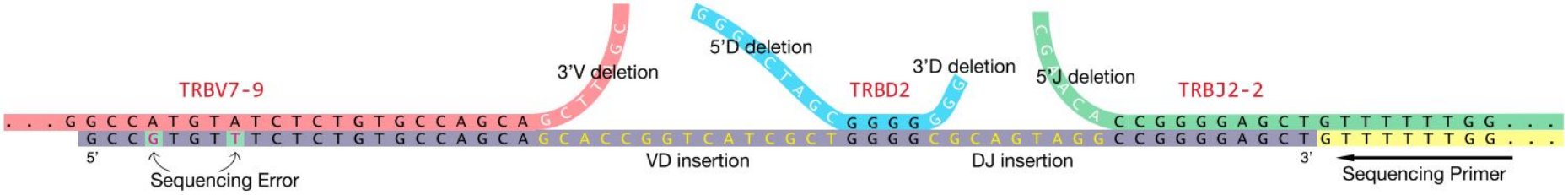



# Репертуары антигенных рецепторов

Какие признаки мы можем получить?

1. Нуклеотидная последовательность CDR3
2. Белковая последовательность CDR3
3. Используемые VDJ сегменты



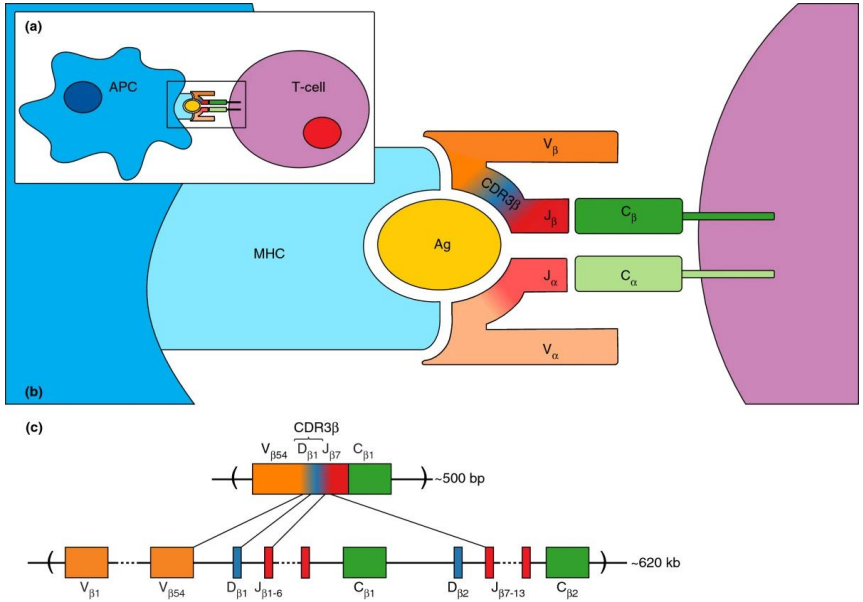




Statistical inference of the generation probability of T-cell receptors from sequence repertoires

# Репертуары антигенных рецепторов

Какие признаки мы можем получить?

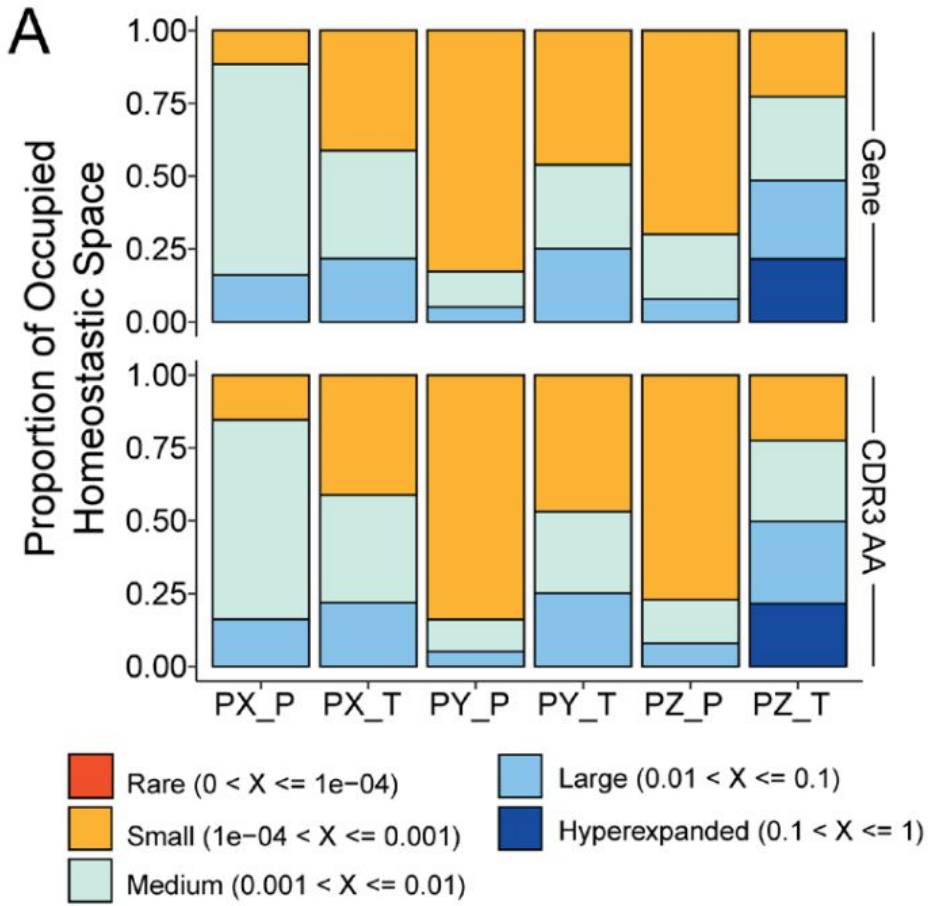
1. Нуклеотидная последовательность CDR3
2. Белковая последовательность CDR3
3. Используемые VDJ сегменты

