Skoltech

Skolkovo Institute of Science and Technology

MASTER'S THESIS

IDENTIFICATION OF MELANOMA-ASSOCIATED T CELL CLONOTYPES IN A MURINE MODEL

Master's Educational Program: Life Sciences

Student: _____ Kseniia Lupyr
signature

Research Advisor: _____ Dmitry Chudakov
signature
PhD, Associate Professor

Co-Advisor: _____ Olga Britanova
signature
PhD, Senior research fellow

Moscow 2022

Skoltech

Skolkovo Institute of Science and Technology

МАГИСТЕРСКАЯ ДИССЕРТАЦИЯ

ОПРЕДЕЛЕНИЕ МЕЛАНОМА-АССОЦИИРОВАННЫХ Т КЛЕТОЧНЫХ КЛОНОТИПОВ В МЫШИНОЙ МОДЕЛИ

Магистерская образовательная программа: Науки о жизни

Студент: _____ Ксения Лупырь
signature

Научный руководитель: _____ Дмитрий Чудаков
signature д.б.н., доцент

Со-руководитель: _____ Ольга Британова
signature к.б.н., старший научный сотрудник

Москва 2022

Identification of melanoma-associated T cell clonotypes in a murine model

Kseniia Lupyr

Submitted to the Skolkovo Institute of Science and Technology on June 2, 2022

ABSTRACT

T cells are crucial players in antitumor immune response. CD4+ T cells shape tumour microenvironment and specifically guide the cytotoxic T cell activity. Remarkable specificity is achieved by a hypervariable T cell receptor or TCR that recognises the peptide in the complex with MHC on the surface of the host cell. T cell target activation is the basis for immunotherapeutic approaches targeting various types of cancer including melanoma. Despite mouse models are commonly used in cancer research, there is still a dearth of knowledge about melanoma-specific TCRs. In this study B16F10 associated murine TCRs were identified by a novel computational tool — TCRgrapher. The method is based on identifying TCRs forming clusters of homologous CDR3 sequences whose generation probability might exceed the baseline expectation. It allows for identifying clonotypes exposed to the challenge without a control sample. TCRgrapher showed a high performance yield as compared to the established methods. Hence, it proved its efficiency to be potentially used in cancer research, in the area of targeted treatment of autoimmunity, for analysis of immune response after vaccination and many other areas.

Keywords: T cell, TCR repertoire, melanoma, B16F10, vaccination, convergent recombination, statistical inference, public immune response

Research Advisor:
Name: Dmitry Chudakov
Degree, title: PhD, Associate Professor

Co-advisor :
Name: Olga Britanova
Degree, title: PhD, Senior research fellow

TABLE OF CONTENTS

1 INTRODUCTION

Immunity is a complex multilevel system protecting various organisms against pathogens and maintaining homeostasis. There are two branches of immunity: innate and adaptive. The first one uses information about pathogens collected in the course of evolution. It is the first line of defence that reacts immediately to any pathogen yet it is not specific. The second one is fine-tuned by the pathogens an organism meets throughout its life, thus it is more specialised. T cells and B cells are the key players of adaptive immunity. However, searching for highly specific types of T or B cells tailored to a particular abnormal condition is still a challenge for contemporary research.

Cancer is one of many abnormal conditions that brings malfunction of immunity to the limelight. Evading immune destruction is an important hallmark of cancer. Research in this area has had significant progress in the last two decades (Hanahan and Weinberg 2011).

T cells largely contribute to the immune response not only against intracellular pathogens such as viruses and intracellular bacteria but also against cancer cells. This ability underlies cancer immunotherapy — an actively developing method for treating cancer.

This work focuses on T cells as they orchestrate the immune response in case of cancerogenesis and are the main cancer cell killers. Billions of T cells have unique receptors that can recognize any pathogen that was met or will be met — T cell receptors or TCRs. The set of all TCRs in an organism is called a TCR repertoire. Each person has his/her TCR repertoire depending on the life history since T cells proliferate and differentiate once they meet a pathogen.

We are living in an era of transition from conventional therapies toward immuno-oncology (Kuryk et al. 2020). The discovery of T cell checkpoint inhibitors recognized by a Nobel Prize in 2018 has already revolutionised the paradigm of oncology treatment for some cancer types and gave rise to many treatments harnessing the natural processes of the immune system.

It is important to recognize the specific TCRs involved in cancer immune response both for basic and applied immune research. In immunotherapy, the knowledge about specific TCRs can be used for developing less toxic checkpoint inhibitor therapies and adoptive T cell therapy. Moreover, the database with TCRs associated with anti-cancer immune response can be used by scientists to check the presence of such TCRs in repertoires in appropriate research.

However, research in this area faces two important challenges. First, the number of TCRs with established specificity is limited. Second, the quality of available data is not always properly controlled (Bagaev et al. 2020). Murine models are one of the most common animal models used in research. So,

the issue of particular importance is to create a database of murine TCRs with known melanoma peptide specificity and to design a method to identify such TCRs.

This study aimed to reveal the key melanoma-associated murine TCRs. To this end, the following objectives were set:

- Develop an R library for the identification of specific TCR from murine repertoires
- Compare developed tool with existing methods
- Make a procedure working with a set of samples
- Identify melanoma-associated TCRs from the dataset

The dataset is TCR repertoires from mice vaccinated with fourteen B16 melanoma peptides. The design allowed identifying TCRs associated with a particular peptide.

We have devised a new approach to process the data efficiently. The existing methods for searching the specific TCRs have several limitations and disadvantages. For instance, experimental methods used in vitro, MHC multimer assays (MHC — major histocompatibility complex), require knowledge about the individual HLA type, peptide processing, and binding to MHC (Davis, Altman, and Newell 2011; Glanville et al. 2017). Modern high-throughput sequencing opens an opportunity to read the individual repertoire and get information about specific TCRs.

Recent evidence suggests that a low number of TCRs is shared among individuals (Shugay et al. 2013). Additionally, similar TCRs can recognize the same MHC-peptide complex (Venturi et al. 2006; Bagaev et al. 2020; Venturi et al. 2006; Bagaev et al. 2020). These two facts drive us to the idea that different donors can have different TCRs associated with the same condition. Furthermore, shared TCRs can be the ones that are highly likely shared within a population and not associated with a condition of interest. Thus, identifying the condition-associated TCRs by mining public TCRs from a large cohort is not effective and can lead to false-positive results.

To solve the mentioned issues a method based on the probability of TCR generation and selection was created (Pogorelyy et al. 2019; Pogorelyy, Minervina, Chudakov, et al. 2018). The next step is taking into account the likelihood of sequences recognizing the same MHC-peptide complex. This idea was implemented in the ALICE (Antigen-specific Lymphocyte Identification by Clustering of Expanded sequences) algorithm (Pogorelyy et al. 2019) which uses TCR generation and selection statistics for the human model. An R library TCRgrapher based on the same idea that operates both

with human and mouse models was developed and was used in this work. Our approach can be used further not only in oncology research but also in diagnostics and intelligent vaccine development.

2 LITERATURE REVIEW

In an attempt to present a complete picture of the scientific background of the subject this chapter begins with overviewing of the general information about T cells and T cell receptors relevant to the research. It proceeds by discussing the ideas lying beneath the identification of specific TCRs from the TCR repertoire. In parallel, research methods and exploratory tools based on these ideas are presented in detail. Finally, a research foundation for our experiments is constructed.

2.1 T cells: specific destruction of what is hidden and control of the immune response

As stated in the introduction, the functions of T cells vary from regulation of the immune response to the lysis of potentially dangerous host cells. These two classes of T cells could be distinguished by co-receptors on their surface: CD4 or CD8 (cluster of differentiation 4 or 8).

CD8+ T cells are called cytotoxic T cells. They recognise target peptides presented in complex with MHC class I. Upon activation, CD8+ T cell makes a pore in a cell it contacts, through which apoptosis-inducing peptides are injected inside. Suchwise, cytotoxic T cells are not only the most specific and effective killers of abnormal host cells, their way of acting is very cautious for neighbouring cells. It is the reason why CD8+ T cells are the most popular object for cancer immunotherapy studies (Raskov et al. 2020).

CD4+ T cells coordinate the work of other immune cells. CD8+ T cells are strongly associated with MHC class I, while CD4+ T cells interact with the peptide-MHC II complex. Naive CD4+ T cells are activated through interacting with peptide-MHC class II on the surface of antigen-presenting cells. Depending on the source of the activation signal, Th (T-helper cell) makes decisions about further differentiation and specifically guides activation of CD8+ T cells (Kasatskaya et al. 2020). On the contrary, regulatory T cells or Tregs (regulatory T cells) inhibit the immune activity of conventional CD4+ T cells when the main threat retreats. CD4+ T cells could serve as viable targets for cancer immunotherapy. Since the immune response in the tumour environment is constrained, Treg suppression or Th activation could lead to positive results in cancer therapy (Wing, Tanaka, and Sakaguchi 2019; Tay, Richardson, and Toh 2021).

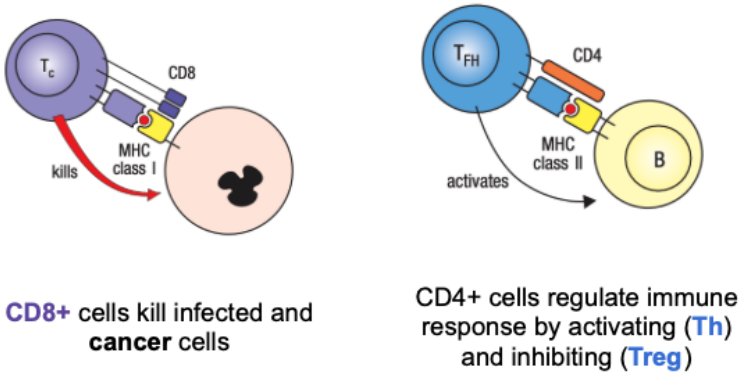


Fig. 1. Two types of T cells (Murphy et al. 2007)

2.2 T cell receptor: a key to T cell specificity

T-cell ability to specifically recognise and bind a wide range of antigens is based on their highly variable antigen-specific TCRs. In the human organism, the naive TCRα/β repertoire diversity reaches 10^8 variants (Qi et al. 2014). Its structure is well established and could be found, for example, in (Garcia et al. 1996). For further TCR repertoire analysis, it is essential to introduce and define the main terms related to its structure, such as TCRβ chain and CDR3 (complementarity-determining region 3).

2.2.1 TCR structure

As one can see in Fig. 2 (A), TCR is formed from two separate germline-encoded polypeptide chains: TCRα and TCRβ. Each chain has a cytoplasmic tail, transmembrane region, and variable (V) and constant (C) regions in its extracellular part. There are three loops in each chain termed complementary-determining regions: CDR1, CDR2 and CDR3 (Fig. 2 (B)).

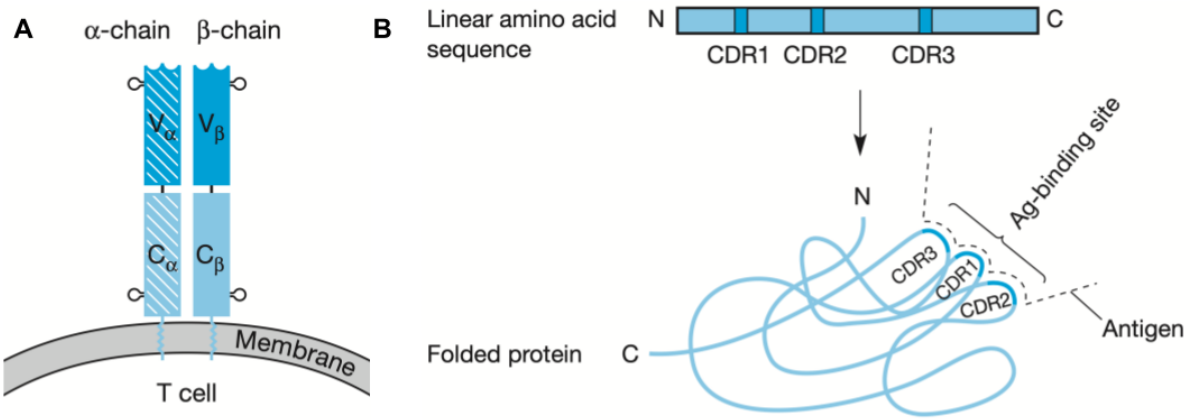


Fig. 2. (A) General structure of TCR (B) Complementary-determining regions in the variable segment (Murphy et al. 2007; Wood 2006)

There are 70 V and 61 J segments in human α-chain locus (Vα and J β). The β-chain locus contains two D segments in addition to 52 Vβ and 13 Jβ segments. During the process called V(D)J-recombination TCRα and TCRβ chains are formed by joining one V, one D (in the case of β-chain) and one J segments, which are quasi-stochastically chosen from the sets of segments. Further, random trimming and insertions of nucleotides add diversity to junction regions.

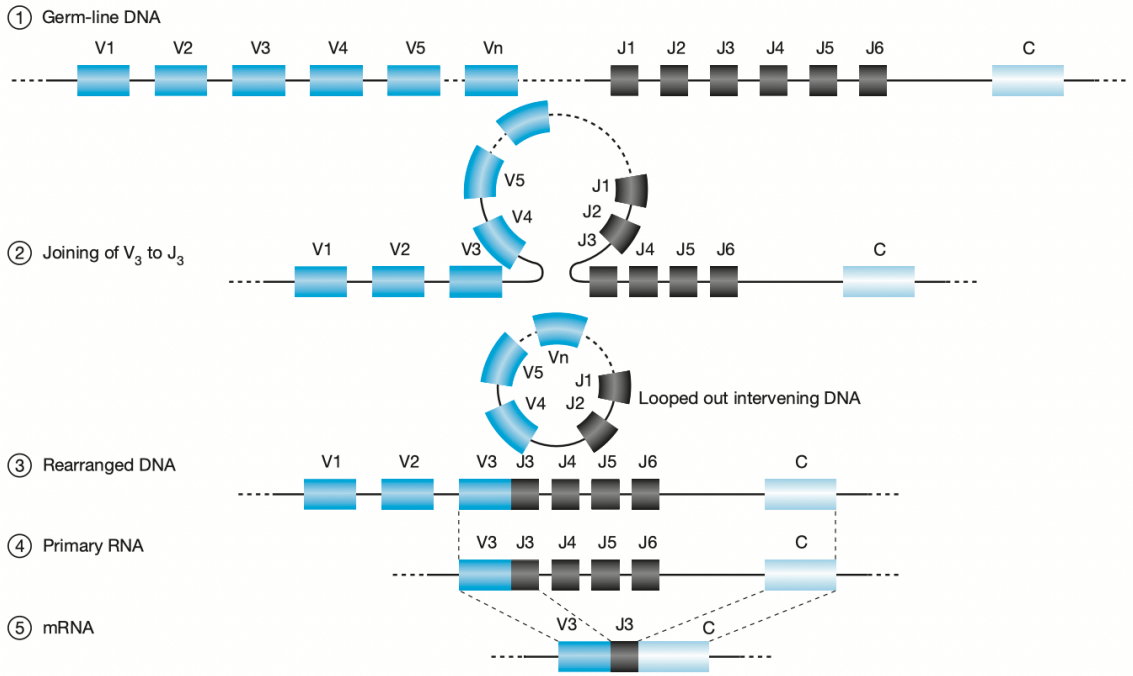


Fig. 3. V(D)J recombination (Wood 2006)

CDR1 and CDR2 depend on V-gene usage. They are mostly responsible for establishing contact with MHC in the peptide-MHC complex (Sim et al. 1996). While the CDR3 region largely contributes to epitope recognition in the MHC context (Egorov et al. 2018). CDR3 includes VD and DJ junction regions. With respect that these junctions have trimmed and added nucleotides, CDR3 mostly provides the diversity of T cell receptors.

This work focuses on the TCRβ CDR3 repertoires and uses the IMGT (The International Immunogenetics Information System) CDR3 definition. According to IMGT the CDR3 is delimited by (but does not include) the highly conservative anchor positions 2nd-CYS 104 and J-PHE or J-TRP 118 ("IMGT Unique Numbering for Immunoglobulin and T Cell Receptor Variable Domains and Ig Superfamily V-like Domains" 2003).

2.2.2 Rearrangements of the TCRα and TCRβ loci

Rearrangements of TCRα and TCRβ loci go sequentially with the TCRβ locus going first. After it, the TCRβ chain is expressed as a part of pre-TCR. If the functional TCRβ chain was not synthesised, which statistically could happen, apoptosis is activated inside the cell. However, the cell could avoid death if a successful rearrangement is performed on the second chromosome (Fig. 4).

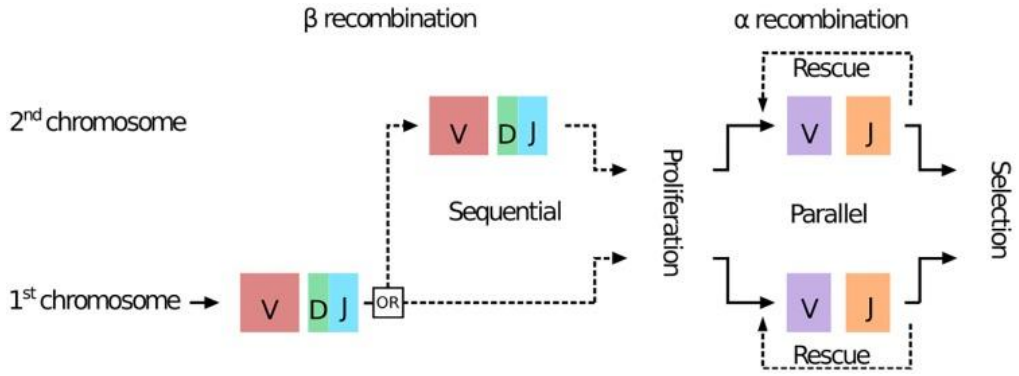


Fig. 4. Rearrangement of the TCRα and TCRβ on homologous chromosomes (Dupic et al. 2019)

Pre-TCR expression induces signals which stop β-chain rearrangement. The same signals lead to cell proliferation and expression of co-receptors CD4 and CD8. This stage is called β-selection. Hence, every cell with a successfully rearranged β-chain proliferates and can give rise to a bunch of thymocytes.

When proliferation ends, cells start rearrangement of TCRα locus. This rearrangement can proceed on both chromosomes. Therefore, T cells could express two types of α-chains. It is estimated that about 35% of T cells express two kinds of receptors (Dupic et al. 2019). At this stage, TCR is expressed on the surface of the cell and could contact peptide-MHC (pMHC) inside the thymus. TCRα rearrangement continues until T cells get signals from pMHC or until cell death.

This work explores the TCRβ repertoire. This approach is widely used for several reasons. First, although segments of the TCRα locus have more combination variants, in the real repertoire, α-chain is less variable than β-chain. This is confirmed by stochastic modelling ("How Many Different Clonotypes Do Immune Repertoires Contain?" 2019). Second, the β-chain is largely responsible for contact with the antigen, while the α-strand has more contact with the MHC (Marrack et al. 2008).

2.3 Selection in a thymus

Rearrangements of TCRα and TCRβ loci happen in the thymus — the primary lymphoid organ specialised in T cell maturation. There are two main parts in the thymus: cortex and medulla. These

parts are associated with positive and negative selection respectively. In the first part of the development T cells with synthesised TCR on the surface are tested for their ability to recognise pMHC in the cortex. T cells that successfully passed this stage migrate to the medulla and are checked for their ability to bind self-peptides. Potentially auto-reactive T cells are eliminated. Let's consider the events going on in the thymus in detail. The main intercellular interactions are presented in Fig. 5.

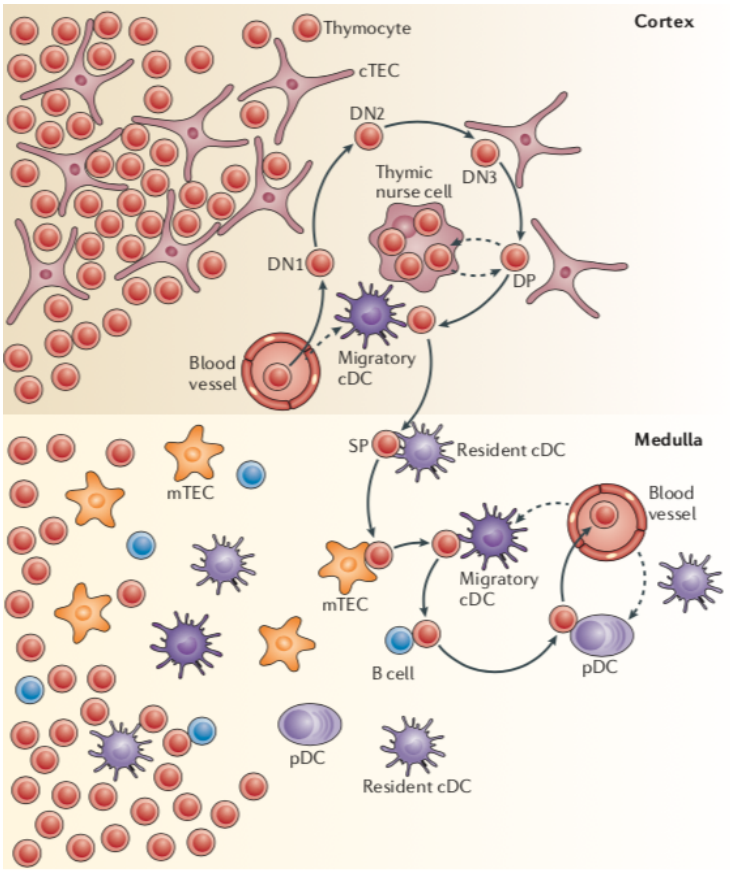


Fig. 5. General cellular interactions in a thymus (Klein et al. 2014)

2.3.1 Positive selection

Firstly, T cell precursors enter the thymus cortex from the bone marrow through blood. At this moment, cells do not have CD4 and CD8 and are called double-negative thymocytes (DN). DN thymocytes rearrange the TCRβ locus and proliferate. After proliferation thymocytes enter a predominant double-positive stage where they express CD4 and CD8, rearrange TCRα locus and finally express TCRαβ on the surface (Sebzda et al. 1999). TCRαβ contacts pMHC on the surface of antigen-presenting cells (APCs). Due to the stochastic formation of TCR, only a low number of them can interact with pMHC molecules. After the contact DP thymocytes get signals essential for their

differentiation into single-positive (SP) thymocytes. If they do not get such signals for 3-4 days in the cortex the apoptosis will be activated. This fate is called death by neglect.