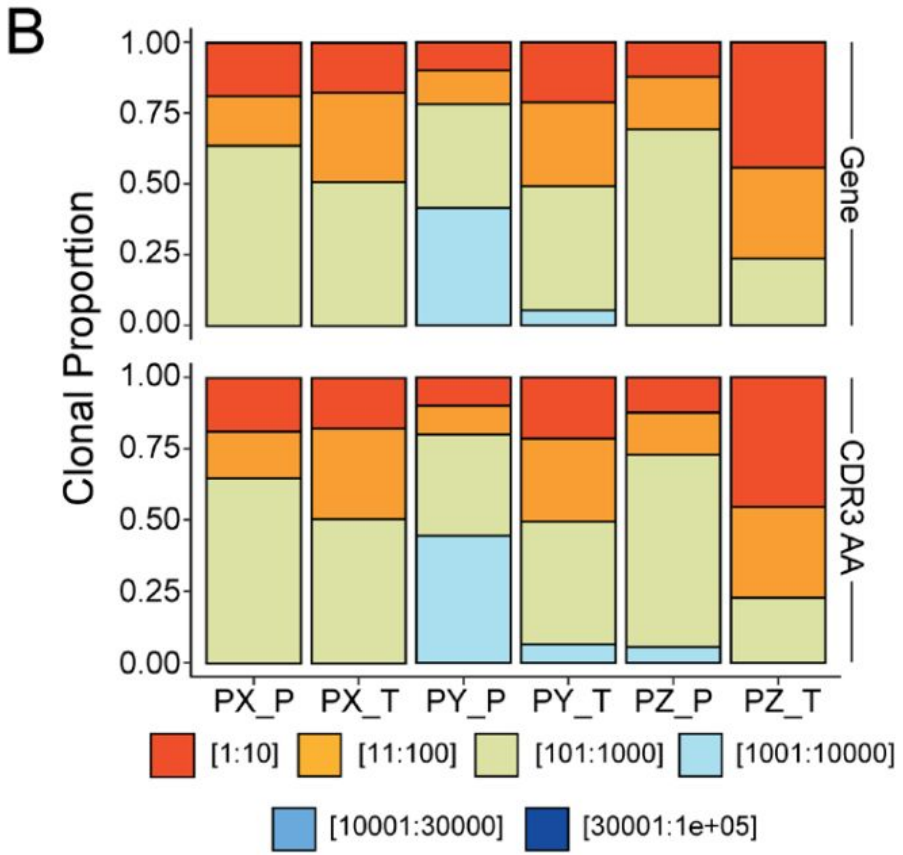



| | cloneCount | cloneFraction | targetSequences | allVHitsWithScore | allDHitsWithScore | allJHitsWithScore | allCHitsWithScore | aaSeqImputedCDR3 |
|---|---|---|---|---|---|---|---|---|
| 0 | 282.0 | 0.194617 | TTTCTGCTGGGTGCGACAGGCCCCTGGACTCGGACTCGAGTGGAT,... | IGHV1-69*00(90.3),IGHV1-69D*00(89.3) | IGHD7-27*00(35),IGHD3-16*00(30) | IGHJ5*00(54),IGHJ4*00(53) | IGHG1*00(155.3),IGHG3*00(141.4) | CARDLNRPIALGDVHW |
| 1 | 126.0 | 0.086957 | CCGAAACTGGGTCCGCCAGCCCCCAGGAAAGGGGCTGGAGTGGATT... | IGHV4-34*00(197.6) | IGHD6-19*00(60) | IGHJ4*00(54.2) | IGHG3*00(21.7) | CATDSSSGWYW |
| 2 | 116.0 | 0.080055 | AACCCACACAGACCCTCACGCTGACCTGCACCTTCTCTGGGTTCTC... | IGHV2-5*00(144.1) | IGHD7-27*00(50),IGHD5-24*00(45) | IGHJ4*00(67.5) | IGHM*00(89.7) | CVYNYWGSSLDYW |
| 3 | 95.0 | 0.065562 | TGGTACAGCCTGGGGGGTCCCTGAGACTCTCCTGTGCAGCCTCTGG... | IGHV3-23*00(156.2) | IGHD2-21*00(54) | IGHJ4*00(69.7) | IGHG1*00(112.9),IGHG3*00(104.1) | CAKDRDRVVVSANLDYW |
| 4 | 70.0 | 0.048309 | CGTGGTACAGCCTGGGGGGTCCCTGAGACTCTCCTGTGCAGCCTCT... | IGHV3-43*00(242.8) | IGHD4-23*00(50),IGHD4-17*00(45) | IGHJ6*00(53.7) | IGHG1*00(28) | CAKDSNYGGYYYGMDVW |

Sequence analysis of T-cell repertoires in health and disease

# Характеристики репертуаров

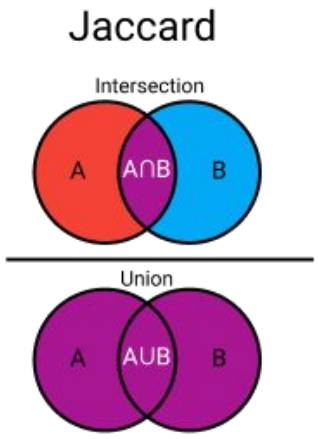


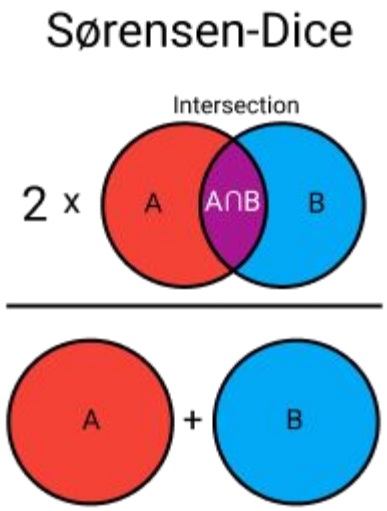

scRepertoire: An R-based toolkit for single-cell immune receptor analysis

# Сравнение репертуаров TCR между образцами

Популярные методы: [T-cell repertoire analysis and metrics of diversity and clonality]

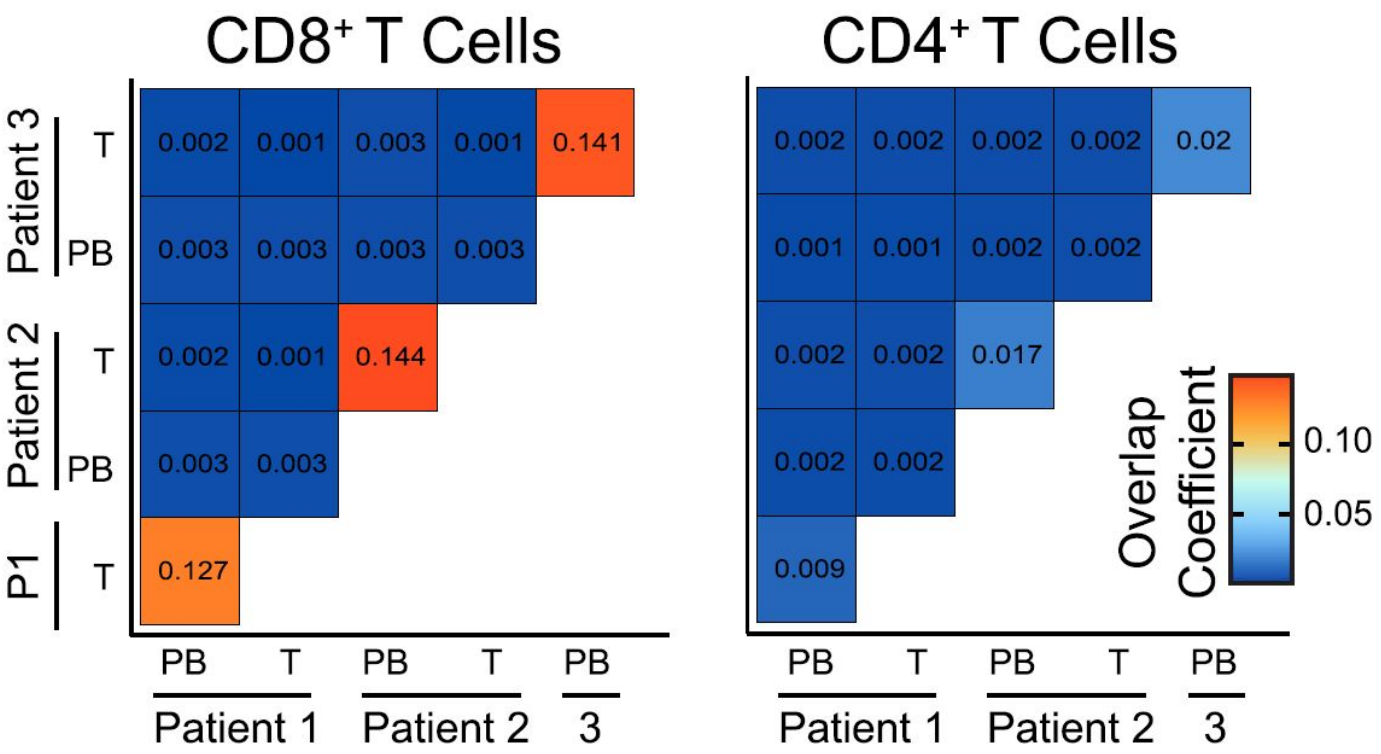



Jaccard index




Sorensen index

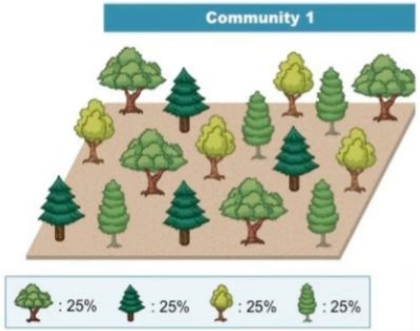
# Перекрывание репертуаров

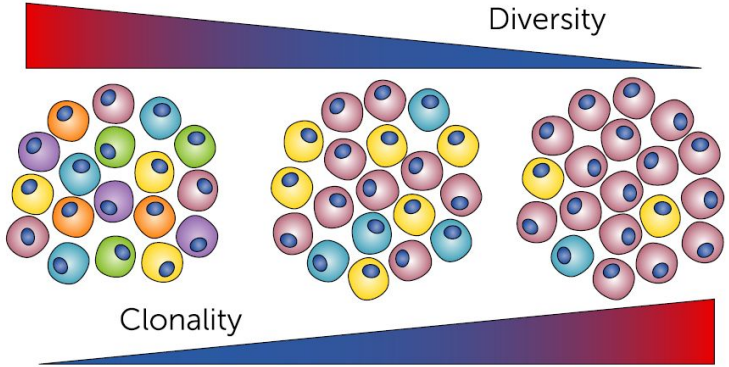


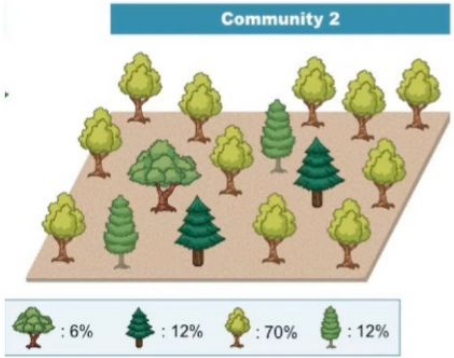

PB - peripheral blood
T - tumour

scRepertoire: An R-based toolkit for single-cell immune receptor analysis

# Клональность и разнообразие




Chao1 index

$$S_1 = S_{obs} + \frac{F_1^2}{2F_2}$$

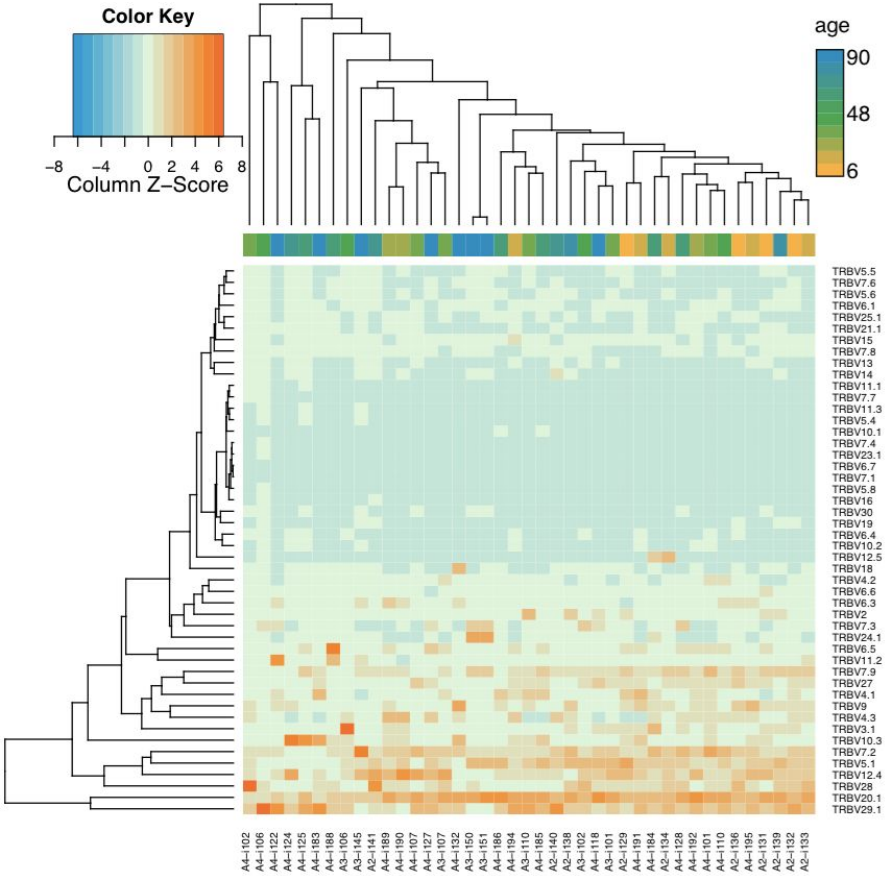
$S_{obs}$ - число видов в выборке
$F_1$ - число синглтонов
$F_2$ - число даблтонов