2.3.2 Negative selection

The positive selection stage is followed by migration from the cortex to the medulla. Migrated SP thymocytes scan APCs, and thymocytes with high affinity to pMHC are eliminated from the TCR repertoire. In other words, if a cell has TCR with a high affinity to self-peptide, at this stage it gets signals for apoptosis (Klein et al. 2014). However, some thymocytes with self-reactive TCRs could differentiate into Tregs. These cells control auto-immune response at the periphery by inhibiting other immune cells. Thus, negative selection ensures the safety of the organism.

Summing up, through both positive and negative selection following each other, only a small amount of thymocytes are released into the periphery. Most cells in this fraction have functional and not auto-reactive TCRs.

2.4 TCR repertoire: revealing nature of the immune response

Recent studies using next-generation sequencing of TCR repertoire gave insights into the composition of immune repertoire and patterns of the immune response. In this chapter, I try to give an overview of current studies about TCRs repertoire generation, selection and convergence.

2.4.1 Generation models

Recent studies using TCR repertoire sequencing provide evidence that V(D)J recombination described above is a complex quasi-stochastic process. Biases in V, D and J gene usage are described in various articles (Zvyagin et al. 2014; Qi et al. 2016; Rubelt et al. 2016; Pogorelyy, Minervina, Touzel, et al. 2018). In humans for both D genes and around half of V and J genes significant association with genotype was found (Russell et al. 2022). The same study reported that V-, D- and J-gene trimming, also, depends on genotype variations. In mice, 80% of biases in TRBJβ usage can be explained by the physical model of chromatin conformation (Ndifon et al. 2012). Therefore, TCRs containing frequently used genes will have a greater generation probability to be in the repertoire than TCRs with rare variants. Statistical composition of immune repertoire can be used in comparative studies and give insights into the individual immune response. Rules of V(D)J recombination can be identified by analysis of non-functional sequences that have not gone through selection. To infer V(D)J recombination model researchers use two approaches: biophysical estimation and machine learning techniques. Let us consider them successively.

Using **biophysical estimation of V(D)J recombination** researchers try to understand the nature of the process in detail and express it in a mathematical model**.** Firstly, different events can lead to the formation of the same TCR. For example, CDR3 read from Fig. 6 (grey box) can be synthesised using different TRBV and TRBD genes with corresponding insertions and deletions. In this case, the probability to be generated for this nucleotide sequence is a combination of probabilities of all possible ways to produce this sequence.

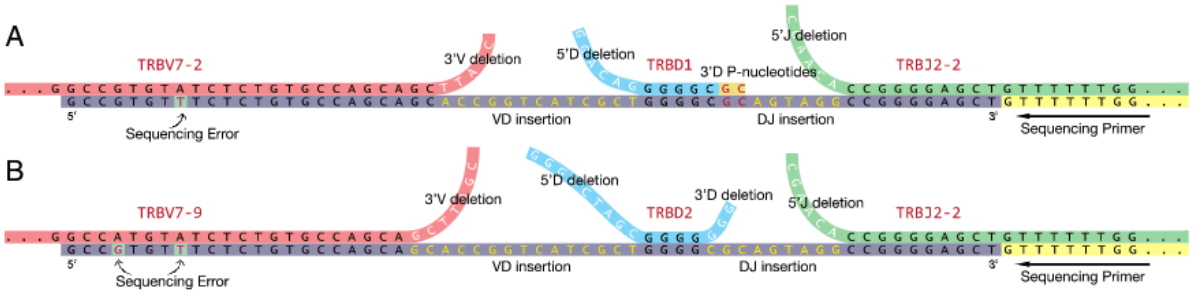


Fig. 6. Example of different scenarios that lead to the same CDR3 nucleotide sequence from (Murugan et al. 2012)

Secondly, convergent TCR*s* are the TCRs that have the same amino acid sequences but different nucleotide sequences. If we are interested in amino acid TCR probability, we should consider all possible ways to produce every nucleotide variant presented in the repertoire.

Further, I will use the term 'scenario' for a unique set of such events following (Sethna et al. 2020). The scenario includes TRBV, TRBD and TRBJ gene choice (V, D, J), deletions in a 3' end of the V gene and in a 5' end of the J gene ($delV$, $delJ$), deletions in both ends of D segment ($del5'D$, $del3'D$), number of palindromic nucleotides for every gene ($palV$, $palJ$, $pal5'D$, $pal3'D$) and inserted sequences between VD and DJ segment ($x_1, \dots, x_{insVD}, y_1, \dots, y_{insDJ}$), where the probability of every nucleotide to be added depends on the previously added nucleotide. Therefore, there are thirteen variables that could affect the probability of the scenario. Biophysical estimation studies aim to find the joint distribution with a minimum set of variables which will match the real one.

(Murugan et al. 2012) found that the following joint distribution catches all significant factors presented in the data:

$$P_{scenario} = P(V)\,P(D,J)\,P(delV \mid V)\,P(delJ \mid J)\,P(del5'D, del3'D \mid D)$$
$$\times\, P(insVD) \prod_{i=1}^{insVD} P_{VD}(x_i \mid x_{i-1}) P(insDJ) \prod_{i=1}^{insDJ} P_{DJ}(y_i \mid y_{i-1})$$

Some elements of the equation are normalised joint or conditional distributions of the variables. From the formula, we can see that choice of V-gene is considered an independent event, while D- and J-gene usage are correlated events. The number of deleted nucleotides depends on the gene if it is deleted. Although, the number of inserted nucleotides depends on the junction if it is added. Furthermore, there is nucleotide bias in the insertions. The probability will be different according to the type of added nucleotide and nucleotide before it. An example of the distribution of the parameters is presented in Fig. 7.

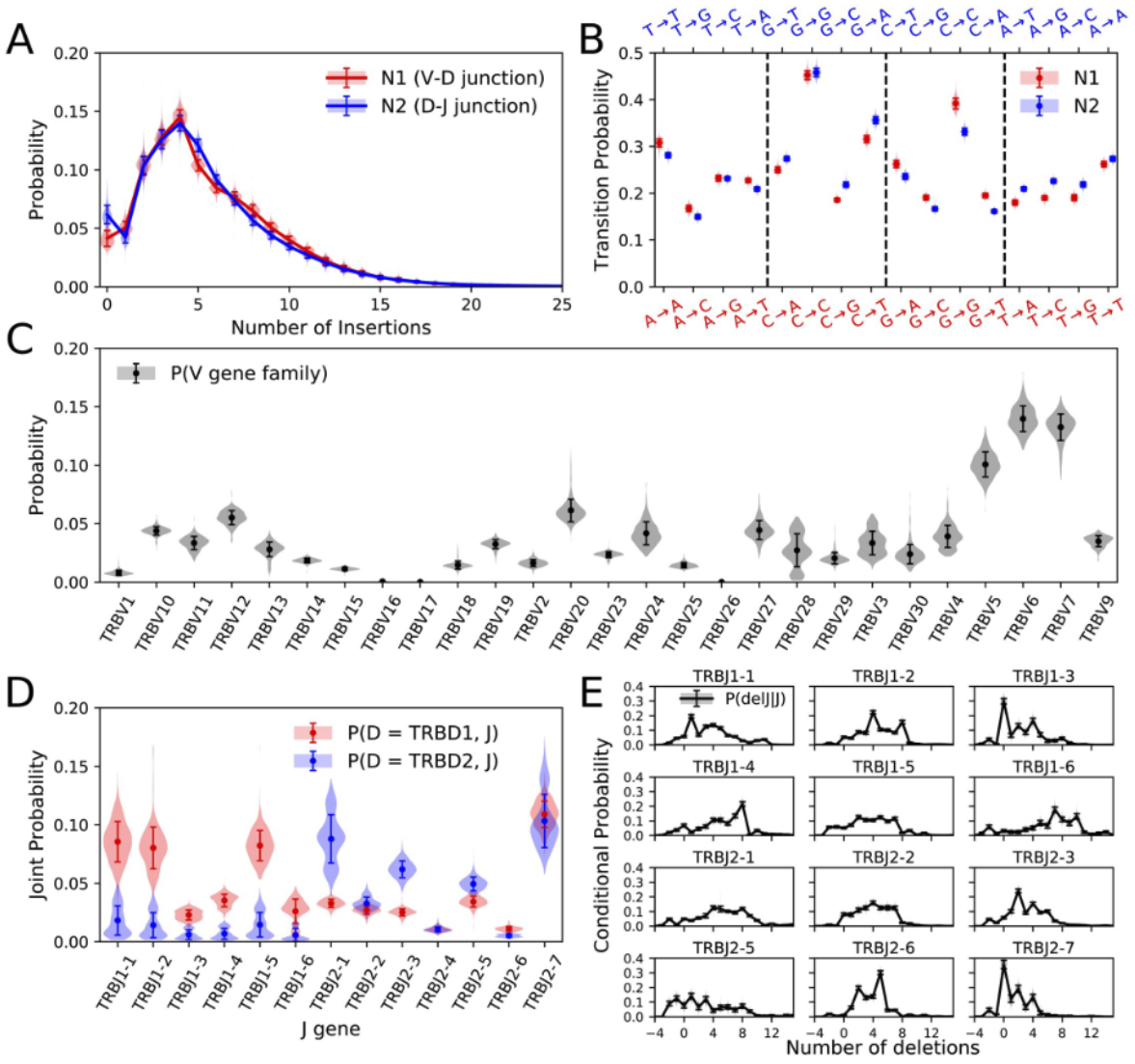


Fig. 7. Distributions of parameters used in recombination scenario probability calculation. The image is taken from (Sethna et al. 2020). There are 651 individuals in the dataset. Median and standard deviations are shown by error bars in all plots. (A) Insertion length distribution. (B) Markov transition probabilities for the inserted nucleotide identities. (C) TRBV gene usage. (D) Joint distribution of TRBD and TRBJ usage (E) Deletions length distributions depending on TRBJ gene

(Marcou, Mora, and Walczak, n.d.) used the model described above to develop IGoR (Inference and Generation of Repertoires) — tool for characterisation of the receptor generation statistics. IGoR finds all possible scenarios for every sequence in the repertoire and calculates probability weights according to the likelihood of the scenarios. As an output, IGoR gives a statistical model of recombination that can be used to generate synthetic sequences. IGoR's pipeline is presented in Fig. 8.

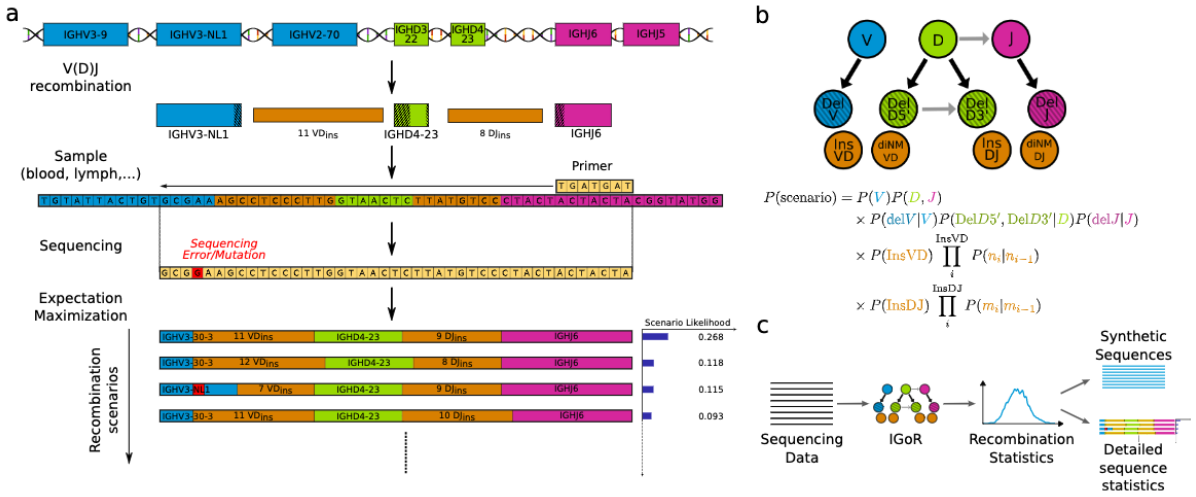


Fig. 8. IGoR's pipeline (A) IGoR makes recombination scenarios for every sequence in the repertoire, calculates the likelihood of each scenario and according to calculations gives weights to the scenarios. (B) Illustration for human TRB likelihood computation that uses Bayesian network of dependencies between the recombination variables (C) IGoR's pipeline consists of three steps. The first step is learning recombination statistics from non-productive sequences. The second step is producing recombination scenarios for every sequence. The third step is a generation of synthetic sequences with learned recombination statistics. Image from (Marcou, Mora, and Walczak, n.d.)

How to calculate the generation probability of every sequence in the repertoire in the amino acid level having inferred from the non-functional data model of recombination? In (Pogorelyy, Minervina, Chudakov, et al. 2018; Pogorelyy et al. 2019) generation probability is estimated using Monte Carlo simulation. First, a large amount of sequences is generated *in silico*. In (Pogorelyy et al. 2019) 100 million TCRs with fixed VJ combinations were generated. Second, generated nucleotide sequences are translated into amino acid sequences. Finally, the frequency of every sequence is calculated by counting. Computations in this case are time and memory consuming. Monte Carlo simulation is sensitive to Poisson sampling noise. So, using this method, the user must be sure that a number of events are enough for the accurate calculation of generation probability.

Another approach is to calculate the occurrence of every sequence in a large control dataset and treat it as a probability to be in the repertoire. The quality of the dataset will be a limitation of this method.

For the fast and accurate computation of generation probability (Sethna et al. 2019) developed a tool named OLGA (Optimized Likelihood estimate of immunoGlobulin Amino-acid sequences). Using dynamic programming computations allows OLGA to be much faster than Monte Carlo simulation which is in agreement with OLGA results. By default, OLGA applies the generation probability model from (Marcou, Mora, and Walczak, n.d.), but as an option, it can use any model in IGoR output format.

A different approach to estimating generation probability is to use **machine learning techniques**. (Davidsen et al. 2019) proposed fitting variational autoencoder (VAE) models parameterized by deep neural networks to predict TCR frequency in the repertoire and other parameters. VAE models are widely used with large data with non-linear distribution and interactions between covariates. An illustration of how VAE works is presented in Fig. 9. Overview of key principles is in Fig. 9 description.

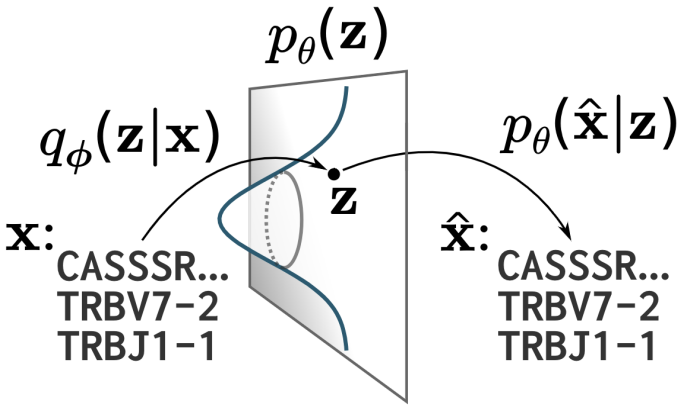


Fig. 9. Principle of VAE operation. Image from (Davidsen et al. 2019). VAE consists from two main parts: encoder $q_\phi(z|x)$ and decoder $p_\theta(\hat{x}|z)$. Encoder embedded input object x (TCR sequence) into 20-dimensional latent space $p_\theta(z)$. Decoder deconverts $\hat{x}$ from latent space. VAE encode and decode objects with high fidelity ($x \approx \hat{x}$). After VAE is trained, parameters of latent space can be used to generate random z' by Monte Carlo importance sampling. Then, z' is decoded producing new synthetic sequences. Generation probability of sequence of interest is calculated by counting its frequency in the synthetic repertoire

(Davidsen et al. 2019) showed that VAE models can predict TCR cohort frequency. The result is slightly different from OLGA output. Authors assumed that VAE captures more characteristics of TCRs sequences than OLGA in their implementation. However, because of parameterization by neural networks the result is not directly interpretable. Authors projected 20-dimensional latent space on the surface with PCA (principal component analysis) and found that projection structured according to

TRBV, TRBJ and CDR3 length. Other arguments remain unrevealed. In contrast to (Davidsen et al. 2019), (Isacchini et al. 2020) reported that biophysical methods outperformed VAE model, but noted that both approaches worked quite well.

2.4.2 Selection models

T cells with synthesised TCR go through positive and negative selection. Resulting naive TCR repertoire differs from the TCR repertoire in the cortex. Since in real datasets T cells were exposed to the selection we should have a way to take it into account. Previous methods based on the VAE model can be used to predict TCR frequency in repertoire after selection. In this section we will discuss three methods based on different techniques.

First method was proposed in (Elhanati et al. 2014). Authors introduce term 'selection factor' for every TCR sequence and define it as $Q = P_{post} / P_{pre}$, where Q — selection factor, P_{post} — TCR probability to be in the repertoire after selection, P_{pre} — TCR probability to be in the repertoire before selection. They found that the following model for Q mostly captures all important features of selection and gives good results:

$$Q = P_{post}(\bar{\tau}, V, J) / P_{pre}(\bar{\tau}, V, J) = \frac{1}{Z} q_L q_{VJ} \prod_{i=1}^{L} q_{i;L}(a_i),$$

Where $\bar{\tau}$ — CDR3 sequence, V, J — V- and J- genes, $(a_1, ..., a_L)$ is CDR3 amino acid sequence with length L, q_L represents the power of selection depending on the length of CDR3, q_{VJ} — depending on VJ combination, $q_{i;L}(a_i)$ express composition of selection according to amino acid and its position.

Schematic representation of parameters' selection procedure is shown in Fig. 10. (B). Parameters are iteratively modified using expectation maximisation algorithm until P_{pre} will fit P_{post} distribution. As a result, every sequence has its own selection factor that shows whether this sequence is less represented in the final naive repertoire or not. This model is used in SONIA. The tool takes TCRs sequences as an input and calculates selection factor for every sequence.

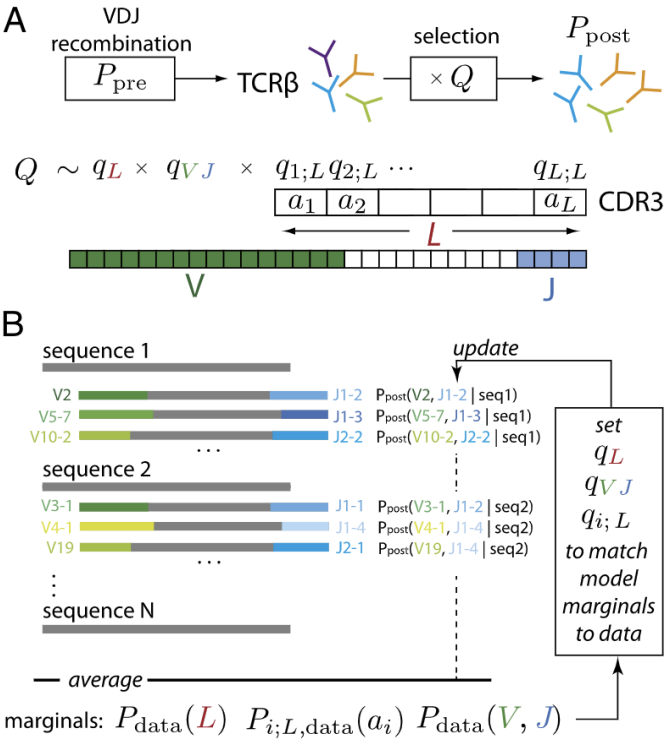


Fig. 10. Graphical representation of the method proposed by (Elhanati et al. 2014).

Interestingly, the study showed correlation between generation probability and probability to pass through selection. Authors assumed that it could tell us about the evolutionary influence of natural selection on generation mechanics.

Additionally, this fact leads us to the idea to use universal selection factor for all sequences. In (Elhanati et al. 2018) it was estimated from a convergent curve. Number of unique amino acid CDR3 sequences depends on the number of unique nucleotide sequences. This dependence is explained by selection factor $q = 1/Q$. By this method the following selection factors were obtained: $q = 0.16 \pm 0.03$ for mice and $q = 0.037 \pm 0.002$ for humans.

The last method for selection estimation in this section works on machine learning techniques. (Isacchini et al. 2021) presented soNNia — a tool for inferring selection models using deep neural networks (DNNs). The working principle is presented in Fig. 11.

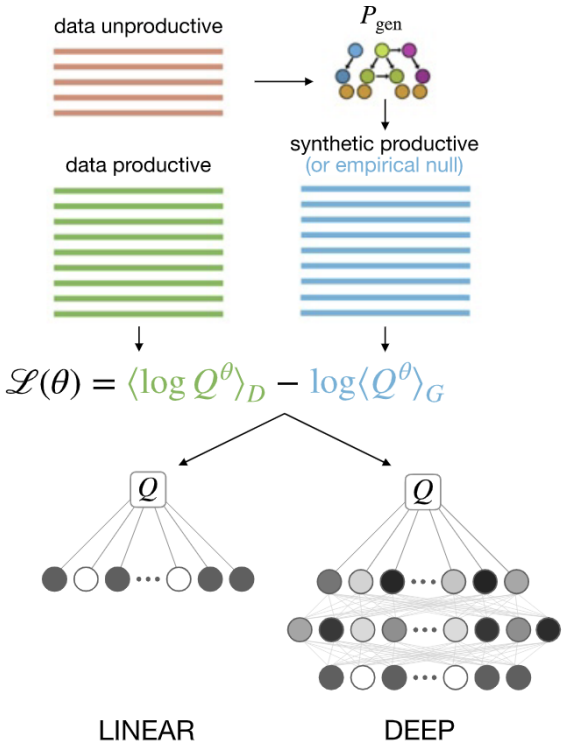


Fig. 11. A schematic representation of the soNNia working principle. As an input, soNNia takes IGoR output and targeted functional sequences. To infer the selection model soNNia maximised the mean log-likelihood of the data. The equation for the likelihood is presented in the plot. soNNia takes three CRD3 features as an input: TRBV and TRBJ usage, CDR3 length and CDR3 amino acid composition in the form of a binary special matrix. In the first place, features go through three independent neural networks. In the dense layer, outputs are combined and transformed. This way decreases the number of parameters in DNN and makes it possible to compare the contributions of three features. The output of the DNN is logQ. Weights of the DNN are iteratively modified to maximise the mean log-likelihood. Image from (Isacchini et al. 2021).