Shannon-Wiener index

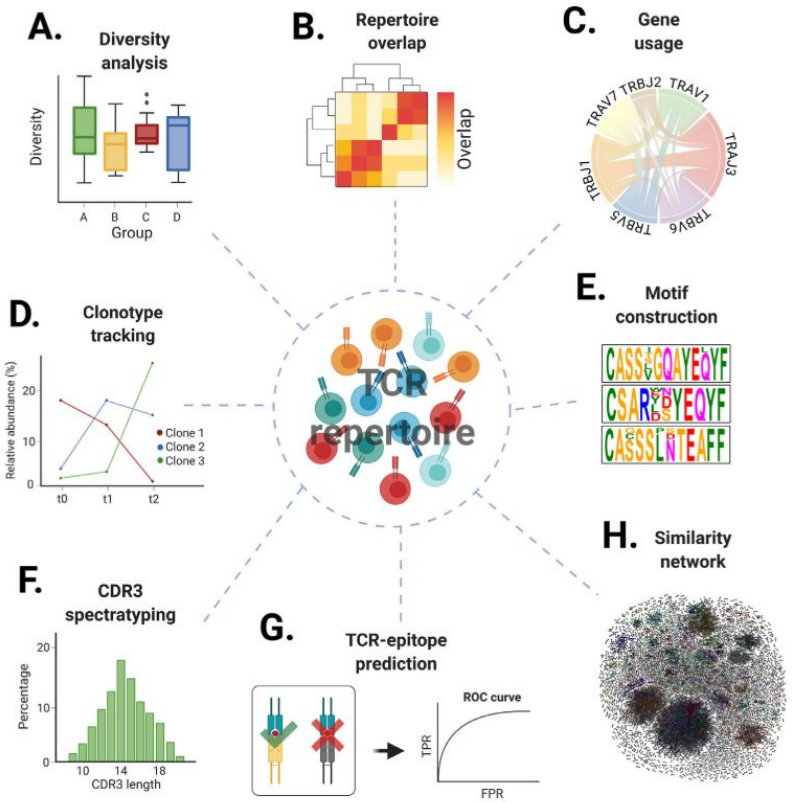
$$H' = \sum_{i=1}^{s} (p_i)(\ln p_i)$$

$p_i$ - доля представителей одного вида

# VDJ usage

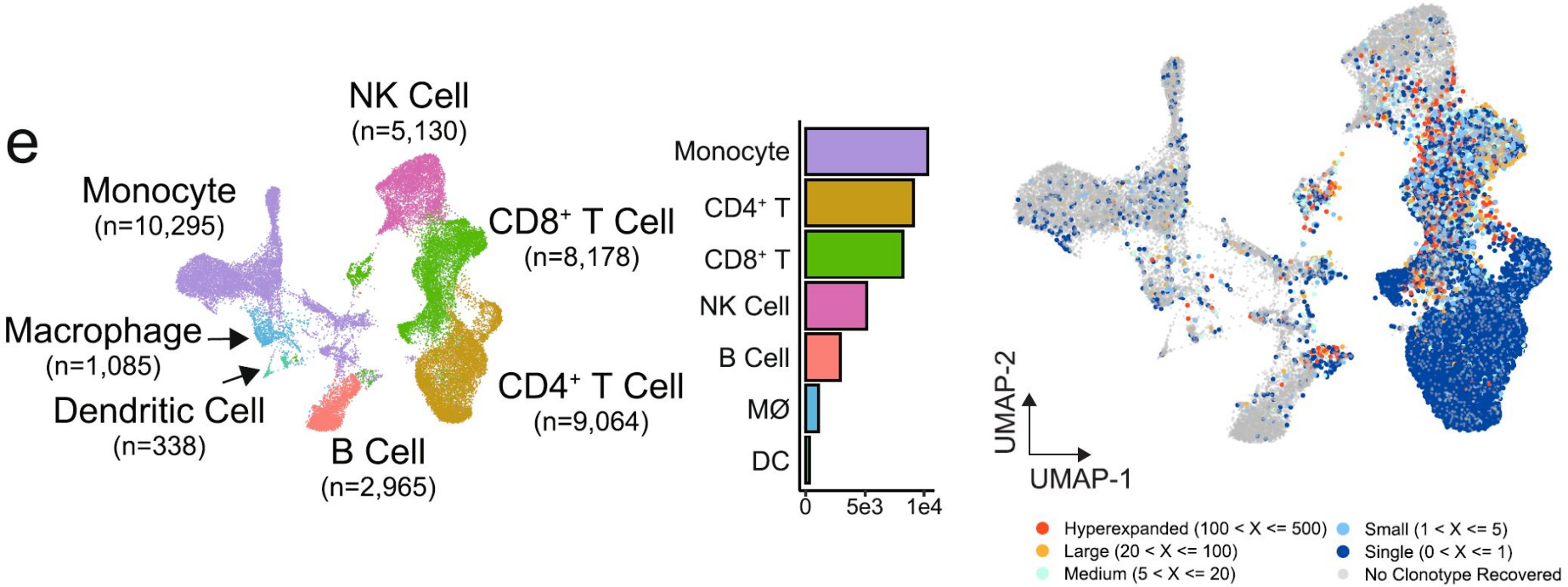

Color Key
−8 −4 0 2 4 6 8
Column Z−Score
age
90
48
6
TRBV5.5
TRBV7.6
TRBV5.6
TRBV6.1
TRBV25.1
TRBV21.1
TRBV15
TRBV7.8
TRBV13
TRBV14
TRBV11.1
TRBV7.7
TRBV11.3
TRBV5.4
TRBV10.1
TRBV7.4
TRBV23.1
TRBV6.7
TRBV7.1
TRBV5.8
TRBV16
TRBV30
TRBV19
TRBV6.4
TRBV10.2
TRBV12.5
TRBV18
TRBV4.2
TRBV6.6
TRBV6.3
TRBV2
TRBV7.3
TRBV24.1
TRBV6.5
TRBV11.2
TRBV7.9
TRBV27
TRBV4.1
TRBV9
TRBV4.3
TRBV3.1
TRBV10.3
TRBV7.2