DNN-based soNNia was compared with the original SONIA and showed significant improvement in the model. The authors emphasised the role of nonlinear factors in the selection model. They, also, compared distributions between CD4+, CD8+, generated and functional T cell subsets. The difference between generated and functional subsets was conspicuous while the difference between CD4+ and CD8+ subsets was minor. Since negative selection follows positive selection and is accompanied by CD4+ and CD8+ differentiation, researchers assumed that their model mostly captured positive selection features. A limitation of the approach is its dependency on a large amount of data.

2.4.3 Convergent immune response

Over the last 30 years, restricted diversity in TCR repertoire was described in various cases: infections, cancer, autoimmunity, and allergy (Miles, Douek, and Price 2011). Now we have sufficient evidence that the immune system responds to the challenge by numerous homologous TCRs (Dash et

al. 2017). As an illustration, I choose a plot from (Pogorelyy, Minervina, Touzel, et al. 2018). In this work immune response to yellow fever vaccination was investigated. In Fig. 12 one can see a graph describing one thousand most abundant clonotypes from donor S1 on day 15 after vaccination. Every node corresponds to one TCRβ amino acid sequence. Nodes connect clonotypes that differ by one or two amino acids. Clonotypes coloured blue were present in the repertoire before vaccination with comparable frequency. Clonotypes coloured yellow are expanded ones. As it might be seen, expanded clonotypes form dense clusters of similar TCRs. This feature of T cell immune response can be used to identify specific TCRs. Specific ways to do so will be discussed below.

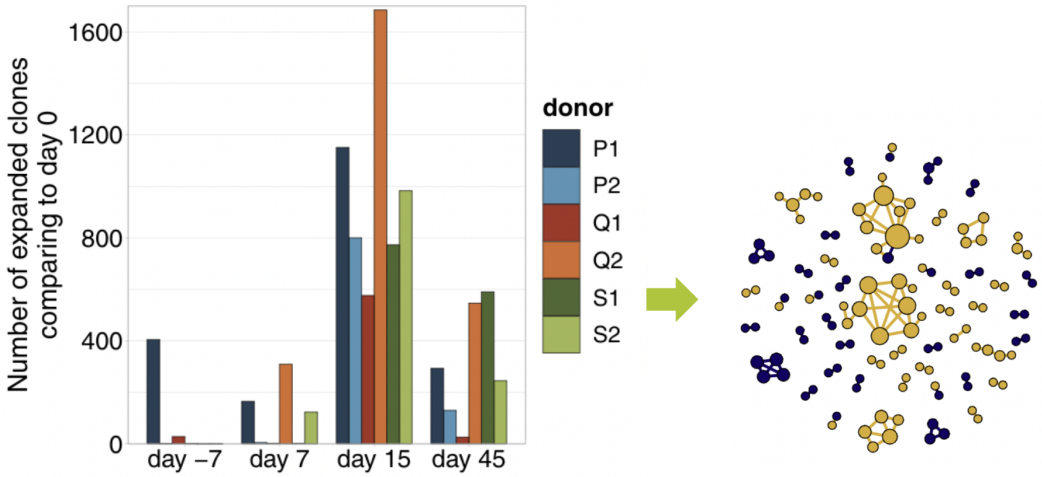


Fig. 12. Yellow fever vaccination. Images from (Pogorelyy, Minervina, Touzel, et al. 2018).

2.5 TCR repertoire: a valuable source for TCR specificity identification

Numerous studies use next-generation sequencing (NGS) to analyse TCR repertoires. Using NGS, great progress has been made in various research areas, for example, in the treatment of cancer, autoimmune diseases, the study of the immune response in organ transplantation, vaccination, and others (Davis, Tato, and Furman 2017). However, the current ability to extract the clinically valuable data from TCR repertoires is limited due to the restricted number of TCRs with known specificity (Bagaev et al. 2020). Meanwhile, this knowledge is in high demand for diagnostics, cancer immunotherapy, target treatment of autoimmune disease and vaccine development. Becoming a more common tool, TCR repertoire sequencing could be a valuable source for target immune specificity experiments. It could also be used to profile disease states and analyse the immune response to a vaccine or immunotherapy.

The current section describes methods for identifying epitope-specific TCRs only from TCR repertoire sequencing *without knowledge about epitope and donor's MHC*. They can be tentatively divided into three groups: methods using similarity to previously established specific TCRs, methods applying TCRs generation and selection models and methods based on searching homologous TCRs clusters.

2.5.1 How to identify epitope-specific TCR? Established specificity

TCRmatch (Gielis et al. 2019) and TCRex (Chronister et al. 2021) are the most popular approaches based on the robust idea — analyse TCR-pMHC binding data to identify common patterns.

(Gielis et al. 2019) developed an open web tool 'TCRmatch' that takes TCRβ CDR3 as an input, matches the input sequence with the Immune Epitope Database and finds TCR from the database with the highest probability to bind the same epitope as an input. A flowchart describing the TCRmarch pipeline is presented in Fig. 13 (A). (Gielis et al. 2019) analysed metrics of sequence similarity to identify their ability to predict if two TCRs have the same epitope specificity. As a result, developed metrics 'TCRmatch' uses "comprehensive comparison of all possible k-mers using BLOSUM62 observed frequency matrix" showed the best performance compared with TCRdist (Dash et al. 2017), alignment score and Levenshtein distance. TCRdist will be described further. One of the results from the article is shown in Fig. 13 (B). TCRmatch allows users to annotate epitope specificity only from the TCRβ repertoire. However, the performance is limited by the size of available datasets of TCR-pMHC crystal structures.

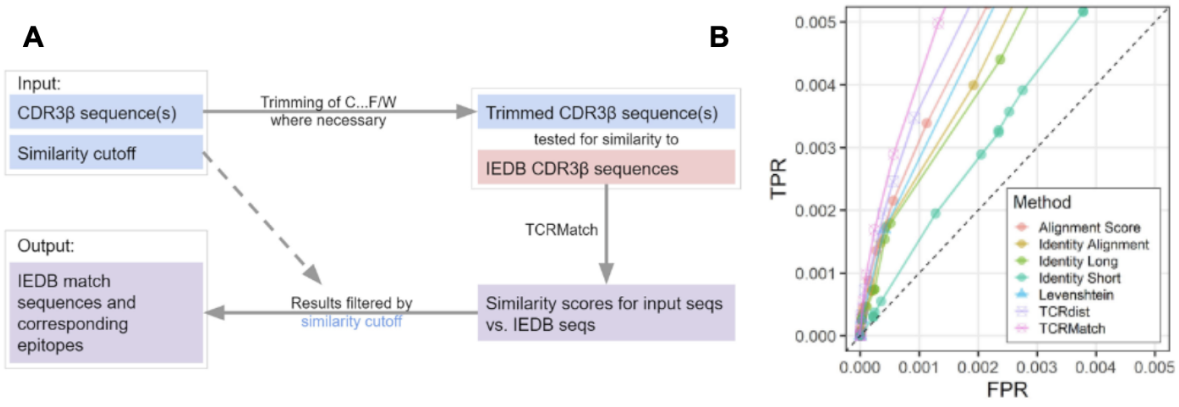


Fig. 13. TCRmatch **(A)** Pipeline of TCRmatch **(B)** Comparison of similarity metrics by true-positive rate (TPR) and false-positive rate (FPR)

(Chronister et al. 2021) presented TCRex — web tool for TCR-epitope binding prediction. It uses a random forest-based model trained on physicochemical properties to identify epitope-binding TCRs

in a repertoire dataset. Moreover, epitope specificity enrichment analysis allows detecting epitopes with a significantly higher number of specific TCRs in the repertoire than expected in a representative control TCR repertoire. Therefore, clonotypes that are widely shared across individuals are eliminated from the analysis. As TCRmatch described above, the main limitation of TCRex is a restricted number of studied epitopes. Taking into consideration the growing number of epitopes with known specificity (Bagaev et al. 2020) both methods can be very prominent in the future.

2.5.2 How to identify epitope-specific TCR? Frequency comparison

If the experiment design involves control groups or samples from time points before antigen exposure, tools for differential expression analysis can be used, for example, EdgeR (Robinson, McCarthy, and Smyth 2010) and DeSeq2 (Love, Huber, and Anders 2014). In this context, every unique TCR amino acid sequence should be treated as a ‘gene’. This approach allows us to estimate frequency variability in control and immunised groups and reveal expanded clonotypes. It was applied in the already mentioned study about yellow fever vaccination (Pogorelyy, Minervina, Touzel, et al. 2018).

Obviously, ReqSeq data alter from RNA-seq data and require its own statistical approaches to estimate frequency variability in the group and the sample. In the same study, (Pogorelyy, Minervina, Touzel, et al. 2018) proposed a Bayesian statistical framework for searching strongly and significantly expanded clonotypes. RNA (ribonucleic acid) molecule count distribution is usually estimated as negative binomial distribution by standard tools for RNA-seq analysis. The framework presented in the study used two-step estimation with negative binomial distributed cell count distribution and Poisson distributed RNA molecule count distribution. It gave better results, especially on low-number pair-count statistics. However, both approaches provided significant discrepancies in high-number pair count statistics with a better result for the two-step framework. The comparison is in Fig. 14.

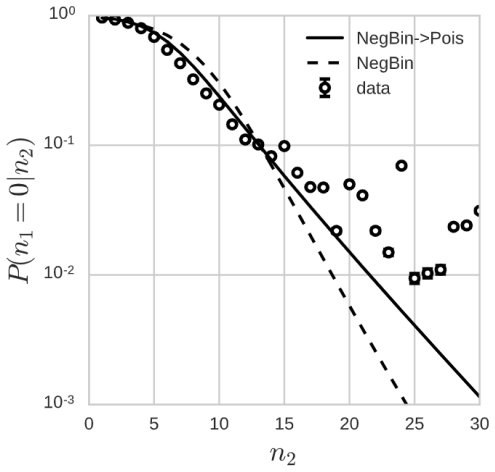


Fig. 14. Conditional count distributions of learned one and two-step models for donor S1 day 0 timepoint biological replicates. The plot from supplementary (Pogorelyy, Minervina, Touzel, et al. 2018).

Nevertheless, cell count and RNA molecule distributions are not the only factors which make difference between TCR repertoire sequencing and RNA sequencing. Observer clone frequency is exposed to sampling issues and varies from the real ones in the whole organism. Therefore, clonotype frequency is a noisy function. Correction for the noise was a work subject of (Puelma Touzel, Walczak, and Mora 2020). The proposed method consisted of three steps. Firstly, the characteristics of the frequency distribution and noise model were inferred from several samples obtained at the same time. Secondly, with data obtained on the first step subsequent time points were analysed to get the parameters of the evolution operator. After these two steps were made, the authors had data for expanding clonotype identification with a correction to stochastic and biological variations in the repertoire. The authors claim that the inferred model was conservative for donors and could be used in experiments without a control sample. However, it was sensitive to different protocols.

2.5.3 How to identify epitope-specific TCR? Probability vs sharing

As previously stated, every TCR can be assigned with its probability to be in the repertoire. 'Public' TCRs that are widely shared across donors have high generation and selection probabilities. So, by mining TCRs from a large cohort, we will highly likely get not specific 'public' TCRs as a result. To avoid it, a statistical approach should be used. In this chapter methods used for both sharing analysis and statistical composition of the repertoires will be described.

(Pogorelyy, Minervina, Chudakov, et al. 2018) provided a statistical framework for identification of specific TCRs from a small cohort of patients without a control sample. Pipeline is graphically presented in Fig. 15. For every shared sequence in the cohort P_{gen} and P_{data} are estimated

and compared. In the paper P_{gen} was calculated using a recombination model proposed in (Murugan et al. 2012). P_{data} was estimated using sharing patterns. Comparison of these two values allows us to test the hypothesis that sharing pattern is explained only by generation probability of the TCRs. If the null hypothesis could be refused, TCR was identified as condition-specific.

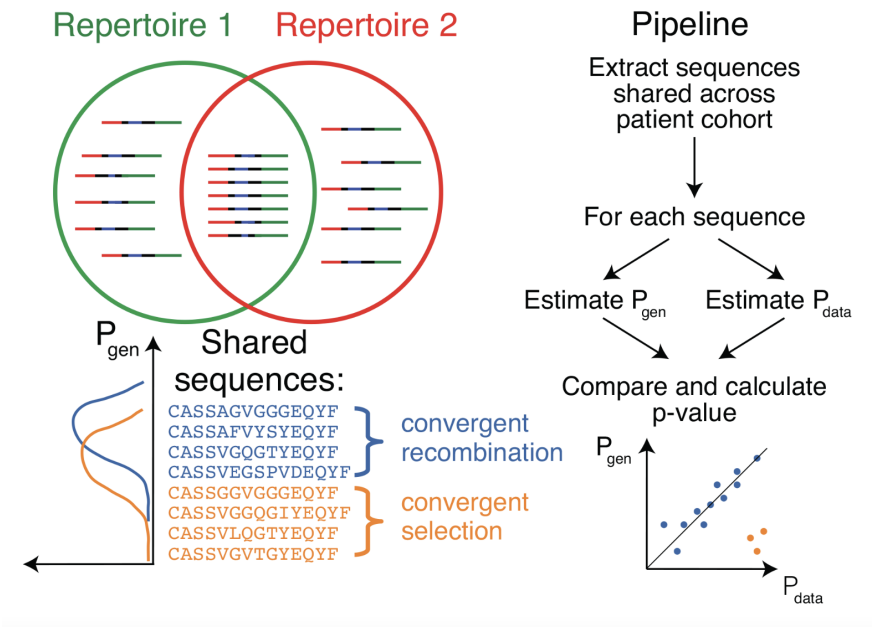


Fig. 15. Graphical representation of the method proposed by (Pogorelyy, Minervina, Chudakov, et al. 2018).

The idea of comparison between TCR probability to be in the repertoire and sharing between individuals can be realised with different methods for P_{gen} estimation described in chapter 'Generation models'. Moreover, several tools were developed for 'publicity' estimation, for example, 'PUBLIC' classifier (Elhanati et al. 2018) and machine learning-based method proposed by (Greiff et al. 2017). Summarising, searching for 'over-shared' clonotypes is a powerful approach for identification of common immune response patterns. However, using it leads to the loss of information about specific private responses.

2.5.4 How to identify epitope-specific TCR? Clusters of homologous TCRs

A commonly accepted paradigm is that TCRs recognising the identical epitope often have similar sequences (Meysman et al. 2019). The question naturally arises: how to measure similarity?

The first answer is to introduce a function that defines the distance between two TCR sequences. In (Dash et al. 2017) the distance between TCRs was computed using Hamming distance with gap penalty that captures the difference in length with a correction to similarity by BLOSUM62.

CDR1, CDR2 and CDR3 have different weights with priority to CDR3. Using this metric (Dash et al. 2017) demonstrated that the TCR repertoire has epitope-specific clusters of homologous TCRs with conservative core features. Identification of these features is a key to revealing TCR epitope recognition patterns. Distance calculation was followed by TCRs clusterization by distance and TCR frequency comparison with background TCR repertoire. The bioinformatics approach for the identification of specific TCR was validated by several functional assays and showed good performance.

The second tool is GLIPH2 (Huang et al. 2020) which searches for similarity using global and local CDR3 alignment. In clusters based on local alignment, CDR3 sequences differ by not more than three amino acids. GLIPH2 identifies specific clusters by comparing the numbers of clones in the cluster between a sample and control naive dataset by Fisher exact test. Analysing response to Mycobacterium tuberculosis (Huang et al. 2020) indicated that only 36% of enriched specific TCRs were clustered.

The third algorithm — TCRNET (Pogorelyy and Shugay 2019) implemented as a part of VDJtools (Shugay et al. 2015). Firstly, it counts the number of neighbours for every TCR, wherein a neighbour is a TCR with fewer mismatches than set by the user. Then it makes a comparison to the control defined by the user according to Poisson or binomial distribution.

Analogous method implemented in the tcrdist3 package (Mayer-Blackwell et al. 2021). Comparison is carried through Fisher's exact test and TCRdist metric is used to define neighbours. In the study, the authors provide recommendations for the choice of the radius — the threshold for the distance. Sequences with a distance less than the radius are considered neighbours.

2.5.5 How to identify epitope-specific TCR? ALICE

Together, the studies and tools presented above form a wide landscape of methods for the identification of specific TCRs. Considering their requirements and limitations one can choose the most suitable method for a particular experiment. Methods that search for enrichment to previously established TCRs in the future can be used, for example, in clinical patient screening. Expanded clonotypes identification with accurate experiments gives clear results with a very low level of false-positive results. On the other hand, the number of identified clonotypes is small compared to cluster searching methods. Sharing analysis gives fine results if we are interested in common response patterns. However, the essential size of the cohort depends on the antigen and can be quite large in the case of low antigen immunogenicity. Moreover, a lot of responded clonotypes are unique to individuals

and sharing analysis can not be used if one is interested in an individual immune response. A cluster searching approach is a powerful tool. But public clonotypes form large clusters too. Comparison to the control requires a large reference dataset and is sensitive to its quality.

One niche remains unoccupied which is an analysis of individual samples without control that combines a cluster searching approach and the power of TCR generation and selection models. For this purpose, ALICE (Pogorelyy et al. 2019) was developed. It is an object of special interest because the method used in the thesis is based on the ALICE statistical model. In a few words, ALICE identifies TCRβ CDR3 clonotypes that have statistically significantly more neighbours than expected. Neighbour is TCRβ CDR3 sequence with one amino acid mismatch or identical amino acid clonotype with different nucleotide sequence. Expected number of neighbours is estimated using generation probabilities of all possible neighbours. Let us consider the ALICE statistical model in detail.

ALICE statistical model. The null hypothesis is the following: the number of neighbours (d) of the given sequence (σ) is the same as in random repertoire, where the neighbour is a sequence with the same VJ combination as in the given sequence and having one or null amino acid mismatches. Probability to have the same number of neighbours as the given sequence or higher is Poisson distributed:

$$P(d \mid \sigma) = e^{-\lambda} \frac{\lambda^d}{d!}, (1)$$

$$\lambda = n \sum_{\sigma'}^{\sigma} Q\, P_{gen}(\sigma'), (2)$$

where d — number of neighbours, σ — the given sequence, σ' — sequence similar to the given with one mismatch, n — number of unique sequences with a given VJ combination; Q — selection factor, the same as 1/q, where q is the fraction of sequences pass the selection; P_{gen} — generation probability.

Including abundance information. Basic ALICE pipeline uses only information about the number of neighbours, not taking into account the number of reads per each neighbour. To include that information Pogorelyy suggested to replace number of neighbours *(d)* by the sum of transformed reads number over neighbours:

$$s = \sum_{i=1}^{d} f(c_i),$$

where c_i — number of reads accounter for the i^{th} neighbour; $f(c)$ — transformation.

There are several variants of transformation: $f(c) = c$ corresponds to the sum of reads over all neighbours; $f(c) = log(c)$ gives the sum of their logarithms. A number of reads per clonotype usually follows a power law. So, it can be useful to use a logarithmic scale.

To calculate $P(s|\sigma)$ the following definition was used

$$P(s|\sigma) = \sum_{d} P(s|d)P(d|\sigma),$$

where $P(d|\sigma)$ was calculated as described in (1). $P(s|d)$ is a probability density function (PDF) of random variables sum. As is known, the PDF of the sum of two independent random variables is the convolution of their two PDFs

$$f_{x+y}(z) = \int_{-\infty}^{+\infty} f_x(x)\, f_y(z - x)\, dx$$

$P(s|d)$ is a d-fold convolution of $P_f(f)$ — PDF of reads in a given sample. For example, $P(s|d = 1) = P_f(f)$ or $P(s|d = 2) = \sum_f P_f(f)\, P_f(2f - f) = \sum_f P_f(f)\, P_f(f)$.

Limitations. ALICE is a powerful approach for the identification of epitope-specific TCRs that can work with single repertoire snapshots. This thesis expands the capabilities of the method. In the original ALICE implementation generation probability is calculated by counting TCR frequency in Monte Carlo simulated dataset, which is a quite long process. The high-performance tool OLGA can calculate the generation probability of any sequence much faster. Using OLGA was suggested in (Pogorelyy et al. 2019), but parallel computations were not realised in the script. Furthermore, ALICE operates only with a human model. As murine models are in high demand in science, there was a necessity for a tool that operates with different models. Finally, ALICE was implemented in the form of an R script and is not user-friendly. To overcome these issues one of the objectives of this thesis is to develop an R library based on ALICE statistical model that will work fast both with human and murine models.

2.6 Immune response for vaccination by tumour peptides

Genome instability drives tumour evolution and helps it avoid immune response. Variable driving mutations make one-targeted drug therapy inefficient in many cases. Multiepitope tumour vaccination is a promising immunotherapy. (Kreiter et al. 2015) provided an individual approach to

cancer treatment using epitope vaccination in a murine model. This section overviews this work and previous study.

B16F10 murine melanoma cell line has been widely used for a syngeneic transplantation melanoma model for decades (Patton et al. 2021). (Castle et al. 2012) made the first study that identified somatic point mutations in B16F10 cells. Mutations expectedly occurred in the classical anti-oncogenes including *p53* and genes participating in proto-oncogenic signalling pathways related to proliferation, apoptosis, migration and adhesion. For 50 validated mutations, wild-type peptides and peptides with mutations were designed. Peptides included CD8+ and CD4+ epitopes. In consequence, sixteen out of fifty mutation coding peptides displayed an immune response in immunised mice. Furthermore, eleven out of sixteen peptides had a stronger immune response against peptides with mutation than against wild-type peptides. Two mutated peptides were tested for anti-tumoural activity and showed tumour growth inhibition.

In (Kreiter et al. 2015) the work was continued. Twenty one B16 melanoma mutation-coding peptides were used for mice peptide and mRNA immunisation. Surprisingly, the immune response was mainly CD4+. The authors showed that vaccination with CD4+ specific neo-antigens reshaped the tumour microenvironment, activated cytotoxic immune response and caused tumour growth inhibition.

Peptides used in this thesis project are from these two articles. Table with exact sequences one can find in the methods section. The database of TCRs specific to the B16F10 peptides will be of interest to future research. For instance, it can be useful in immunotherapy studies to identify immune response specificity.

2.7 Summary

T cells play a versatile yet an essential role in antitumor immune response. Cytotoxic T cells accurately and specifically kill cancer cells, while CD4+ T cells modulate their activity. Tumour-specific CD4+ T cells' stimulation activates CD8+ T cells and inhibits tumour growth. This remarkable role of CD4+ T cells in immune response orchestrating merits specific attention. Yet there is no comprehensive picture of murine TCRs specific to the melanoma. In the present study, a cohort of mice was vaccinated with fourteen B16 melanoma peptides targeting CD4+ immune response. Then, TCRβ repertoires were obtained. To identify melanoma-specific TCRs I developed an R library based on the ALICE statistical model. It was necessary because the basic ALICE implementation works only with human TCR repertoires. Additionally, the developed tool turned out to work much faster and is easier to use.