TRBV5.1
TRBV12.4
TRBV28
TRBV20.1
TRBV29.1
A4-I102
A4-I106
A4-I122
A4-I124
A4-I125
A4-I183
A4-I188
A3-I106
A3-I145
A2-I141
A4-I189
A4-I190
A4-I107
A4-I127
A3-I107
A4-I132
A3-I150
A3-I151
A4-I186
A4-I194
A3-I110
A4-I185
A2-I140
A2-I138
A3-I102
A4-I118
A3-I101
A2-I129
A4-I191
A4-I184
A2-I134
A4-I128
A4-I192
A4-I101
A4-I110
A2-I136
A4-I195
A2-I131
A2-I139
A2-I132
A2-I133

# Что еще можно делать с репертуарами антигенных рецепторов?



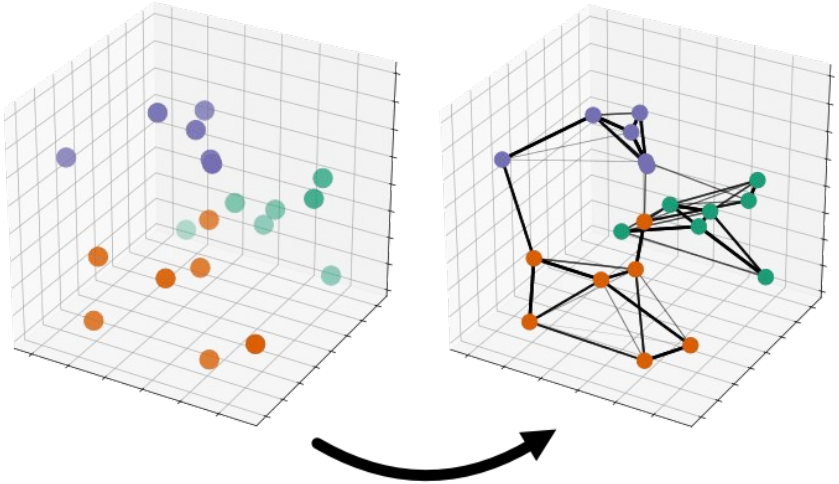

[Recent advances in T-cell receptor repertoire analysis: Bridging the gap with multimodal single-cell RNA sequencing]

# Использование данных о репертуарах адаптивных иммунных рецепторов совместно с данными экспрессии генов

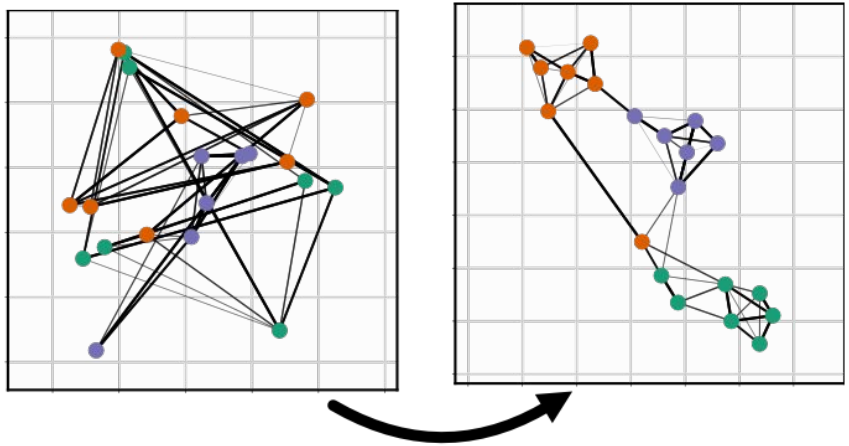



scRepertoire: An R-based toolkit for single-cell immune receptor analysis

# UMAP

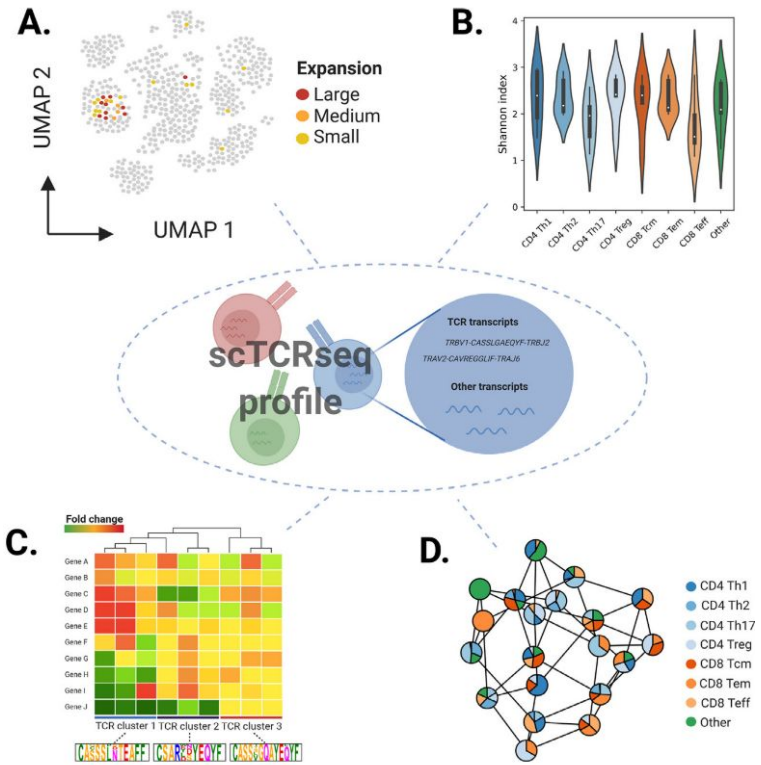


Compute a graphical representation
of the dataset



Learn an embedding that preserves
the structure of the graph

# Использование данных о репертуарах адаптивных иммунных рецепторов совместно с данными экспрессии генов



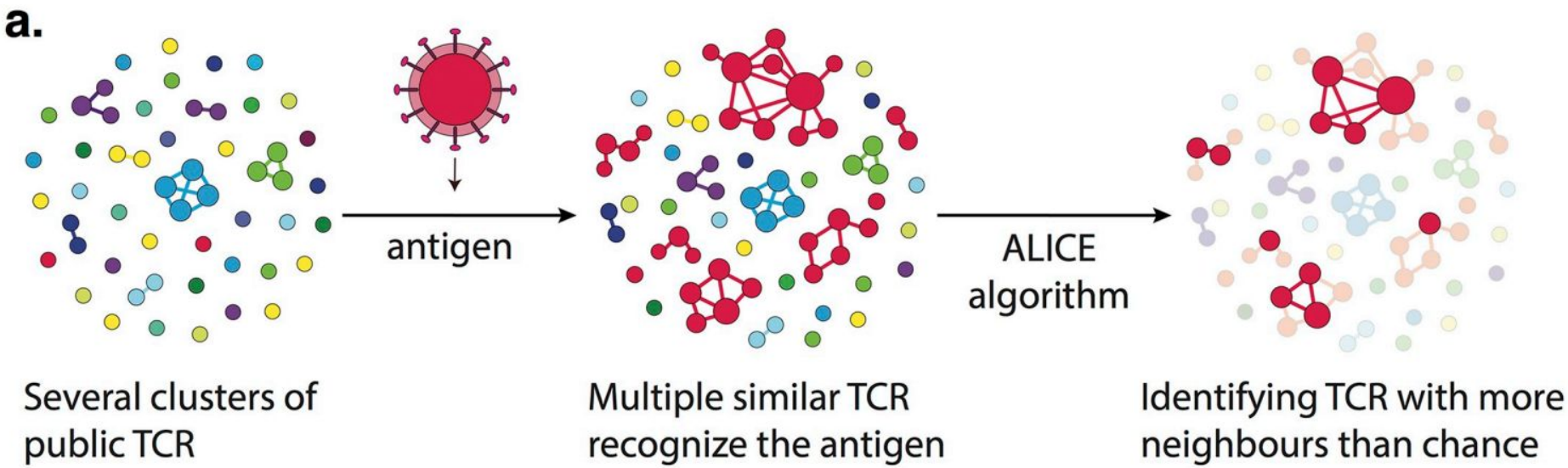

[Recent advances in T-cell receptor repertoire analysis: Bridging the gap with multimodal single-cell RNA sequencing]

# Вероятностная модель сборки TCR

$$
\begin{aligned}
P(\text{scenario}) = {} & P(V)P(J|V)\,P(D|V,J)P(\text{del}V|V) \\
& \times\ P(\text{del}J|J)\,P(\text{delD5'}|D)\,P(\text{delD3'}|\text{delD5'},D) \\
& \times\ P(\text{insVD}) \prod_{i}^{\text{InsVD}} P(n_i|n_{i-1}) \\
& \times\ P(\text{insDJ}) \prod_{i}^{\text{InsDJ}} P(n_i|n_{i-1})
\end{aligned}
$$

Statistical inference of the generation probability of T-cell receptors from sequence repertoires

# ALICE (Antigen-specific Lymphocyte Identification by Clustering of Expanded sequences)




Ищем обогащенные клоны в репертуаре, используя синтетический репертуар в качестве бэкграунда