Various tools for identifying the epitope specific TCRs from the TCR repertoire without any specific knowledge about the epitope and MHC were described above. ALICE statistical model combines the strength of probabilistic approach with specific knowledge about homologous TCRs cluster formation. It allows identifying a large number of TCRs involved in the immune response. Thus, the analysis can be carried out without control samples. The resulting melanoma-associated TCR database will be used by me and my colleagues in follow-up immunotherapy study and might be of interest for other researchers as well.

3 METHODS

There are two datasets. To test developed tool TCRβ repertoires of mice vaccinated with Sputnik were used. Identification of melanoma-associated TCRs was made by analysis of mice immunised with B16F10 melanoma peptides. For clarity, the following chapters are divided into two parts according to the datasets.

3.1 Mice vaccinated with Sputnik V

3.1.1 Experiment design

Female 6-8 weeks old C57BL/6 mice were vaccinated with 50 µl in each thigh muscle (100 µl in total). Vaccination was done twice with a two weeks interval between injections. The control group was vaccinated with a placebo, the treatment group with Sputnik V. Control and treatment groups contained 7 and 8 mice respectively. 70-120 µl of blood was collected from the retro-orbital sinus into EDTA (ethylenediaminetetraacetic acid) tubes for isolation of mononuclear cells at every time point. Detailed description of the experiment timeline is shown in Fig. 16.

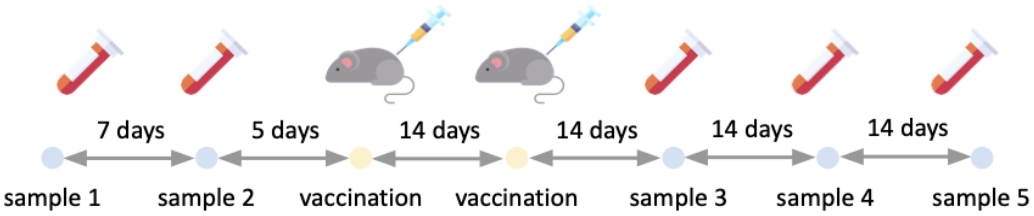


Fig. 16. Experiment timeline. Samples were prepared in five time points: (1) 12 days before the first injection of placebo/vaccine (2) 5 days before the first injection of placebo/vaccine (3) 2 weeks after the second injection (4) 4 weeks after the second injection (5) 6 weeks after the second injection

3.1.2 Library preparation and next-generation sequencing

RNA was extracted from PBMC (peripheral blood mononuclear cell) with RNeasy Micro reagent kit (Qiagen). Then, TCRβ cDNA (complementary deoxyribonucleic acid) libraries were prepared with MiLaboratory reagent kit. Repertoires' amplicons were purified with AMPure-XP (Beckman Coulter, #A63881). All was done according to the manufacturers' protocols. Onwards, TCRβ repertoires were sequenced on Illumina NextSeq and MiSeq platforms.

3.1.3 Data processing

MIGEC (Shugay et al. 2014) was used for UMI (unique molecular identifier) extraction and estimation of sequence number per UMI. To delete potentially erroneously read sequences only

sequences with two and more reads per UMI were taken into analysis. Alignment to the reference was made with MiXCR (Bolotin et al. 2015). Finally, functional sequences were filtered with VDJtools (Shugay et al. 2015). Finally, basic statistics such as number of UMI per sample and number of unique clonotypes were calculated with vdjtools CalcBasicStats command. Data processing pipeline and used commands are present in Fig. 17.

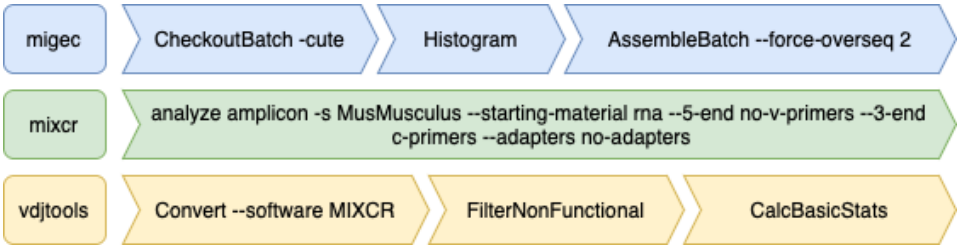


Fig. 17. Data processing pipeline. Used programs are in the rectangles, commands in the arrows. Software and its commands have the same colour

3.1.4 Generation probability model

To inference the generation probability model IGoR (Marcou, Mora, and Walczak 2018) was used on a set of unproductive sequences collected from control and vaccinated mice and 48 additional mice in two time points. There were 63 mice overall. The resulting dataset has 131849 nucleotide sequences in total.

3.1.5 Clonotype tables processing

TCRgrapher pipeline is sensitive to the size of the sample. So, to compare results obtained from different samples all tables of functional clonotypes were normalised to the equal number of unique nucleotide sequences. 4426 was chosen as the largest threshold. Basic statistics of unnormalised samples are present in Supplementary Table 1.

3.1.6 Identification of specific clonotypes by TCRgrapher

TCRgrapher analysis was performed with the following parameters: Q = 6.27 (Q - Selection factor. 1/Q sequences pass selection in the thymus. The values were taken from (Elhanati et al. 2018)), cores = 30, thres_counts = 1, N_neighbors_thres = 1, p_adjust_method = "BH", chain = "mouseTRB", stats = "OLGA". This choice of parameters allows you to save all sequences in the analysis. ‘Model’ parameter was chosen according to the task. The function ‘pval_with_abundance’ was used to recalculate p-values including abundance information.

3.1.7 Identification of specific clonotypes by TCRdist3

Analysis was performed according to the tcrdist3 documentation https://tcrdist3.readthedocs.io/en/latest/. Additionally ‘pandas’ and ‘numpy’ python3 libraries were used.

Individual samples. To find clonotypes with significantly more neighbours after vaccination TCR repertoires from the third, the fourth and the fifth time points were separately compared to the merged TCR repertoires from the first and the second time points. This procedure was repeated for every mouse from control and vaccinated groups. TCR repertoires normalised up to 4426 most abundant clonotypes were used for the analysis. Every TCR repertoire was imported with function import_vdjtools with parameters: “chain = ‘beta’”, “organism = ’mouse’”, “db_file = 'alphabeta_gammadelta_db.tsv'”. Additional column ‘time’ was added for every dataset. Tables from the first and the second time points were marked with ‘before’ in the ‘time’ column. Table with TCR repertoire after vaccination was marked with ‘after’. Three tables were concatenated, TCRrep object was analysed with compute_distances method. Finally, ‘neighborhood_diff’ was used with parameters ``knn_radius = 16” and “test_method='fishers'”. It means that for every clonotype its number of neighbours (CDR3 sequences with distance less than 16 according to tcrdist metric) were compared using Fisher’s exact test. The uniform radius equal to 16 was used according to recommendations in (Mayer-Blackwell et al. 2021). Additionally, ‘calc_radii’ function was used to calculate the optimal radius for every clonotype separately. The mean radius was used by the ‘neighborhood_diff’ function leaving other parameters the same.

Merged samples. Full unnormalised TCR repertoires of eight mice vaccinated with Sputnik V were merged together. Clonotypes with identical nucleotide sequences from different mouses were considered as distinct clonotypes. The analysis was similar to that described above. The datasets from points before vaccination were combined into a common table. Every time point after vaccination was compared to a combined dataset from the first two time points. To calculate distances function ‘compute_sparse_rect_distances’ were used with parameter ‘radius=16’ to reduce computational time and memory usage.

3.1.8 Identification of specific clonotypes by GLIPH2

GLIPH2 is a web tool available on http://50.255.35.37:8080. It analyses TCR repertoire for shared CDR3 motifs that are enriched compared to their expected frequencies in the naive TCR repertoire. As it said on the website: “GLIPH returns significant motif lists, significant TCR

convergence groups, and for each group, a collection of scores for that group indicating enrichment for motif, V-gene, CDR3 length and proliferation count". GLIPH2, also, returns Fisher's exact test results for every cluster. Input is compared with reference naive TCR repertoire.

Individual samples. Normalised by top 4426 clonotypes TCR repertoires of seven mice from control group and eight mice vaccinated with Sputnik V were analysed in three time points after vaccination by GLIPH2 using web interface.

Merged samples. Unnormalised TCR repertoires of vaccinated mice were pooled together at every time point. Clonotypes from distinct mice were not combined together. Merged samples were analysed using GLIPH2 web interface.

3.1.9 Identification of expanded clonotypes by EdgeR

Primarily, a count table was produced. For this purpose full unnormalised individual murine TCR repertoires were used. Unnormalised counts of clonotypes with the same nucleotide sequence were added up. Clonotype count is a number of UMI per clonotype unique nucleotide sequence. In the count table every column corresponded to one sample. There were seven control samples and eight samples from the vaccinated group. Every mouse had five time points: two points before vaccination and three points after vaccination. Then, the EdgeR object was analysed with the following design: 'model.matrix(~group + group:time, data=metadata)'. First step was data filtration with the 'filterByExpr' function with parameters 'min.count = 2' , 'min.total.count = 5', 'large.n = 2', 'min.prop = 0.5'. Second step was normalisation by the 'calcNormFactors' function and the trimmed mean of the M-values (TMM) method. Third step was dispersion estimation by the 'estimateDisp' function using the 'trended' dispersion estimation method. To find significantly expanded clonotypes the 'glmQLFTest' was performed.

3.2 Mice immunised with B16 melanoma peptides

3.2.1 Experiment design

Mice were immunised by B16F10 melanoma peptides with Freund's adjuvant in the pads of their hind paws. The dose was divided into two paws. Two and a half weeks later, popliteal lymph nodes were collected. T cells were sorted into three subsets: CD8+ T cells, T-helper cells and regulatory T cells. Experiment design schematic representation is in Fig. 18.

Fig. 18. Schematic representation of experiment design with mice immunised by B16F10 melanoma peptides

Peptides names and amino acid sequences are presented in Table 1. There were fourteen peptides. One peptide per group. Five peptides (p5, p17, p20, p30, p48) have two experimental groups. One group consists of five mice, except control and one of the group vaccinated with p17. These two groups have three mice per group. Control group was injected with adjuvant only.

Table 1. Peptides used for immunisation

peptide id	peptide name	sequence
p5	B16-M05-Eef2	FVVKAYLPVNESFAFTADLRSNTGGQA
p12	B16-M12-Gnas	TPPPEEAMPFEFNGPAQGDHSQPPLQV
p17	B16-M17-Tnpo3	VVDRNPQFLDPVLAYLMKGLCEKPLAS
p20	B16-M20-Tubb3	FRRKAFLHWYTGEAMDEMEFTEAESNM
p22	B16-M22-Asf1b	PKPDFSQLQRNILPSNPRVTRFHINWD
p25	B16-M25-Plod1	STANYNTSHLNNDVWQIFENPVDWKEK
p27	B16-M27-Obsl1	REGVELCPGNKYEMRRHGTTHSLVIHD
p28	B16-M28-Ppp1r7	NIEGIDKLTQLKKPFLVNNKINKIENI
p30	B16-M30-Kif18b	PSKPSFQEFVDWENVSPELNSTDQPFL
p33	B16-M33-Pbk	DSGSPFPAAVILRDALHMARGLKYLHQ
p44	B16-M44-Cpsf3l	EFKHIKAFDRTFANNPGPMVVFATPGM
p47	B16-M47-Rpl13a	GRGHLLGRLAAIVGKQVLLGRKVVVVR
p48	B16-M48-Def8	SHCHWNDLAVIPAGVVHNWDFEPRKVS
p50	B16-M50-Sema3b	GFSQPLRRLVLHVVSAAQAERLARAEE

3.2.2 Library preparation and next-generation sequencing

Library preparation and sequencing matched described in 3.1.2 for the dataset of mice vaccinated with Sputnik.

3.2.3 Data processing

Data processing was similar to that described in 3.1.3 for the first dataset.

3.2.4 Generation probability model

Generation probability models were inferred by IGoR (Marcou, Mora, and Walczak 2018) for every T cell subset from non-functional sequences collected from 93 mice. There were 252157 TCRβ CDR3 non-functional sequences obtained from CD8+ T cells; 265110 sequences from Th cells and 93957 form Tregs.

3.2.5 Clonotype tables processing

Every TCRβ CDR3 repertoire was normalised up to 5000 the most abundant clonotypes. Further, repertoires were merged together. Identical clonotypes were not merged. For a group of five mice the resulting clonotype table included 25000 clonotypes.

3.2.6 Identification of specific clonotypes by TCRgrapher

TCRgrapher parameters didn't differ from those used in analysis of mice vaccinated with Sputnik V.

3.3 Code availability

TCRgrapher is available on GitHub https://github.com/KseniaMIPT/tcrgrapher. IGoR inference output is stored in the 'models' directory.

All significant melanoma-associated clonotypes and corresponding additional information available online: https://github.com/KseniaMIPT/B1610-melanoma-associated-murine-TCRs.

3.4 Statistical analysis

Used statistical tests and parameters are written in the figure descriptions.

3.5 Implementation and visualisation

TCRgrapher was turned into an R library according to recommendations from https://r-pkgs.org/index.html. The following supporting packages were used: 'devtools', 'roxygen2'.

TCRgrapher dependencies: 'stats', 'utils', 'stringdist', 'data.table', 'foreach', 'parallel', 'doParallel', 'igraph'.

The following R packages were used for analysis and visualisation: 'ggplot2', 'ggpubr' 'RColorBrewer', 'ggrepel' 'ggseqlogo', 'gridExtra', 'eulerr', 'ggnet'.

4 RESULTS

4.1 TCRgrapher

TCRgrapher, R library for identifying specific TCRs was developed. It yields specific TCRs from a single TCRβ CDR3 repertoire based on the ALICE statistical model described in the LITERATURE REVIEW section. In brief, TCRgrapher finds TCRs with significantly more neighbours than by chance, where the neighbour is a CDR3 sequence with one amino acid mismatch or without it. TCRgrapher operates with murine, human or any custom model. For calculating the generation probability TCRgrapher uses the high-performance tool OLGA. Parallel computations are available.

4.1.1 TCRgrapher pipeline

In this section I provide a detailed explanation of how the TCR grapher works and present its pipeline (see flowchart in Fig. 19).

TCRgrapher takes as an input a clonotype table with nucleotide and amino acid CDR3 TCR sequences, V and J segments and the number of counts for every sequence. The pipeline can operate both with human α and β chains models and mouse β chain model. As an option, any custom model can be used. Data passes through several filtering steps: filtering of unproductive sequences, filtering by V and J segments present in the model and by the number of counts. After calculating neighbours for each sequence, they are filtered by the number of neighbours.

To estimate the expected mean number of neighbours λ for sequence σ a dataset with all possible neighbours is generated. Then, P_{post} is calculated for the original sequence and every generated one. To get a probability under the condition of having a given VJ combination the sum is divided by the probability of VJ combination according to Bayes' law. Then, the mean is calculated by the formula (2).

There are two computational options for P_{post} calculation. First, P_{gen} could be calculated for every sequence by OLGA and then be multiplied by selection factor Q common to all sequences. After multiplication, we get the probability to be in the repertoire under the condition to be selected with probability $q = 1/Q$.

The second option is to calculate P_{gen}, Q and P_{post} by SONIA. Additionally in both cases, P_{post} should be divided by the probability of the VJ combination since this condition was not taken into account in the generation of all possible mismatches.

Finally, p-value estimation is performed using formulas (1) and (2). Adjusted p-values can be obtained with any method available by the 'p.adjust' function from 'stats' R package. Output is the same clonotype table as an input extended by additional columns with statistics.

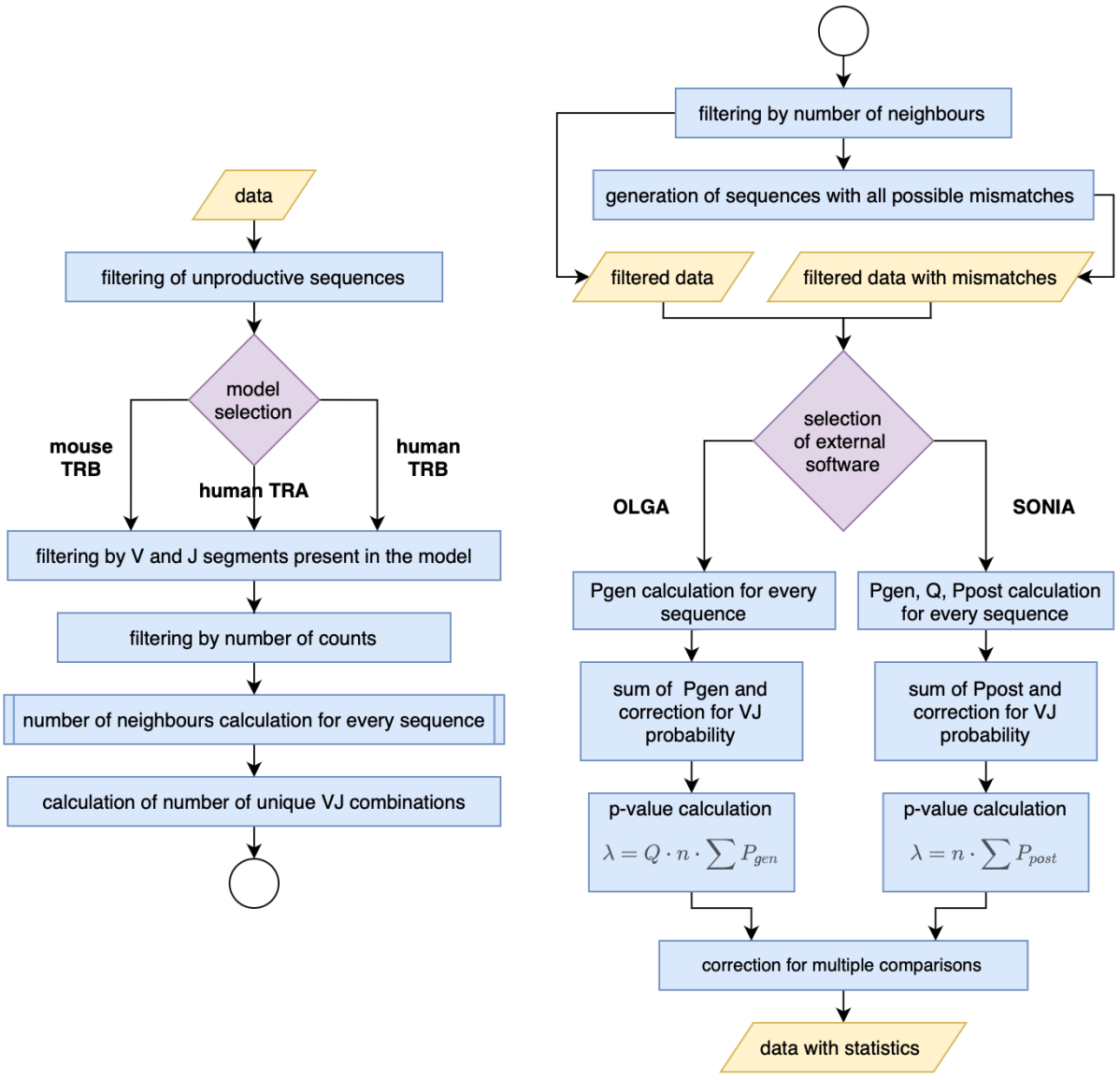
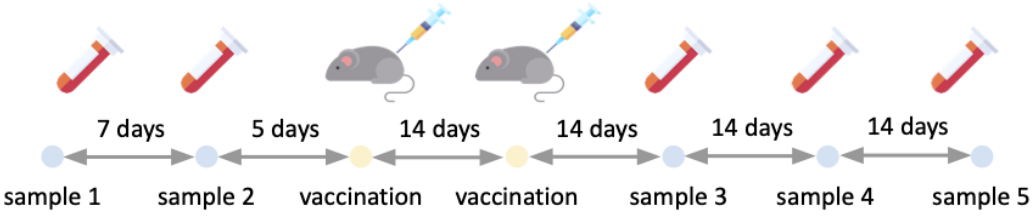


Fig. 19. TCRgrapher pipeline.

4.2 TCRβ CDR3 repertoires of mice vaccinated with Sputnik V

To test TCRgrapher performance the mice vaccinated with Sputnik V were used. Using ELISPOT (Enzyme-Linked ImmunoSpot) Sputnik has independently been shown to elicit a strong T cell response. There were seven mice in the control group and eight mice in the vaccinated group. TCRβ CDR3 repertoires obtained from PBMC (peripheral blood mononuclear cell) were prepared as described in the METHODS section. Only the sequences with two or more reads per UMI were taken into analysis to avoid any influence of sequencing errors on the repertoire diversity. The number of functional clonotypes in the final tables ranged from 4426 to 40411 clonotypes per mouse. Statistics of clonotype and UMI number per sample is shown in Appendix Table 1.



Experiment timeline. Samples were prepared in five time points: (1) 12 days before the first injection of placebo/vaccine (2) 5 days before the first injection of placebo/vaccine (3) 2 weeks after the second injection (4) 4 weeks after the second injection (5) 6 weeks after the second injection

4.2.1 Own generation probability model performed better than standard OLGA model

By default TCRgrapher works with the standard OLGA generation probability model. As an option any model in the format of IGoR output can be used. Individual murine TCR repertoires were analysed by both ways. The comparison of total frequency of significant clonotypes between control and vaccinated groups is shown in Fig. 20.

As a result, using the own generation probability model trained on non-functional sequences of mice used in analysis gave better separation of the control and immunised groups than standard OLGA generation probability model. The OLGA standard model didn't give statistically significant difference between group medians at 5% confidence level, while 'own' model showed statistically significant results.

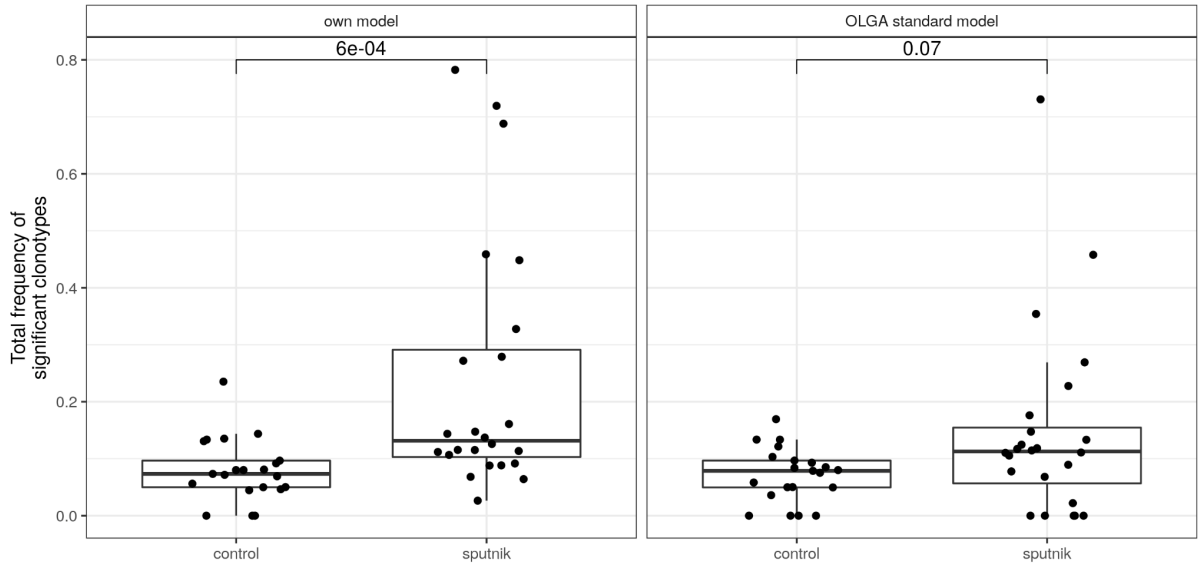


Fig. 20. Total frequency distribution of significant clonotypes between control group and group vaccinated with Sputnik. Clonotype was identified as significant if adjusted p-value was less than 0.05. Each dot represents one mouse in one of three time points after vaccination. Three dots for each mouse from seven in the control group and eight in the vaccinated group. Samples were normalised up to 4426 most abundant clonotypes. Analysis was made using the TCR CDR3 generation probability model inferred from non-functional sequences of 63 C57BL/6 mice ('own model') and using standard OLGA model for mouse TCRβ chain ('OLGA standard model'). To compare the medians between two groups the Wilcoxon rank sum test was used. Received p-values are pictured on the top of the plots.

4.2.2 Individual murine TCR repertoire analysis by TCRgrapher and other tools

TCR repertoires of seven mice from the control group and eight vaccinated mice were analysed by three tools: TCRgrapher, GLIPH2 and TCRdist3. In both groups specific TCRs were identified in three time points after vaccination. Analysis with TCRdist3 required comparison with repertoires before vaccination. GLIPH2 compares imputed TCR repertoire with its own reference dataset of naive TCRs. TCRgrapher identified specific TCRs using its own statistical model. Detailed description of analysis one can find in the METHODS section. Additionally, the working principle of GLIPH2 and TCRdist are presented in the LITERATURE REVIEW section. For every sample the total frequency of clonotypes identified as significant was calculated. In Fig. 21. one can see the comparison of distributions of shares occupied by significant clonotypes after vaccination between the control group and the group immunised with Sputnik.

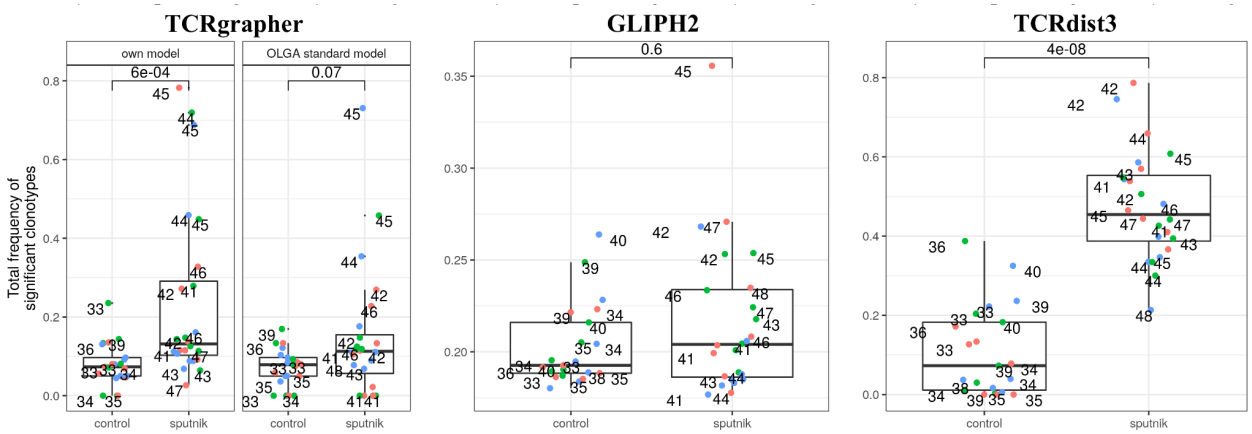


Fig. 21. Distribution of total frequency of significant clonotypes between control and Sputnik vaccinated groups. Each dot represents one mouse in one of three time points after vaccination. Three dots for each mouse from seven in the control group and eight in the vaccinated group. Samples were normalised up to 4426 most abundant clonotypes. To compare the medians between two groups the Wilcoxon rank sum test was used. Received p-values are pictured on the top of the plots. **TCRgrapher result.** Clonotype was identified as significant if adjusted p-value was less than 0.05. Analysis was made using the TCR CDR3 generation probability model inferred from non-functional sequences of 63 C57BL/6 mice ('own model') and using standard OLGA model for mouse TRB chain ('OLGA standard model'). **GLIPH2 and TCRdist3 result.** Clonotype was considered as significant if Fisher's exact test adjusted p-value was less than 0.05. The Benjamini-Hochberg procedure was used for correction for multiple comparisons for all three methods.

GLIPH2 showed poor results without any significant difference between medians of the control and vaccinated groups. TCRgrapher with own generation probability model and TCRdist3 found statistically significant differences between the control group and the group vaccinated with Sputnik at 5% confidence level. Moreover, the median ratio of the repertoire occupied by the potentially vaccine-specific TCRs is higher for TCRdist3 results than for TCRgrapher. Interestingly, both me showed that more than 3500 clonotypes may be involved in the immune response in a mouse.

Additionally, TCRdist3 analysis with mean optimal radius was performed. One can find details in the METHODS section. This type of analysis proved to be worse than using a single conservative radius. Therefore, in the following research only the radius equal to 16 was used.

4.2.3 TCRgrapher: how merging of TCR repertoires from different mice affects the result

Tracking the frequency of clonotypes through all five time points for the vaccinated group is presented in Fig. 22. For five out of eight mice the most abundant clonotype was the same. Its amino acid sequence is highlighted by green. In all TCR repertoires this sequence was not presented before vaccination and occupied a remarkable share of the repertoire. This clonotype is most probably specific for the vaccine. However, in two of three mice this clonotype was not identified as significant. For the largest clonotypes in every mouse the clusters were plotted. Nodes are clonotypes with the

same VJ combination and CDR3 length. Edges connect the clonotypes with one or fewer amino acid mismatches. The picture shows that insignificant clonotypes have one neighbour or do not have it at all, while significant clonotypes are included in the dense cluster. Assuming that the largest clonotypes after vaccination are specific, we can conclude that due to the stochastic nature of TCR repertoire in some cases specific clonotypes do not have enough neighbours to be detected by TCRgrapher. The solution for this issue is to merge several TCR repertoires from different mice together and to analyse the merged samples by TCRgrapher. I assumed that there is an optimal number of samples and TCRs in them, which makes it possible to find the largest specific TCRs and yield reproducible results.

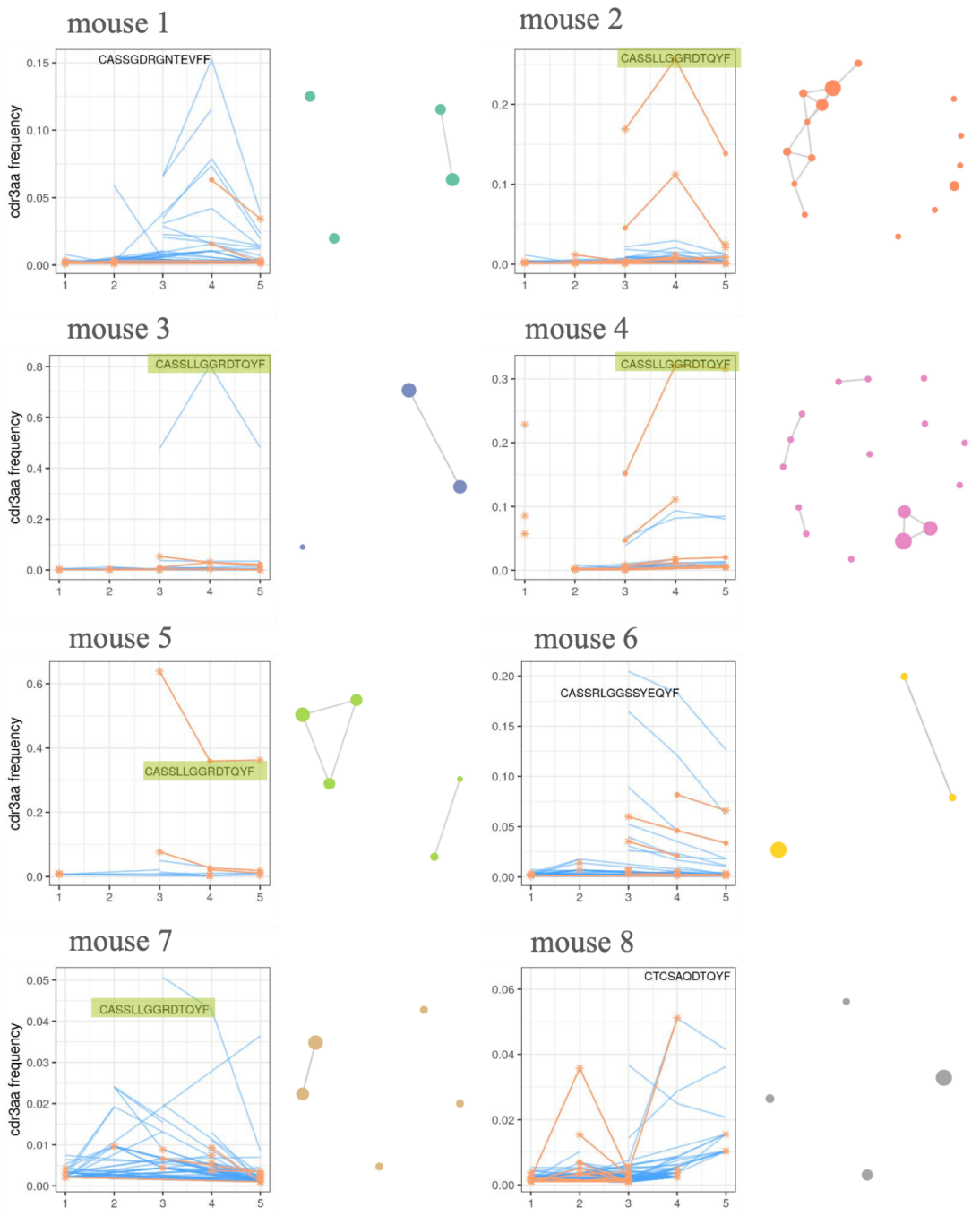


Fig. 22. Tracking the frequency of clonotypes through five time points for mice vaccinated with Sputnik V. Every line corresponds to one amino acid clonotype. Amino acid clonotype frequency in the repertoire is plotted along the y-axis. Time points are plotted along the x-axis. TCR repertoires were normalised up to 4426 most abundant clonotypes. Clonotypes that were identified as significant by TCRgrapher at any time point are colored orange. Clonotype was identified as significant if adjusted p-value was less than 0.05. Insignificant clonotypes are colored blue. If a clonotype was identified as significant in a particular time point, in that time point it was marked with an asterisk. Amino acid sequences of the largest clonotypes are written at the top of the plots for every mouse. Identical sequences are highlighted by green. Next to the tracking plots are the clusters that include the most abundant clonotype. Nodes are unique amino acid sequences with the same VJ combination and the same CDR3 length. Edges connect sequences that differ by one amino acid mismatch.

To find the optimal number of samples I tested all combinations of eight vaccinated mice taken from one to seven times and used TCRgrapher on the merged samples. In Fig. 23 results of the analysis are presented. Normalised number of clusters with clonotypes identified as significant increases remarkably with an addition of mice in the merged repertoire from initial individual repertoire to (merged from) four mice. Starting with five mice per combination, the increase in the number of significant clusters grows slowly. Mean adjusted p-value obtained by TCRgrapher, also, has a saturation point in four mice per merged repertoire. However, the behaviour of the points before vaccination differ in the two graphs. In the first one they behave similar to the points after vaccination, while in the second graph they do not show any dependence on the number of mice in the sample.The mean adjusted p-value falls from one to four mice in analysis, which can be interpreted as follows. Adding each new mouse to the analysis makes it possible to find more rare clonotypes. A subsequent increase in the number of mice greater than four does not affect the mean adjusted p-value.

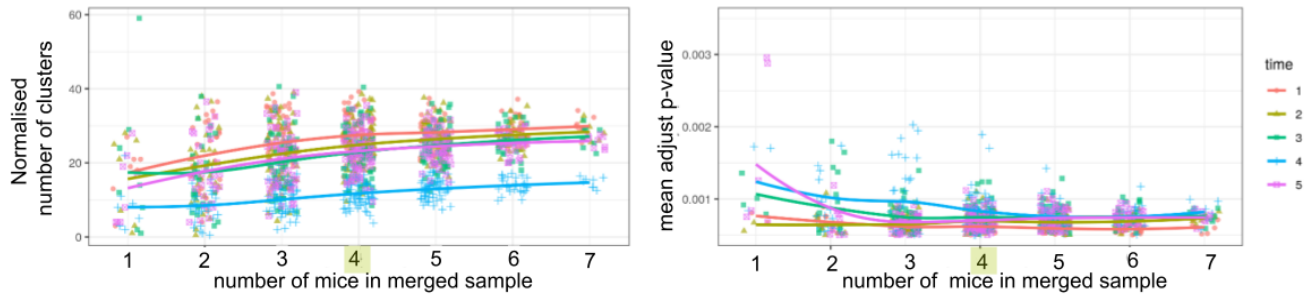


Fig. 23. TCRgrapher analysis of merged TCR repertoires from mice vaccinated with Sputnik V. Number of merged TCR repertoires is plotted along the x-axis for both plots. Every line corresponds to one time point. On the left plot number of clusters with significant clonotypes divided by the number of mice used in repertoires merging is plotted along the y-axis. Clonotype was identified as significant if adjusted p-value in TCRgrapher result was less than 0.05. The Benjamini-Hochberg procedure was used for correction for multiple comparisons. On the right plot mean adjust p-value received by TCRgrapher is plotted along y-axis. Every individual murine TCR repertoire was normalised up to 4426 most abundant clonotypes.

Distribution of the number of neighbours normalised by the number of sequences with the same VJ combination for TCR repertoires combined from one to seven mice is presented in Fig. 24. This picture illustrates how an increase in the number of mice affects the structure of the repertoire. With an increase in the number of mice in a sample, the average number of neighbours decreases along with the variation. There are only small stochastic changes in the distribution after reaching four mice per sample.

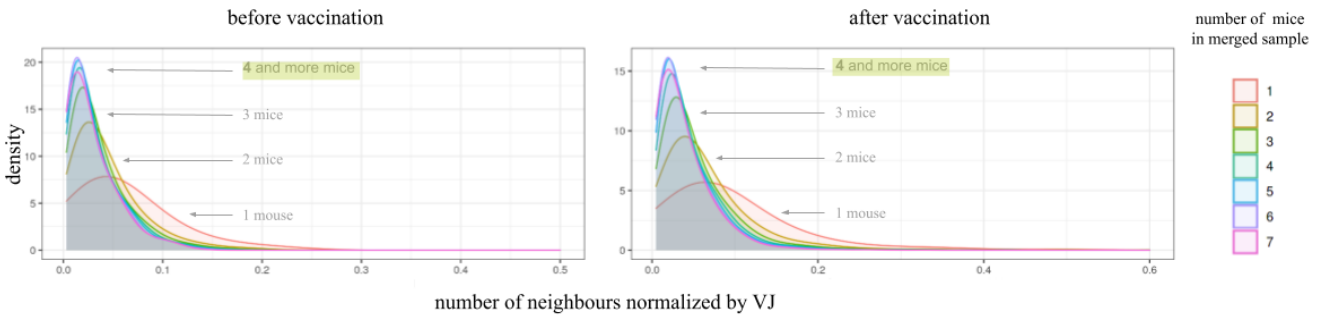


Fig. 24. Distribution of the number of neighbours normalised by the number of sequences with the same VJ combination for TCR repertoires combined from one to seven mice. Neighbour is a TCRβ CDR3 sequence differ from considered by one or zero amino acid mismatch. Sequences should have the same VJ combination and CDR3 length. The left plot presents distributions for merged samples before vaccination. The right plot presents distributions for merged samples after vaccination. Every individual murine TCR repertoire was normalised up to 4426 most abundant clonotypes. Then all combinations from one to seven mice in the merged repertoire were taken.

These results suggest that four mice per sample with 17704 clonotypes in a merged TCR repertoire is enough to yield a reliable result by TCRgrapher. A further increase will not lead to a qualitative improvement in the result.

4.2.4 Analysis of merged TCR repertoires by TCRgrapher and other tools

Full TCR repertoires of eight mice were merged together at each time point. Number of functional TCRβ CDR3 sequences in the merged samples from groups vaccinated with Sputnik are in Table 2.

Table 2. Number of functional TCRβ CDR3 sequences in the merged samples from the group vaccinated with Sputnik. There were eight mice per the group.

Time point	Number of functional sequences
1	200 455
2	169 856
3	170 607
4	101 128
5	193 835

Resulting TCR repertoires were analysed by three tools: TCRgrapher, TCRdist3 and GLIPH2. EdgeR worked on unmerged full TCR repertoires. Details are in the METHODS section. Fig. 25. shows clonotype tracking through five time points. Background TCRs are colored blue. TCRs

identified as significant by the corresponding tools marked with a special colour. GLIPH2 showed poor results. It was unable to identify the largest clonotypes. TCRgrapher, TCRdist3 and EdgeR showed good results, identifying the most abundant TCRs.

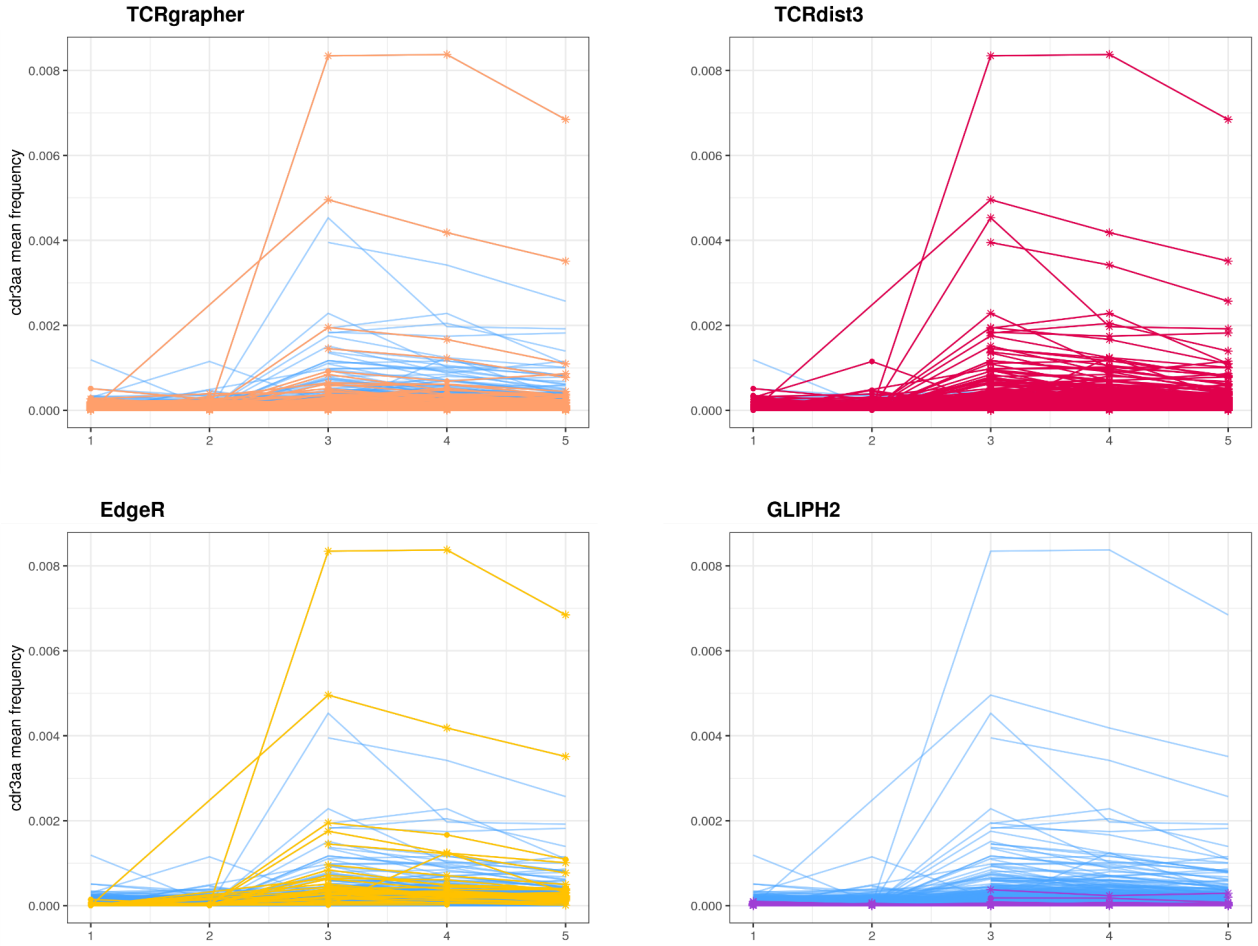


Fig. 25. Tracking the mean frequency of clonotypes of vaccinated group from merged samples through five time points. Full unnormalised TCR repertoires of eight mice vaccinated with Sputnik were merged together in every time point. Every line corresponds to one amino acid clonotype. Amino acid clonotypes mean frequency in the merged repertoire is plotted along the y-axis. Time points are plotted along the x-axis. Insignificant clonotypes are colored blue. If a clonotype was identified as significant in a particular time point, in that time point it was marked with an asterisk. The Benjamini-Hochberg procedure was used for correction for multiple comparisons for all four methods. **TCRgrapher.** Clonotypes that were identified as significant by TCRgrapher at any time point are colored orange. Clonotype was identified as significant if adjusted p-value was less than 0,05. **TCRdist3.** Clonotypes that were identified as significant by TCRdist3 in any time point are colored pink. Clonotype was identified as significant if adjusted p-value obtained by Fisher's exact test was less than 0.05. **EdgeR.** Clonotypes that were identified as significant by EdgeR in any time point are colored yellow. Clonotype was identified as significant if FDR (false discovery rate) was less than 0.05. **GLIPH2.** Clonotypes that were identified as significant by GLIPH2 in any time point are colored purple. Clonotype was identified as significant if adjusted p-value obtained by Fisher's exact test was less than 0.05.