# Репертуары антигенных рецепторов: featurization

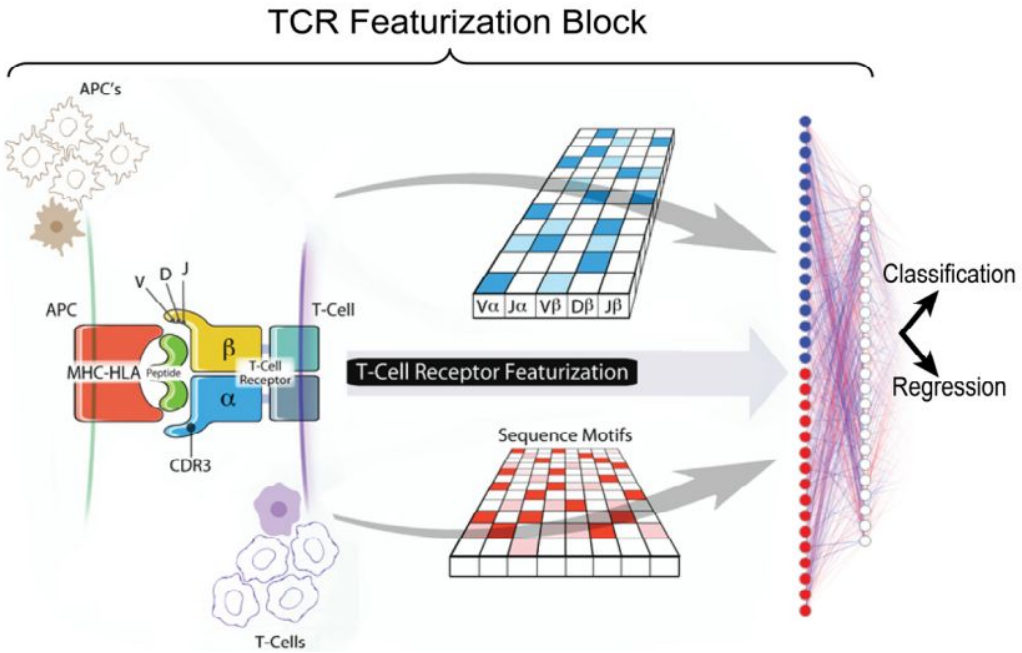
Работа: [DeepTCR is a deep learning framework for revealing sequence concepts within T-cell repertoires]

Задача: классификация TCR (например, антигенная специфичность) или регрессия (аффинность связывания)

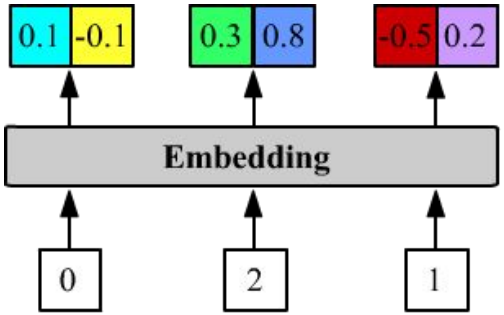
Featurization:

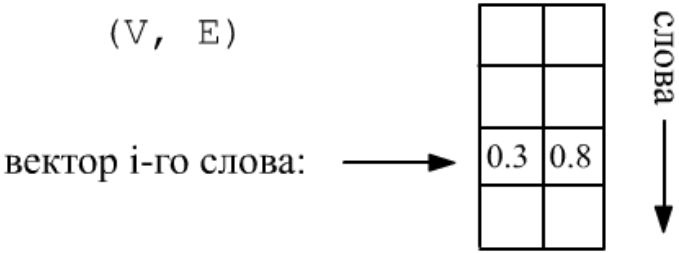
1. CDR3 последовательности $\alpha$ и $\beta$ цепи TCR
2. VDJ usage

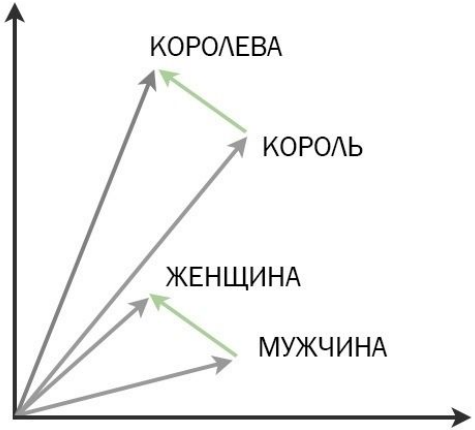
Используется эмбеддинг



# Эмбеддинг


0.1
-0.1
0.3
0.8
-0.5
0.2
Embedding
0
2
1


(V, E)
вектор i-го слова:
0.3
0.8
слова


КОРОЛЕВА
КОРОЛЬ
ЖЕНЩИНА
МУЖЧИНА

# Репертуары антигенных рецепторов: featurization

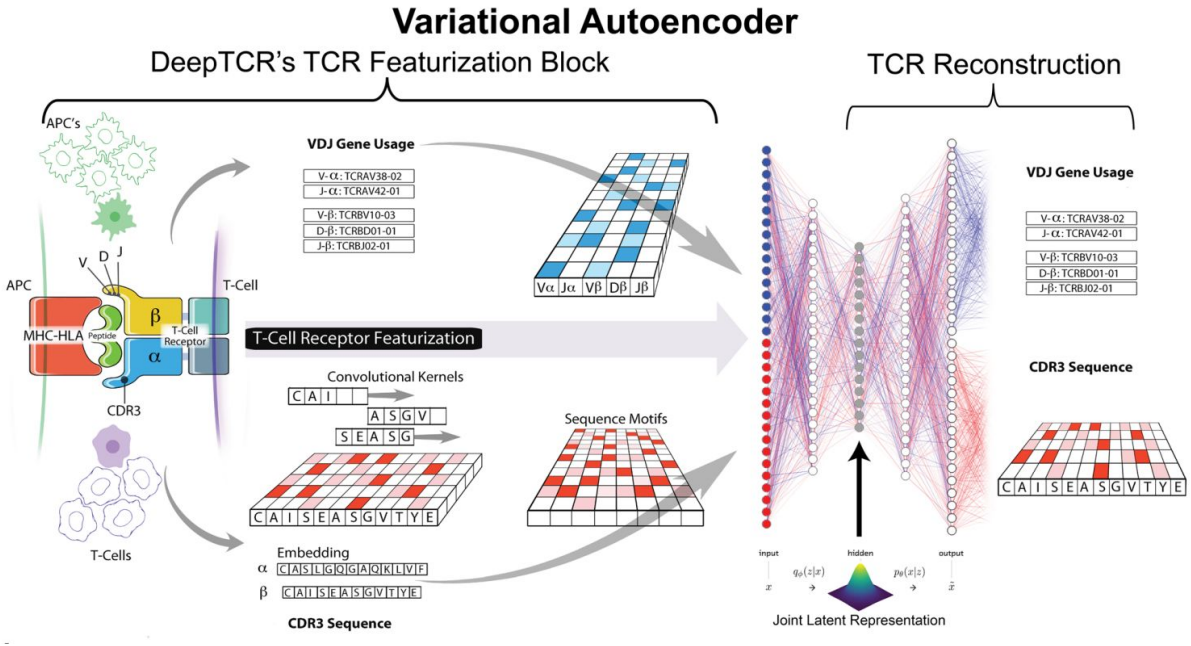
Работа: [DeepTCR is a deep learning framework for revealing sequence concepts within T-cell repertoires]

Задача: получение представлений TCR

Featurization:

1. CDR3 последовательности $\alpha$ и $\beta$ цепи TCR
2. VDJ usage

Что может использоваться помимо эмбеддинга?