TCRdist3 identified the largest number of clonotypes — more than 1500 in every time point. TCRgrapher revealed fewer but still large numbers of clonotypes — more than 800 on average (Fig.

26). EdgeR found 30, 14 and 22 significantly expanded clonotypes in the third, the fourth and the fifth time points respectively. Moreover, all clonotypes were also found by TCRdist3 or TCRgarpher. Surprisingly, TCRgrapher and TCRdist results were quite different. Around 50 clonotypes in every time point after vaccination were identified by both tools. It means that these clonotypes have significantly more neighbours as compared to time points before vaccination and more neighbours than expected according to the ALICE statistical model.

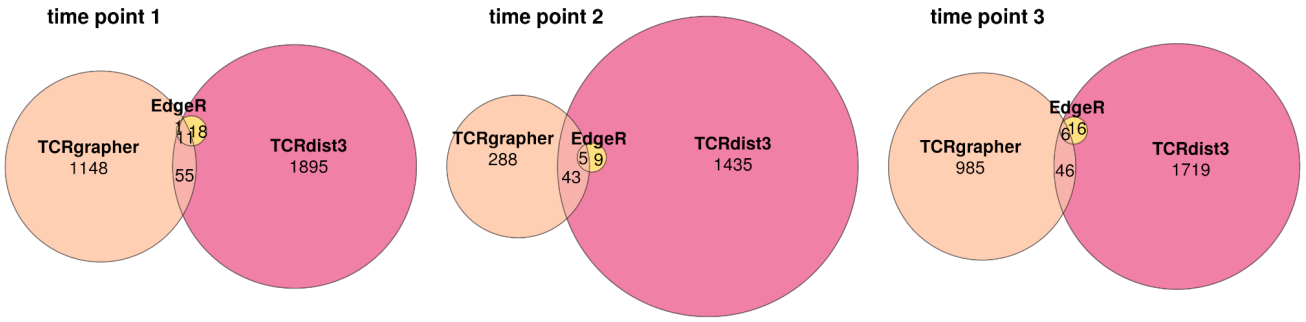


Fig. 26. Venn diagrams showing intersection between clonotypes identified as significant by TCRgrapher, EdgeR and TCRdist3 in three time points after vaccination. **TCRgrapher.** Clonotypes that were identified as significant by TCRgrapher at any time point are colored orange. Clonotype was identified as significant if adjusted p-value was less than 0,05. **TCRdist3.** Clonotypes that were identified as significant by TCRdist3 in any time point are colored pink. Clonotype was identified as significant if adjusted p-value obtained by Fisher's exact test was less than 0.05. **EdgeR.** Clonotypes that were identified as significant by EdgeR in any time point are colored yellow. Clonotype was identified as significant if FDR was less than 0.05. **GLIPH2.** Clonotypes that were identified as significant by GLIPH2 in any time point are colored purple. Clonotype was identified as significant if adjusted p-value obtained by Fisher's exact test was less than 0.05.

It is important to add that TCRgrapher and GLIPH2 parsed the merged data much faster than TCRdist3. When the repertoire size was several hundred sequences, the approximate runtime of TCRdist3 was four days, while TCRgrapher and GLIPH2 solved the problem in a few hours. Moreover, TCRgrapher p-values were recalculated taking into account information about the cluster's frequency. The result was mostly the same. So, all TCRgrapher results presented in this study go without information about clonotypes abundance.

4.2.5 TCRs clusters identified by TCRgrapher

TCRgrapher identified the largest expanded clusters (Fig. 27). However, some small clusters that had stable frequency through all time points were identified as significant too. Such clusters are not likely to be vaccine-associated. They probably represent a common immune challenge encountered by mice prior to vaccination. Therefore, in the case of searching for TCRs specific to a recent challenge, a threshold for the size of clusters should be used.

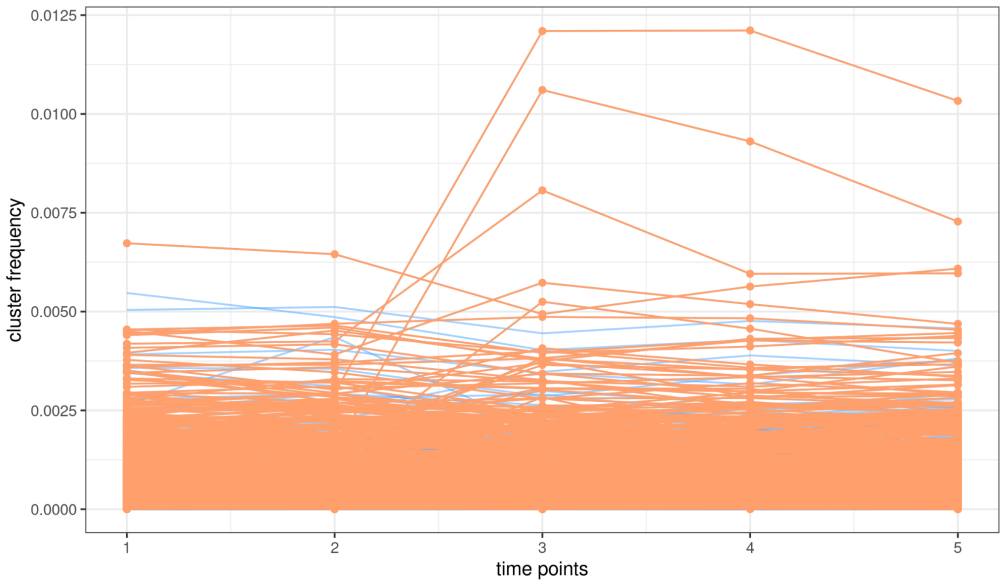


Fig. 27. Tracking the frequency of the clusters with significant clonotypes identified by TCRgrapher through five time points. TCRgrapher analysis were performed on merged TCR repertoires from eight mice from the vaccinated group. Significant clonotypes are colored orange. Insignificant clonotypes are colored blue. Clonotype was identified as significant if adjusted p-value was less than 0,05. Cluster is a group of similar TCRβ CDR3 sequences aligned to the same V- and J- gene, with the same CDR3 length. If one presents clonotypes of a cluster as a graph where edges connect clonotypes with no more than one amino acid mismatch, the graph will be connected.

CASSLLGGRDTQYF is a clonotype with the highest mean frequency after vaccination among eight murine TCR repertoires. After vaccination, expanded clonotypes formed dense core cluster (Fig. 28.).

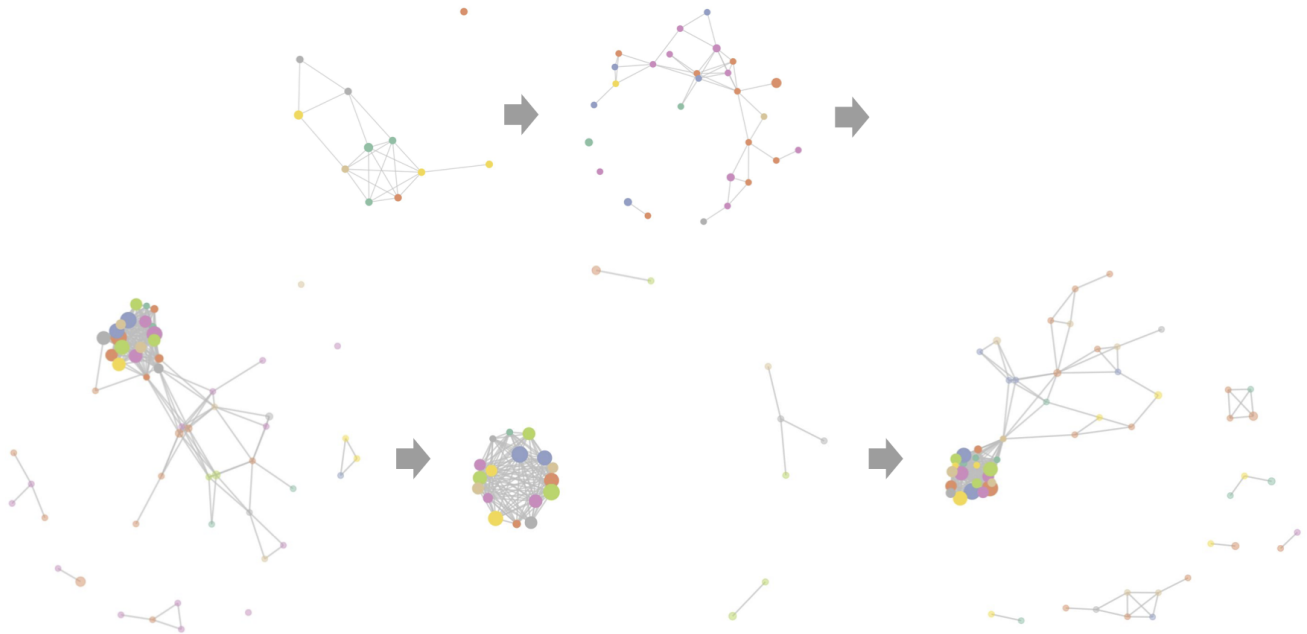


Fig. 28. Transformation of the cluster with the most abundant clonotype CASSLLGGRDTQYF through five time points. Every node corresponds to one amino acid clonotype. Nodes are colored by mouse. There are eight

mice from the group vaccinated with Sputnik V. Edges connect clonotypes differ by one amino acid mismatch or identical with different nucleotide sequences or clonotypes from different donors with one mismatch or without it. All clonotypes were aligned on TRBV16 and TRBJ2-5 and have CDR3 length equal to 14. Clonotype was identified as significant by TCRgrapher if the adjusted p-value was less than 0,05. Transparent clonotypes are insignificant.

4.3 TCRβ CDR3 repertoires of mice immunised with B16F10 melanoma peptides

To identify melanoma-specific T cell clonotypes, TCRβ CDR3 repertoire profiling of mice immunised with fourteen B16F10 melanoma peptides were performed. Basic statistics for repertoires is shown in Appendix Table 2. Overview of the pipeline is presented in Fig. 29. Taking into account the experience gained in the analysis of the mice vaccinated with Sputnik, the following procedure was developed. First, by IGoR TCR generation probability models were inferred for every T cell subset using non-functional sequences. Second, TCR repertoires were normalised up to 5000 most abundant clonotypes and TCR repertoires of mice from one group were merged together. Then, TCRgrapher analysed pooled samples with 'own' generation probability model from the first step. Clonotypes were identified as significant if TCRgrapher adjusted p-value was less than 0.05. Significant clonotypes were checked for intersections. If clonotype was presented in groups vaccinated with different peptides and in both groups was identified as significant, it was deleted from both repertoires. Additionally, clusters formed by significant clonotypes from one mouse were removed. At the end, clusters were filtered by total frequency threshold. The final result included only clusters with a frequency above 0.25%.

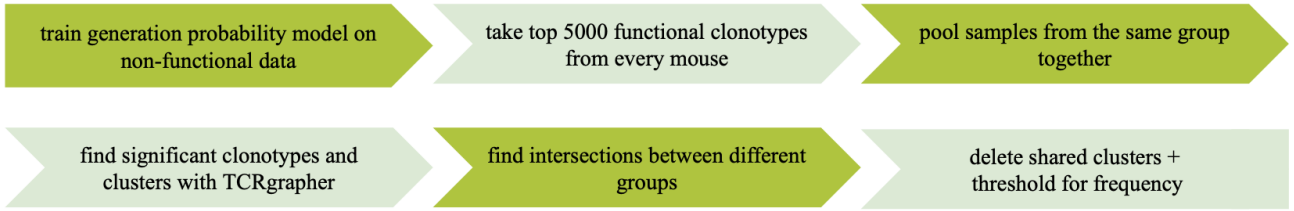


Fig. 29. Analysis pipeline. There were 14 peptides in analysis. Group consisted of five mice, except the control group and one out of two groups vaccinated with p17. Every group was immunised by one peptide. Five peptides had two experimental groups. Threy T cell subsets were collected from one mouse. 288 TCRβ CDR3 repertoires were extracted in total.

The clonotypes determined in this way are highly likely not public, as they have passed several stages of verification. Thanks to the analysis of several mice, the common immune response pattern is found. As it was revealed earlier five mice per sample is enough to get valuable results.

Schematic representation of experiment design with mice immunised by B16F10 melanoma peptides

4.3.1 T-helper cells showed the most variable immune response

TCRgrapher identified the largest number of significant clonotypes in T-helper cells as expected (Fig. 30). All Th groups have more significant clonotypes than control group. Interestingly, using own generation probability model instead of standard one gave fever significant clonotypes. Standard OLGA model identified a lot of clonotypes that were shared across different groups especially in CD8+ T cells. Naturally, mentioned above is also true for the number of clusters of significant clonotypes (Appendix Fig. 2).

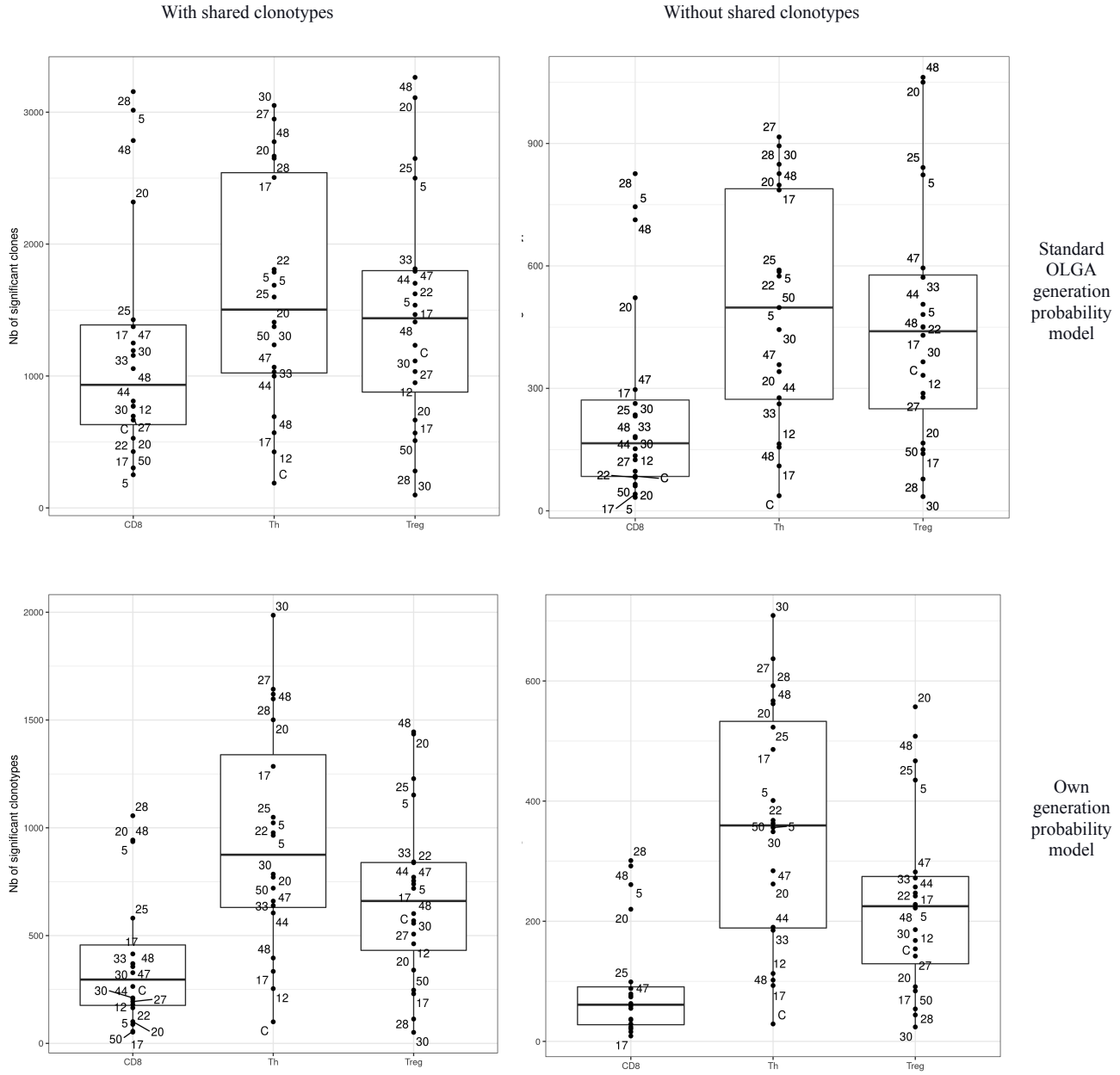


Fig. 30. Distribution of number of significant clonotypes in the samples between T cell subsets. Clonotype was identified as significant by TCRgrapher if the adjusted p-value was lower than 0.05. There were 14 groups immunised with different B16F10 melanoma peptides and control groups. Five peptides have two experimental

groups. Each dot corresponds to one group. There were five mice per group, except control groups and one p17 group.

CD8+ T cells formed the largest clusters of significant clonotypes compared to Th and Tregs. However, these clusters were formed from public clonotypes. After shared clonotypes deletion CD8+ T cells has the sample mean size of the cluster among T cell subsets (Fig. 31).

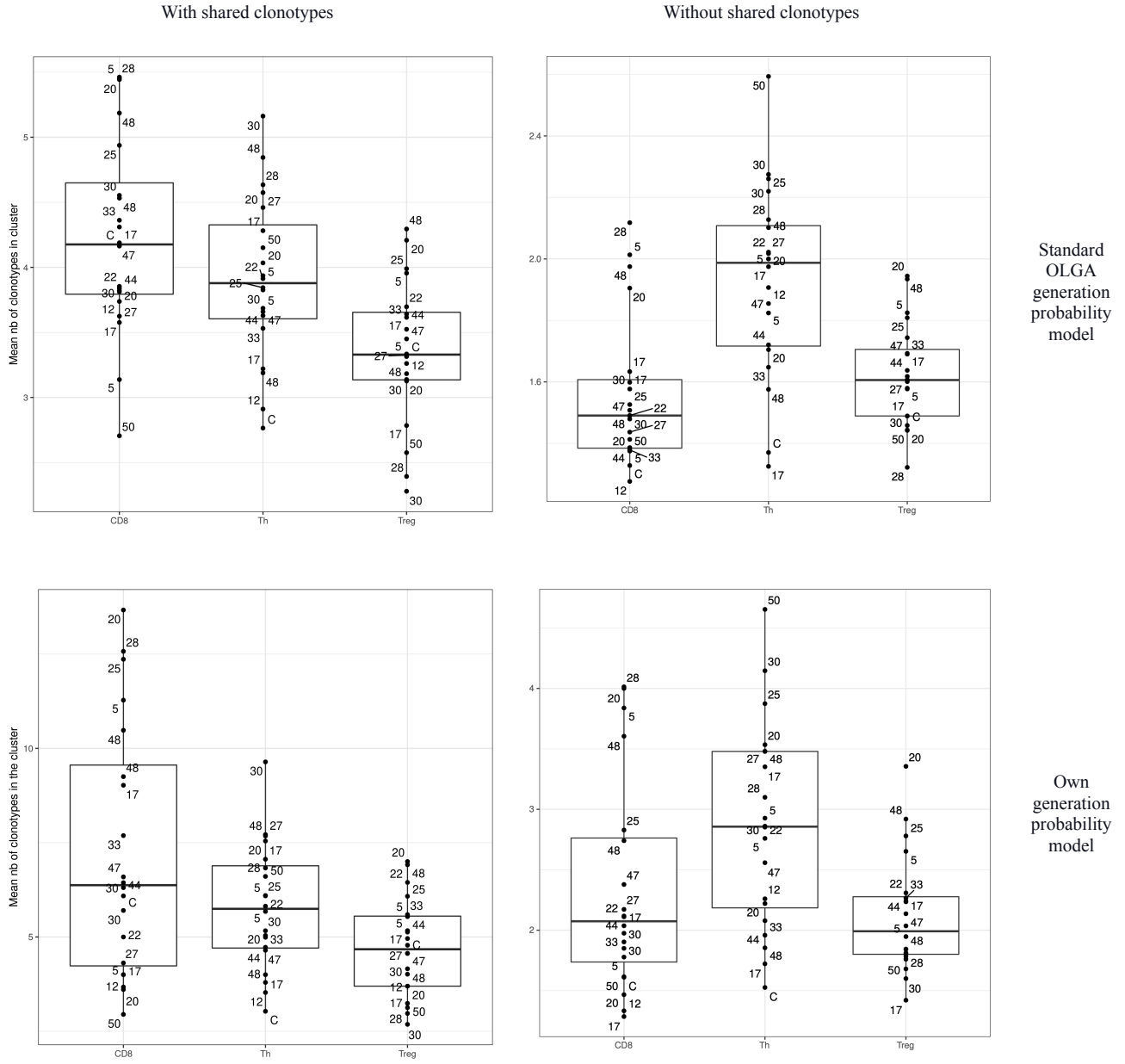


Fig. 31. Distribution of the mean size of the cluster formed by significant clonotypes between T cell subsets. Clonotype was identified as significant by TCRgrapher if the adjusted p-value was lower than 0.05. There were 14 groups immunised with different B16F10 melanoma peptides and control groups. Five peptides have two experimental groups. Each dot corresponds to one group. There were five mice per group, except control groups and one p17 group.

4.3.2 Cluster frequency as a potential marker of specificity

Clusters of significant clonotypes discovered by TCRgrapher had peculiar frequency distribution. While the largest clusters identified in the control sample had relatively equal frequency, the largest peptide-associated clusters have a large lead in relation to the subsequent ones (Fig. 32).