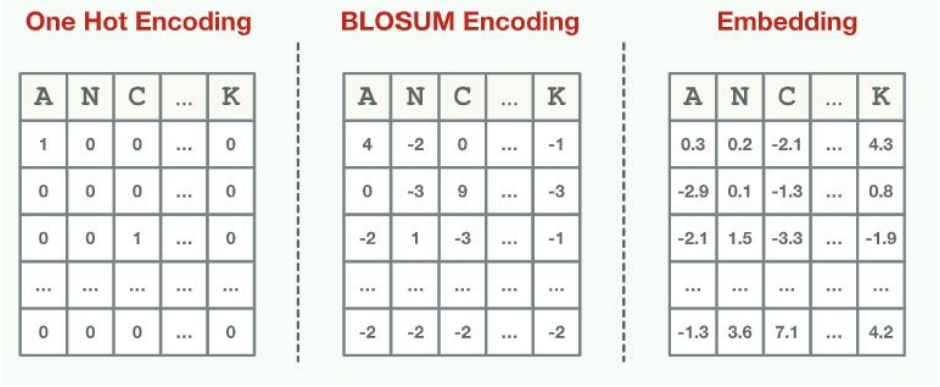
# Предсказание аффинности связывания MHC:белок

Работа: [NetMHCpan, a method for MHC class I binding prediction beyond humans]

Задача: Предсказание аффинности для любой пары пептид:MHC

Featurization:

1. One-hot кодирование
2. BLOSUM кодирование
3. Представление молекулы MHC в виде псевдо-последовательностей



# Использование физических характеристик TCR

Работа: [On the feasibility of mining CD8+ T cell receptor patterns underlying immunogenic peptide recognition]

Задача: Распознавание комплекса пептид:MHC TCR

Featurization:

1. CDR3 Shannon-Wiener diversity
2. длина последовательности TCR CDR3
3. общая масса аминокислот в последовательности CDR3
4. средняя основность, гидрофобность, изоэлектрическая точка CDR3
5. физические характеристики индивидуальных аминокислот

Для отбора признаков используется Boruta

# Boruta

1. Теневые признаки

| | age | height | weight |
|---|---|---|---|
| **0** | 25 | 182 | 75 |
| **1** | 32 | 176 | 71 |
| **2** | 47 | 174 | 78 |
| **3** | 51 | 168 | 72 |
| **4** | 62 | 181 | 86 |

| | income |
|---|---|
| **0** | 20 |
| **1** | 32 |
| **2** | 45 |
| **3** | 55 |
| **4** | 61 |

→

| | age | height | weight | shadow_age | shadow_height | shadow_weight |
|---|---|---|---|---|---|---|
| **0** | 25 | 182 | 75 | 51 | 176 | 75 |
| **1** | 32 | 176 | 71 | 32 | 182 | 71 |
| **2** | 47 | 174 | 78 | 47 | 168 | 78 |
| **3** | 51 | 168 | 72 | 25 | 181 | 72 |
| **4** | 62 | 181 | 86 | 62 | 174 | 86 |

- Измерение важности признаков относительно порога

| | age | height | weight | shadow_age | shadow_height | shadow_weight |
|---|---|---|---|---|---|---|
| **feature importance %** | 39 | 19 | 8 | 11 | 14 | 9 |
| **hits** | 1 | 1 | 0 | - | - | - |

→

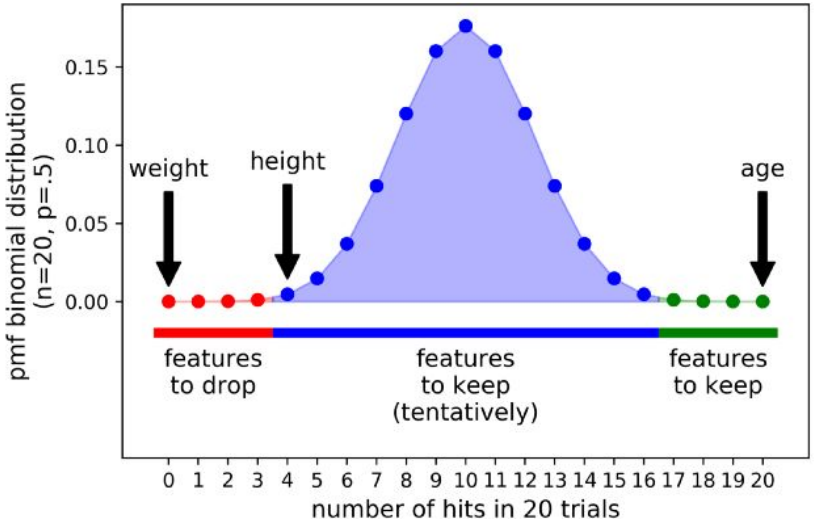
| | age | height | weight |
|---|---|---|---|
| **hits (in 20 trials)** | 20 | 4 | 0 |

[Boruta Explained Exactly How You Wished Someone Explained to You]

# Boruta

- Измерение важности признаков относительно порога

| | age | height | weight | shadow_age | shadow_height | shadow_weight |
|---|---|---|---|---|---|---|
| feature importance % | 39 | 19 | 8 | 11 | 14 | 9 |
| hits | 1 | 1 | 0 | - | - | - |

| | age | height | weight |
|---|---|---|---|
| hits (in 20 trials) | 20 | 4 | 0 |

- Использование биномиального распределения для отбора признаков