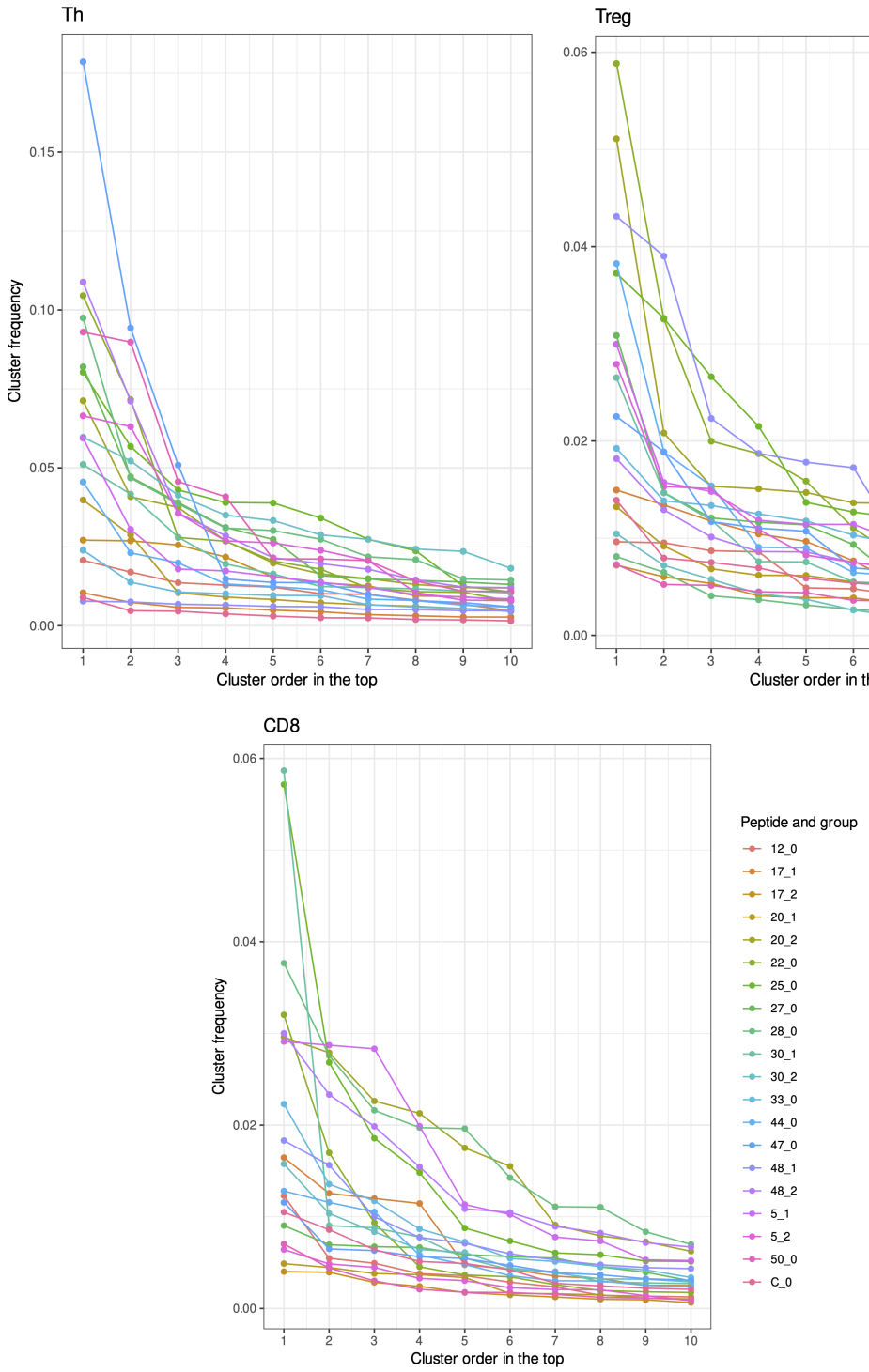


Fig. 32. Cluster frequency distribution of 10 the most abundant clusters of significant TCRs. Clonotype was identified as significant by TCRgrapher if the adjusted p-value was lower than 0.05. Own generation probability model was used for TCRgapher analysis. Each line corresponds to one group.

The maximum frequency of the cluster in the control group is 0.25%. It was selected as the threshold for frequency. To check the validity of the threshold frequency distributions were constructed among all groups of those clusters that were identified as significant. Clusters with total frequency more than threshold showed high specificity while clusters with lower frequency were presented with high frequency in many peptides. Examples of clusters above and below threshold for Th p5 repertoires are in Fig. 33.

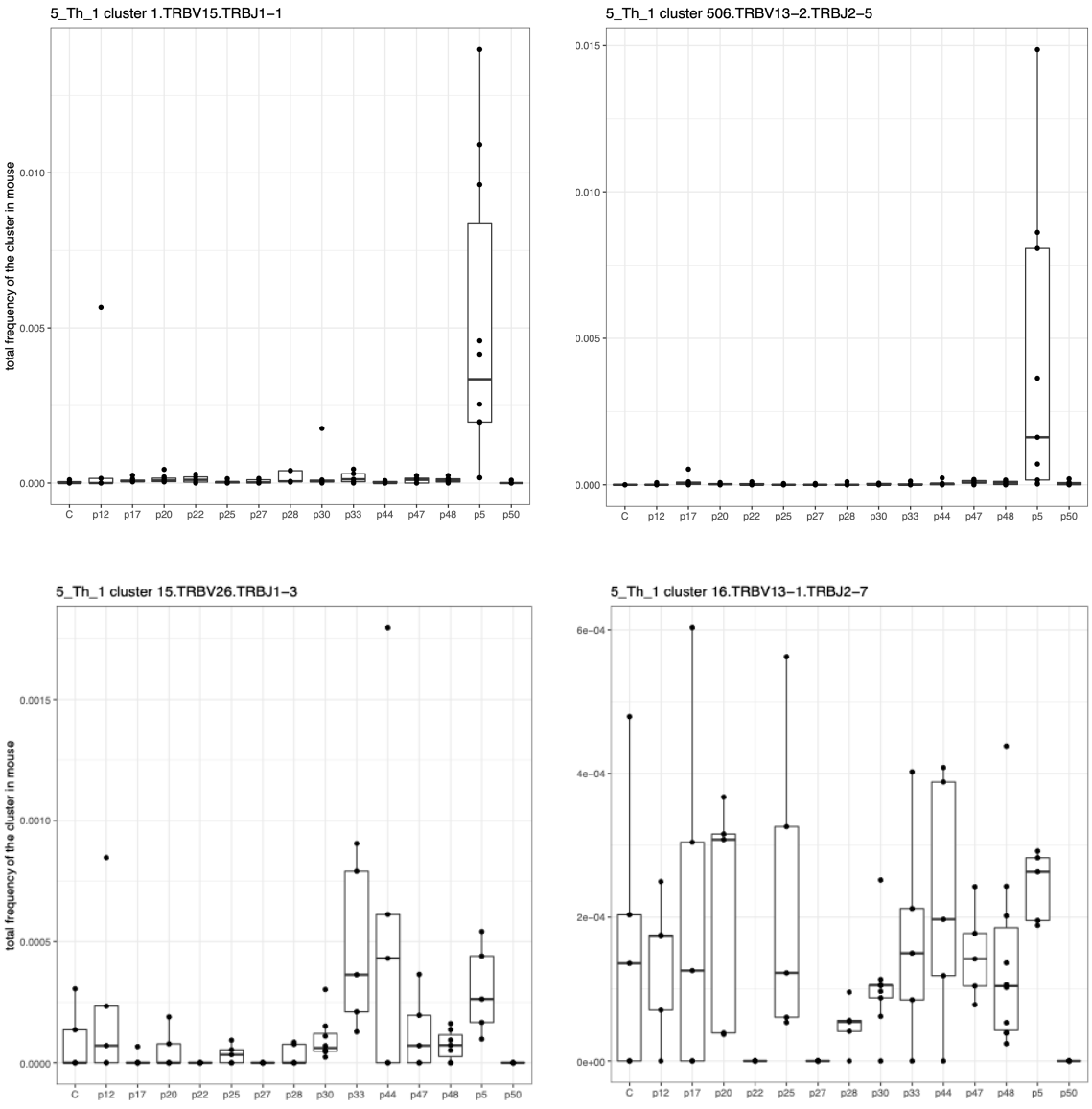


Fig. 33. Distribution of cluster frequency among experimental groups. Every dot corresponds to one mouse. Replicas were merged.

Plots with distribution of cluster frequency for all clusters above threshold are available online:

https://github.com/KseniaMIPT/B1610-melanoma-associated-murine-TCRs.

4.3.3 Melanoma-associated murine TCRs

Most identified melanoma-associated murine TCR clusters originated from Th subset (Appendix table 3). Their number ranged from 1 to 27 with a median equal to 11 (Table 3).

Table .3 Number of clusters in Th cells with total frequency more than 0.0025

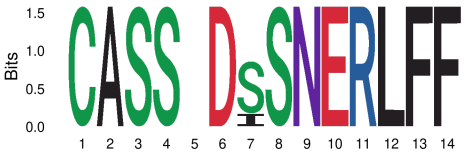
Peptide	Type	Replica	Nb of identified clusters	Peptide	Type	Replica	Nb of identified clusters
12	Th	0	9	33	Th	0	3
17	Th	1	1	44	Th	0	7
17	Th	2	11	47	Th	0	10
20	Th	1	4	48	Th	1	4
20	Th	2	16	48	Th	2	13
22	Th	0	6	5	Th	1	6
25	Th	0	27	5	Th	2	13
27	Th	0	12	50	Th	0	13
28	Th	0	16	C	Th	0	1
30	Th	1	14				
30	Th	2	17				

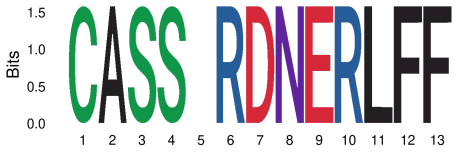
Three out of five peptides with two experimental groups had reproducible results. For the first experimental group vaccinated with p5 five out six significant clusters had matches among significant clusters from the second experimental group (Table 4). Other motifs and tables with significant clonotypes are available online:

https://github.com/KseniaMIPT/B1610-melanoma-associated-murine-TCRs.

Table 4. Motifs of significant clusters with frequency higher than 0.25% found in Th repertoires of the groups vaccinated with p5. Motifs were made by the ‘ggseqlogo’ R package.

Top position	p5 Th group 1	Top position	p5 Th group 2
1	1.TRBV15.TRBJ1-1	1	3.TRBV15.TRBJ1-1
			

2	506.TRBV13-2.TRBJ2-5	3	182.TRBV13-2.TRBJ2-5
			
3	19.TRBV4.TRBJ1-4	4	14.TRBV4.TRBJ1-4
			
4	9.TRBV15.TRBJ2-3	6	32.TRBV15.TRBJ2-3
			
6	254.TRBV19.TRBJ1-3	8	17.TRBV19.TRBJ1-3
			
5	8.TRBV13-2.TRBJ1-4	2	224.TRBV14.TRBJ2-4
			
		5	27.TRBV3.TRBJ2-7
			
		7	27.TRBV13-2.TRBJ2-7

			
		9	46.TRBV31.TRBJ2-7
			
		10	28.TRBV13-1.TRBJ1-3
			
		11	10.TRBV5.TRBJ2-5
			
		12	21.TRBV13-3.TRBJ1-4
			

5 DISCUSSION

As mentioned in the literature review, T cells are crucial players of anti-cancer immune response. Mediating via T cell receptors, cytotoxic T cells specifically recognise neo-antigens in the complex with MHC on the surface of tumour cells and kill them. CD4+ T cells pointedly guide CD8+ T cell activity. Their targeted activation leads to tumour growth inhibition and is used in immunotherapies. While murine models are widely used in cancer research, our ability to extract valuable information from murine TCR repertoires is limited. The present research identified melanoma-associated murine TCRs. These TCRs can be used as a source for functional biological assays. Moreover, they can serve as a marker of antitumor activity to profile disease state, cause of immunotherapy and other purposes.

Melanoma-associated TCRs were identified by a novel tool called TCRgrapher. Its development was the first objective of the study. It is based on the ALICE statistical model and identifies clonotypes with significantly more neighbours than expected. Neighbour is a CDR3 sequence with one amino acid mismatch or without it. Expected baseline are estimated using generation probabilities of all possible neighbours. Unlike ALICE, the developed tool operates with both human and murine models. As an option it can use any custom model. Additionally, it supports parallel computations and uses high-performance methods for calculations. Finally, the form of R library makes it easier to use and write reproducible code. TCRgrapher can be used in many areas including cancer research, autoimmunity and vaccine design and testing.

Sure, it is not the only way to identify condition-associated clonotypes from TCR repertoire without knowledge about epitope and MHC. TCRgrapher was compared with TCRdist3, GLIPH2 and EdgeR using a dataset of mice vaccinated with Sputnik V. Based on the results of the analysis, the following recommendations can be made. TCRdist3 seems to be the best choice if there is a reliable control sample, for instance, several samples before vaccination to identify responded clonotypes. EdgeR can be used if there is a control sample and a small number of identified clonotypes is satisfactory, for example, for further confirmation by functional assays. TCRgrapher is preferable in cases when there is no control sample, which happens in research quite often, especially in human studies. GLIPH2 showed poor results. This may indicate that the naive reference repertoire is not suitable as s control.

For every clonotype TCRdist3 compared the number of neighbours inside the ‘circle’ with a radius equal to 16. This radius corresponds to one or two amino acid mismatches. Its use has worked

pretty well. On the other hand, TCRgrapher compares the real number of neighbours to the expected, considering neighbour as a sequence with no more than one amino acid mismatch. A larger proportion of clonotypes found by TCRdist compared to the TCRgrapher may indicate that in the case of Sputnik vaccination responding clonotypes form large clusters of TCRs that differ by more than one amino acid.

Another important finding was that four mice is mostly likely enough to reproduce valuable results. The need to merge mice for analysis arose after the assumption that the largest clonotypes after vaccination are vaccine-specific. In some cases potentially specific large clonotypes do not have enough neighbours to be detected by TCRgrapher. It could be due to the stochastic nature of immune response.

That four mice are enough are indirectly confirmed by identified melanoma-associated clusters. For three out of five peptides with two experimental groups, motifs of identified clusters were quite similar. However, Sputnik causes a very strong immune response. It goes without saying, but the number of mice could differ for antigens with low immunogenicity.

Moreover, TCRgrapher p-values were recalculated taking into account information about the cluster's frequency. The result was mostly the same. This is consistent with the conclusion obtained in (Pogorelyy et al. 2019). Furthermore, in the thesis four generation probability models we obtained. They performed better than the OLGA standard model, available online and can be used for future research.

The study detects the evidence that frequency of the cluster formed by significant clonotypes is an important marker of cluster specificity. Because when there is no control, many found clusters are specific for recent immune challenges. As anticipated, in the case of a recent immune challenge, the specific clusters should be relatively large. The threshold used in the thesis is most likely not the same for all cases. In this work, it rather reflects the level of the immune response provoked by the adjuvant.

The melanoma-associated clonotypes determined in the study are highly likely not public, as they have passed several stages of verification. Thanks to the analysis of several mice, the common immune response pattern was found. As it was revealed earlier five mice per sample is enough to get valuable results.

Interestingly, CD8+ T cells collected from mice vaccinated with melanoma peptides had a high rate of shared sequence compared to CD4+ T cell subsets. This may be explained by the fact that

MHC-II restricted epitopes are more conservative than MHC-I restricted ones. So, CD8+ T cells have more different possibilities to recognise the same antigen and be cross-reactive.

Identified melanoma associate TCRs were mainly from T-helper subset. This is in line with initial expectations, since peptides were selected to cause CD4+ immune response. Identified melanoma associated TCRs formed a database that is available online. Database of melanoma-specific murine TCRs will be used by me and my colleagues in the future study about anti-CTLA4 therapy.

6 CONCLUSIONS

1. A method, TCRgrapher, for the identification of condition-associated TCRs in human and murine repertoires was developed. TCRgrapher does not require control samples and is high-performance compared with ALICE, GLIPH2 and TCRdist3 especially on large datasets.
2. The size of the sample was determined. TCR repertoire from four mice was sufficient to obtain the majority of antigen-specific TCR beta variants in the given experimental setting.
3. Generation probabilistic models were inferred for the repertoires obtained from PBMC and sorted CD8+, Th and Tregs cell subsets. The models trained on the available data showed better results in comparison with OLGA standard model: they recorded a statistically significant response to the vaccination with Sputnik and decreased false positive rate of identified melanoma-associated clonotypes.
4. Cluster searching approaches (TCRgrapher, TCRdist3) provided much more information than just searching for expanded clonotypes (EdgeR). TCRdist3 seemed to be the best choice if the control group presents.
5. The importance of cluster frequency in determining its specificity in the context of this work was shown. Total frequency threshold for the cluster was adjusted. CDR3 motifs of the TCR clusters specific to B16F10 melanoma antigens above this threshold were identified.
6. Significantly overrepresented TCR clusters in the repertoire of CD8+, Treg and Th cell subsets from the mice vaccinated with melanoma peptides were identified. The largest clusters were in Th, as expected. The mean number of clusters per vaccinated group was 11. Melanoma-associated TCR database with identified clusters is available online https://github.com/KseniaMIPT/B1610-melanoma-associated-murine-TCRs
7. The database will be of interest to melanoma research in a murine model. Me and my colleagues will use it in future research about anti-CTL4 therapy. The methods used for identification of condition-associated TCRs may be applied to other areas: cancer research, targeted autoimmunity treatment, vaccine design and others.

7 AUTHOR CONTRIBUTION

All analysis starting from TCR repertoire extraction was made by the author unless otherwise specified in acknowledgments.

8 ACKNOWLEDGEMENTS

I express sincere thanks to my colleagues for experiment design, library preparation and supervision.

George V. Sharonov — carried out the experiment

Irina. A. Shagina — library preparation and sequencing

Diana V. Yuzhakova — carried out the experiment

Anna V. Izosimova — carried out the experiment

Pavel V. Shelyakin — IGoR generation probability model inference

Olga V. Britanova — supervision

Dmitry M. Chudakov — supervision

9 ABBREVIATIONS

ALICE — Antigen-specific Lymphocyte Identification by Clustering of Expanded sequences
APC — antigen presenting cell
CD4/8 — cluster of differentiation 4/8
cDNA — complementary deoxyribonucleic acid
CDR — complementarity-determining region
DN — double negative
DNN — Deep Neural Networks
DP — double positive
EDTA —ethylenediaminetetraacetic acid
ELISPOT — Enzyme-Linked ImmunoSpot
FDR — false discovery rate
Fig. — figure

IGoR — Inference and Generation of Repertoires
IMGT — The International Immunogenetics Information System
MHC — major histocompatibility complex
NGS — next-generation sequencing
OLGA — Optimized Likelihood estimate of immunoGlobulin Amino-acid sequences
p5 — peptide five
PBMC — peripheral blood mononuclear cell
PCA — principal component analysis
pMHC — peptide-major histocompatibility complex
RNA —ribonucleic acid
SP — single positive
TCR — T cell receptor
Th — T-helper cell

TMM — The Trimmed Mean of the M-values
TPR — true positive rate
TRA — T cell receptor alpha chain
TRB — T cell receptor beta chain
Treg — regulatory T cell
UMI — unique molecular identifier
VAE — Variational Autoencoder

10 REFERENCES

Bagaev, Dmitry V., Renske M. A. Vroomans, Jerome Samir, Ulrik Stervbo, Cristina Rius, Garry Dolton, Alexander Greenshields-Watson, et al. 2020. “VDJdb in 2019: Database Extension, New Analysis Infrastructure and a T-Cell Receptor Motif Compendium.” *Nucleic Acids Research* 48 (D1): D1057–62.

Bolotin, Dmitriy A., Stanislav Poslavsky, Igor Mitrophanov, Mikhail Shugay, Ilgar Z. Mamedov, Ekaterina V. Putintseva, and Dmitriy M. Chudakov. 2015. “MiXCR: Software for Comprehensive Adaptive Immunity Profiling.” *Nature Methods* 12 (5): 380–81.

Castle, John C., Sebastian Kreiter, Jan Diekmann, Martin Löwer, Niels van de Roemer, Jos de Graaf, Abderraouf Selmi, et al. 2012. “Exploiting the Mutanome for Tumor Vaccination.” *Cancer Research* 72 (5): 1081–91.

Chronister, William D., Austin Crinklaw, Swapnil Mahajan, Randi Vita, Zeynep Koşaloğlu-Yalçın, Zhen Yan, Jason A. Greenbaum, et al. 2021. “TCRMatch: Predicting T-Cell Receptor Specificity Based on Sequence Similarity to Previously Characterized Receptors.” *Frontiers in Immunology* 12 (March): 640725.

Dash, Pradyot, Andrew J. Fiore-Gartland, Tomer Hertz, George C. Wang, Shalini Sharma, Aisha Souquette, Jeremy Chase Crawford, et al. 2017. “Quantifiable Predictive Features Define Epitope-Specific T Cell Receptor Repertoires.” *Nature* 547 (7661): 89–93.

Davidsen, Kristian, Branden J. Olson, William S. DeWitt, Jean Feng, Elias Harkins, Philip Bradley, and Frederick A. Matsen. 2019. “Deep Generative Models for T Cell Receptor Protein Sequences.” *eLife*. https://doi.org/10.7554/elife.46935.

Davis, Mark M., John D. Altman, and Evan W. Newell. 2011. “Interrogating the Repertoire: Broadening the Scope of peptide–MHC Multimer Analysis.” *Nature Reviews Immunology*. https://doi.org/10.1038/nri3020.

Davis, Mark M., Cristina M. Tato, and David Furman. 2017. “Systems Immunology: Just Getting Started.” *Nature Immunology* 18 (7): 725–32.

Dupic, Thomas, Quentin Marcou, Aleksandra M. Walczak, and Thierry Mora. 2019. “Genesis of the αβ T-Cell Receptor.” *PLoS Computational Biology* 15 (3): e1006874.

Egorov, Evgeny S., Sofya A. Kasatskaya, Vasiliy N. Zubov, Mark Izraelson, Tatiana O. Nakonechnaya, Dmitriy B. Staroverov, Andrea Angius, et al. 2018. “The Changing Landscape of Naive T Cell Receptor Repertoire With Human Aging.” *Frontiers in Immunology* 9 (July): 1618.

Elhanati, Yuval, Anand Murugan, Curtis G. Callan Jr, Thierry Mora, and Aleksandra M. Walczak. 2014. “Quantifying Selection in Immune Receptor Repertoires.” *Proceedings of the National Academy of Sciences of the United States of America* 111 (27): 9875–80.

Elhanati, Yuval, Zachary Sethna, Curtis G. Callan Jr, Thierry Mora, and Aleksandra M. Walczak. 2018. “Predicting the Spectrum of TCR Repertoire Sharing with a Data-Driven Model of Recombination.” *Immunological Reviews* 284 (1): 167–79.

Garcia, K. Christopher, K. Christopher Garcia, Massimo Degano, Robyn L. Stanfield, Anders Brunmark, Michael R. Jackson, Per A. Peterson, Luc Teyton, and Ian A. Wilson. 1996. “An αβ T Cell Receptor Structure at 2.5 Å and Its Orientation in the TCR-MHC Complex.” *Science*. https://doi.org/10.1126/science.274.5285.209.

Gielis, Sofie, Pieter Moris, Wout Bittremieux, Nicolas De Neuter, Benson Ogunjimi, Kris Laukens, and Pieter Meysman. 2019. “Detection of Enriched T Cell Epitope Specificity in Full T Cell Receptor Sequence Repertoires.” *Frontiers in Immunology* 10 (November): 2820.

Glanville, Jacob, Huang Huang, Allison Nau, Olivia Hatton, Lisa E. Wagar, Florian Rubelt, Xuhuai Ji, et al. 2017. “Identifying Specificity Groups in the T Cell Receptor Repertoire.” *Nature* 547 (7661): 94–98.

Greiff, Victor, Cédric R. Weber, Johannes Palme, Ulrich Bodenhofer, Enkelejda Miho, Ulrike Menzel, and Sai T. Reddy. 2017. “Learning the High-Dimensional Immunogenomic Features That Predict Public and Private Antibody

Repertoires." *Journal of Immunology* 199 (8): 2985–97.

Hanahan, Douglas, and Robert A. Weinberg. 2011. "Hallmarks of Cancer: The Next Generation." *Cell*. https://doi.org/10.1016/j.cell.2011.02.013.

"How Many Different Clonotypes Do Immune Repertoires Contain?" 2019. *Current Opinion in Systems Biology* 18 (December): 104–10.

Huang, Huang, Chunlin Wang, Florian Rubelt, Thomas J. Scriba, and Mark M. Davis. 2020. "Analyzing the Mycobacterium Tuberculosis Immune Response by T-Cell Receptor Clustering with GLIPH2 and Genome-Wide Antigen Screening." *Nature Biotechnology* 38 (10): 1194–1202.

"IMGT Unique Numbering for Immunoglobulin and T Cell Receptor Variable Domains and Ig Superfamily V-like Domains." 2003. *Developmental and Comparative Immunology* 27 (1): 55–77.

Isacchini, Giulio, Zachary Sethna, Yuval Elhanati, Armita Nourmohammad, Aleksandra M. Walczak, and Thierry Mora. 2020. "Generative Models of T-Cell Receptor Sequences." *Physical Review. E* 101 (6-1): 062414.

Isacchini, Giulio, Aleksandra M. Walczak, Thierry Mora, and Armita Nourmohammad. 2021. "Deep Generative Selection Models of T and B Cell Receptor Repertoires with soNNia." *Proceedings of the National Academy of Sciences of the United States of America* 118 (14). https://doi.org/10.1073/pnas.2023141118.

Kasatskaya, Sofya A., Kristin Ladell, Evgeniy S. Egorov, Kelly L. Miners, Alexey N. Davydov, Maria Metsger, Dmitry B. Staroverov, et al. 2020. "Functionally Specialized Human CD4 T-Cell Subsets Express Physicochemically Distinct TCRs." *eLife* 9 (December). https://doi.org/10.7554/eLife.57063.

Klein, Ludger, Bruno Kyewski, Paul M. Allen, and Kristin A. Hogquist. 2014. "Positive and Negative Selection of the T Cell Repertoire: What Thymocytes See (and Don't See)." *Nature Reviews. Immunology* 14 (6): 377–91.

Kreiter, Sebastian, Mathias Vormehr, Niels van de Roemer, Mustafa Diken, Martin Löwer, Jan Diekmann, Sebastian Boegel, et al. 2015. "Mutant MHC Class II Epitopes Drive Therapeutic Immune Responses to Cancer." *Nature* 520 (7549): 692–96.

Kuryk, Lukasz, Laura Bertinato, Monika Staniszewska, Katarzyna Pancer, Magdalena Wieczorek, Stefano Salmaso, Paolo Caliceti, and Mariangela Garofalo. 2020. "From Conventional Therapies to Immunotherapy: Melanoma Treatment in Review." *Cancers* 12 (10). https://doi.org/10.3390/cancers12103057.

Love, Michael I., Wolfgang Huber, and Simon Anders. 2014. "Moderated Estimation of Fold Change and Dispersion for RNA-Seq Data with DESeq2." *Genome Biology* 15 (12): 550.

Marcou, Quentin, Thierry Mora, and Aleksandra M. Walczak. 2018. "High-Throughput Immune Repertoire Analysis with IGoR." *Nature Communications* 9 (1): 561.

———. n.d. "IGoR: A Tool for High-Throughput Immune Repertoire Analysis." https://doi.org/ 10.1101/141143.

Marrack, Philippa, James P. Scott-Browne, Shaodong Dai, Laurent Gapin, and John W. Kappler. 2008. "Evolutionarily Conserved Amino Acids That Control TCR-MHC Interaction." *Annual Review of Immunology* 26: 171–203.

Mayer-Blackwell, Koshlan, Stefan Schattgen, Liel Cohen-Lavi, Jeremy C. Crawford, Aisha Souquette, Jessica A. Gaevert, Tomer Hertz, Paul G. Thomas, Philip Bradley, and Andrew Fiore-Gartland. 2021. "TCR Meta-Clonotypes for Biomarker Discovery with tcrdist3 Enabled Identification of Public, HLA-Restricted Clusters of SARS-CoV-2 TCRs." *eLife*. https://doi.org/10.7554/elife.68605.

Meysman, Pieter, Nicolas De Neuter, Sofie Gielis, Danh Bui Thi, Benson Ogunjimi, and Kris Laukens. 2019. "On the Viability of Unsupervised T-Cell Receptor Sequence Clustering for Epitope Preference." *Bioinformatics* 35 (9): 1461–68.

Miles, John J., Daniel C. Douek, and David A. Price. 2011. "Bias in the αβ T-Cell Repertoire: Implications for Disease Pathogenesis and Vaccination." *Immunology and Cell Biology* 89 (3): 375–87.

Murphy, Kenneth M., Paul Travers, Charles Janeway, and Mark Walport. 2007. *Janeway's Immunobiology, International*

Student Edition. Garland Publishing.

Murugan, Anand, Thierry Mora, Aleksandra M. Walczak, and Curtis G. Callan Jr. 2012. "Statistical Inference of the Generation Probability of T-Cell Receptors from Sequence Repertoires." *Proceedings of the National Academy of Sciences of the United States of America* 109 (40): 16161–66.

Ndifon, Wilfred, Hilah Gal, Eric Shifrut, Rina Aharoni, Nissan Yissachar, Nir Waysbort, Shlomit Reich-Zeliger, Ruth Arnon, and Nir Friedman. 2012. "Chromatin Conformation Governs T-Cell Receptor Jβ Gene Segment Usage." *Proceedings of the National Academy of Sciences of the United States of America* 109 (39): 15865–70.

Patton, E. Elizabeth, Kristen L. Mueller, David J. Adams, Niroshana Anandasabapathy, Andrew E. Aplin, Corine Bertolotto, Marcus Bosenberg, et al. 2021. "Melanoma Models for the next Generation of Therapies." *Cancer Cell* 39 (5): 610–31.

Pogorelyy, Mikhail V., Anastasia A. Minervina, Dmitriy M. Chudakov, Ilgar Z. Mamedov, Yuri B. Lebedev, Thierry Mora, and Aleksandra M. Walczak. 2018. "Method for Identification of Condition-Associated Public Antigen Receptor Sequences." *eLife* 7 (March). https://doi.org/10.7554/eLife.33050.

Pogorelyy, Mikhail V., Anastasia A. Minervina, Mikhail Shugay, Dmitriy M. Chudakov, Yuri B. Lebedev, Thierry Mora, and Aleksandra M. Walczak. 2019. "Detecting T Cell Receptors Involved in Immune Responses from Single Repertoire Snapshots." *PLoS Biology* 17 (6): e3000314.

Pogorelyy, Mikhail V., Anastasia A. Minervina, Maximilian Puelma Touzel, Anastasiia L. Sycheva, Ekaterina A. Komech, Elena I. Kovalenko, Galina G. Karganova, et al. 2018. "Precise Tracking of Vaccine-Responding T Cell Clones Reveals Convergent and Personalized Response in Identical Twins." *Proceedings of the National Academy of Sciences of the United States of America* 115 (50): 12704–9.

Pogorelyy, Mikhail V., and Mikhail Shugay. 2019. "A Framework for Annotation of Antigen Specificities in High-Throughput T-Cell Repertoire Sequencing Studies." *bioRxiv*. https://doi.org/10.1101/676239.

Puelma Touzel, Maximilian, Aleksandra M. Walczak, and Thierry Mora. 2020. "Inferring the Immune Response from Repertoire Sequencing." *PLoS Computational Biology* 16 (4): e1007873.

Qi, Qian, Mary M. Cavanagh, Sabine Le Saux, Hong NamKoong, Chulwoo Kim, Emerson Turgano, Yi Liu, et al. 2016. "Diversification of the Antigen-Specific T Cell Receptor Repertoire after Varicella Zoster Vaccination." *Science Translational Medicine* 8 (332): 332ra46.

Qi, Qian, Yi Liu, Yong Cheng, Jacob Glanville, David Zhang, Ji-Yeun Lee, Richard A. Olshen, Cornelia M. Weyand, Scott D. Boyd, and Jörg J. Goronzy. 2014. "Diversity and Clonal Selection in the Human T-Cell Repertoire." *Proceedings of the National Academy of Sciences*. https://doi.org/10.1073/pnas.1409155111.

Raskov, Hans, Adile Orhan, Jan Pravsgaard Christensen, and Ismail Gögenur. 2020. "Cytotoxic CD8+ T Cells in Cancer and Cancer Immunotherapy." *British Journal of Cancer* 124 (2): 359–67.

Robinson, Mark D., Davis J. McCarthy, and Gordon K. Smyth. 2010. "edgeR: A Bioconductor Package for Differential Expression Analysis of Digital Gene Expression Data." *Bioinformatics* 26 (1): 139–40.

Rubelt, Florian, Christopher R. Bolen, Helen M. McGuire, Jason A. Vander Heiden, Daniel Gadala-Maria, Mikhail Levin, Ghia M. Euskirchen, et al. 2016. "Individual Heritable Differences Result in Unique Cell Lymphocyte Receptor Repertoires of Naïve and Antigen-Experienced Cells." *Nature Communications* 7 (March): 11112.

Russell, Magdalena L., Aisha Souquette, David M. Levine, Stefan A. Schattgen, E. Kaitlynn Allen, Guillermina Kuan, Noah Simon, et al. 2022. "Combining Genotypes and T Cell Receptor Distributions to Infer Genetic Loci Determining V(D)J Recombination Probabilities." *eLife*. https://doi.org/10.7554/elife.73475.

Sebzda, E., S. Mariathasan, T. Ohteki, R. Jones, M. F. Bachmann, and P. S. Ohashi. 1999. "Selection of the T Cell Repertoire." *Annual Review of Immunology* 17: 829–74.

Sethna, Zachary, Yuval Elhanati, Curtis G. Callan, Aleksandra M. Walczak, and Thierry Mora. 2019. "OLGA: Fast Computation of Generation Probabilities of B- and T-Cell Receptor Amino Acid Sequences and Motifs."

Bioinformatics 35 (17): 2974–81.

Sethna, Zachary, Giulio Isacchini, Thomas Dupic, Thierry Mora, Aleksandra M. Walczak, and Yuval Elhanati. 2020. "Population Variability in the Generation and Selection of T-Cell Repertoires." *PLoS Computational Biology* 16 (12): e1008394.

Shugay, Mikhail, Dmitriy V. Bagaev, Maria A. Turchaninova, Dmitriy A. Bolotin, Olga V. Britanova, Ekaterina V. Putintseva, Mikhail V. Pogorelyy, et al. 2015. "VDJtools: Unifying Post-Analysis of T Cell Receptor Repertoires." *PLoS Computational Biology* 11 (11): e1004503.

Shugay, Mikhail, Dmitriy A. Bolotin, Ekaterina V. Putintseva, Mikhail V. Pogorelyy, Ilgar Z. Mamedov, and Dmitriy M. Chudakov. 2013. "Huge Overlap of Individual TCR Beta Repertoires." *Frontiers in Immunology* 0. https://doi.org/10.3389/fimmu.2013.00466.

Shugay, Mikhail, Olga V. Britanova, Ekaterina M. Merzlyak, Maria A. Turchaninova, Ilgar Z. Mamedov, Timur R. Tuganbaev, Dmitriy A. Bolotin, et al. 2014. "Towards Error-Free Profiling of Immune Repertoires." *Nature Methods* 11 (6): 653–55.

Sim, B. C., L. Zerva, M. I. Greene, and N. R. Gascoigne. 1996. "Control of MHC Restriction by TCR Valpha CDR1 and CDR2." *Science* 273 (5277): 963–66.

Tay, Rong En, Emma K. Richardson, and Han Chong Toh. 2021. "Revisiting the Role of CD4 T Cells in Cancer Immunotherapy-New Insights into Old Paradigms." *Cancer Gene Therapy* 28 (1-2): 5–17.

Venturi, Vanessa, Katherine Kedzierska, David A. Price, Peter C. Doherty, Daniel C. Douek, Stephen J. Turner, and Miles P. Davenport. 2006. "Sharing of T Cell Receptors in Antigen-Specific Responses Is Driven by Convergent Recombination." *Proceedings of the National Academy of Sciences of the United States of America* 103 (49): 18691–96.

"Website." n.d. https://doi.org/10.1016/S1074-7613(00)80250-2.

Wing, James B., Atsushi Tanaka, and Shimon Sakaguchi. 2019. "Human FOXP3 Regulatory T Cell Heterogeneity and Function in Autoimmunity and Cancer." *Immunity* 50 (2): 302–16.

Wood, Peter John. 2006. *Understanding Immunology*. Pearson Education.

Zvyagin, Ivan V., Mikhail V. Pogorelyy, Marina E. Ivanova, Ekaterina A. Komech, Mikhail Shugay, Dmitry A. Bolotin, Andrey A. Shelenkov, et al. 2014. "Distinctive Properties of Identical Twins' TCR Repertoires Revealed by High-Throughput Sequencing." *Proceedings of the National Academy of Sciences*. https://doi.org/10.1073/pnas.1319389111.

11 APPENDIX

Appendix Table 1. Basic statistics for unnormalised tables of functional clonotypes of mice vaccinated with Sputnik V

Mouse	Group	Time point	Number of UMI	Number of clonotypes	Mouse	Group	Time point	Number of UMI	Number of clonotypes
1	control	1	20911	17309	8	Sputnik	1	31456	25639
1	control	2	40388	30851	8	Sputnik	2	29322	21089
1	control	3	20211	15920	8	Sputnik	3	10729	7052
1	control	4	7626	7122	8	Sputnik	4	11288	9395
1	control	5	30750	26688	8	Sputnik	5	31863	25286
2	control	1	7212	6437	9	Sputnik	1	44332	36797
2	control	2	6091	5374	9	Sputnik	2	39706	31150
2	control	3	28025	21176	9	Sputnik	3	62738	40411
2	control	4	6044	5568	9	Sputnik	4	14332	12009
2	control	5	39043	31547	9	Sputnik	5	75565	57194
3	control	1	14998	12911	10	Sputnik	1	24183	17841
3	control	2	12618	11415	10	Sputnik	2	35879	27220
3	control	3	21588	18375	10	Sputnik	3	22731	17210
3	control	4	8998	8354	10	Sputnik	4	9282	7689
3	control	5	23803	20593	10	Sputnik	5	26988	21454
4	control	1	11407	9795	11	Sputnik	1	7062	6373
4	control	2	29654	23640	11	Sputnik	2	49784	40052
4	control	3	26275	22254	11	Sputnik	3	47331	35637
4	control	4	20341	16066	11	Sputnik	4	7875	6459
4	control	5	14328	12628	11	Sputnik	5	15648	12797
5	control	1	12732	11228	12	Sputnik	1	17139	15135
5	control	2	4867	4426	12	Sputnik	2	5357	4932
5	control	3	7235	6333	12	Sputnik	3	10985	8638
5	control	4	22393	19837	12	Sputnik	4	21958	16890
5	control	5	12203	10799	12	Sputnik	5	12877	10872
6	control	1	23628	20008	13	Sputnik	1	40367	33798
6	control	2	8285	7144	13	Sputnik	2	16331	14280
6	control	3	24368	21183	13	Sputnik	3	11436	8660
6	control	4	48872	40281	13	Sputnik	4	19064	14708
6	control	5	33743	27417	13	Sputnik	5	29227	23595

7	control	1	27353	23232	14	Sputnik	1	28811	24688
7	control	2	28066	24360	14	Sputnik	2	12440	10926
7	control	3	18390	15529	14	Sputnik	3	20347	15992
7	control	4	51679	41923	14	Sputnik	4	17802	13422
7	control	5	39415	33474	14	Sputnik	5	40419	31580
					15	Sputnik	1	47214	40183
					15	Sputnik	2	23144	20207
					15	Sputnik	3	45461	37007
					15	Sputnik	4	24989	20556
					15	Sputnik	5	13081	11057

Appendix Table 2. Basic statistics for unnormalised tables of functional clonotypes of mice immunised with B16F10 melanoma peptides

Peptide	**Mouse**	**Type**	**Number of UMI**	**Number of clonotypes**	**Peptide**	**Mouse**	**Type**	**Number of UMI**	**Number of clonotypes**
Control	m1	CD8	41315	37850	p30	m4.1	CD8	20129	18467
Control	m2	CD8	34541	32520	p30	m4.2	CD8	26541	24753
Control	m3	CD8	42419	38745	p30	m5.1	CD8	25240	23783
Control	m1	Th	18784	16556	p30	m5.2	CD8	66236	58382
Control	m2	Th	22077	20036	p30	m1.1	Th	20712	16750
Control	m3	Th	9834	9110	p30	m1.2	Th	43802	33508
Control	m1	Treg	31105	17101	p30	m2.1	Th	28411	22233
Control	m2	Treg	31253	17954	p30	m2.2	Th	41245	36354
Control	m3	Treg	35506	19873	p30	m3.1	Th	19848	14395
p12	m1	CD8	41677	38802	p30	m3.2	Th	45443	35464
p12	m2	CD8	30062	28377	p30	m4.1	Th	32122	21980
p12	m3	CD8	30469	27846	p30	m4.2	Th	85474	68201
p12	m4	CD8	31285	28198	p30	m5.1	Th	26385	22328
p12	m5	CD8	6586	5936	p30	m5.2	Th	37917	32239
p12	m1	Th	20017	17568	p30	m1.1	Treg	6678	4258
p12	m2	Th	17272	15868	p30	m1.2	Treg	19636	10563
p12	m3	Th	17113	13893	p30	m2.1	Treg	1763	1432
p12	m4	Th	14101	11184	p30	m2.2	Treg	17435	9349
p12	m5	Th	4723	4064	p30	m3.1	Treg	3507	2586
p12	m1	Treg	19939	11386	p30	m3.2	Treg	27522	14527
p12	m2	Treg	25133	15793	p30	m4.1	Treg	3908	2852

p12	m3	Treg	22433	9892	p30	m4.2	Treg	17574	8685
p12	m4	Treg	10804	6201	p30	m5.1	Treg	3460	2547
p12	m5	Treg	1221	863	p30	m5.2	Treg	8000	5791
p17	m1	CD8	22357	20845	p33	m1	CD8	30504	27125
p17	m1	CD8	65050	58247	p33	m2	CD8	26154	24465
p17	m2	CD8	17211	15953	p33	m3	CD8	25600	22952
p17	m2	CD8	57297	50374	p33	m4	CD8	33467	30678
p17	m3	CD8	19073	17797	p33	m5	CD8	32150	29081
p17	m3	CD8	34450	31994	p33	m1	Th	9943	8107
p17	m4	CD8	10490	10054	p33	m2	Th	32982	27581
p17	m5	CD8	21065	19768	p33	m3	Th	16453	13939
p17	m1.1	Th	32860	28194	p33	m4	Th	23465	20411
p17	m1.2	Th	52219	38894	p33	m5	Th	33306	26955
p17	m2.1	Th	29843	24860	p33	m1	Treg	5580	3841
p17	m2.2	Th	40657	31991	p33	m2	Treg	36116	19193
p17	m3.1	Th	23842	21317	p33	m3	Treg	20485	10286
p17	m3.2	Th	34900	28468	p33	m4	Treg	33188	17452
p17	m4.1	Th	28031	23575	p33	m5	Treg	18332	11056
p17	m5.2	Th	46527	37331	p44	m1	CD8	27025	25243
p17	m1.1	Treg	15389	8619	p44	m2	CD8	31814	29810
p17	m1.2	Treg	61321	31453	p44	m3	CD8	33859	31527
p17	m2.1	Treg	13410	8254	p44	m4	CD8	33062	30742
p17	m2.2	Treg	39937	20483	p44	m5	CD8	20443	19540
p17	m3.1	Treg	12548	8082	p44	m1	Th	24493	19284
p17	m3.2	Treg	35021	20648	p44	m2	Th	20295	17234
p17	m4.1	Treg	6125	4554	p44	m3	Th	23191	18496
p17	m5.1	Treg	13547	8569	p44	m4	Th	25230	21710
p20	m1.1	CD8	31914	29672	p44	m5	Th	26135	21179
p20	m1.2	CD8	82804	71186	p44	m1	Treg	36139	17596
p20	m2.1	CD8	25726	23907	p44	m2	Treg	27331	14805
p20	m2.2	CD8	49454	44048	p44	m3	Treg	25234	12981
p20	m3.1	CD8	12679	12014	p44	m4	Treg	18115	10882
p20	m3.2	CD8	28248	24802	p44	m5	Treg	17532	10745
p20	m4.1	CD8	17203	16223	p47	m1	CD8	50120	43382
p20	m4.2	CD8	59530	52741	p47	m2	CD8	36674	33508
p20	m5.1	CD8	15534	14509	p47	m3	CD8	34454	30364
p20	m5.2	CD8	48547	44196	p47	m4	CD8	27225	24650

p20	m1.1	Th	47011	33914	p47	m5	CD8	42872	32854
p20	m1.2	Th	54165	44149	p47	m1	Th	32960	24164
p20	m2.1	Th	27229	23751	p47	m2	Th	25515	20686
p20	m2.2	Th	48797	37492	p47	m3	Th	11255	8934
p20	m3.1	Th	25605	20534	p47	m4	Th	19172	14608
p20	m3.2	Th	46756	35201	p47	m5	Th	28146	21499
p20	m4.1	Th	15832	13198	p47	m1	Treg	33860	16526
p20	m4.2	Th	35046	28999	p47	m2	Treg	9471	5868
p20	m5.1	Th	37554	28645	p47	m3	Treg	27318	13805
p20	m5.2	Th	35700	29765	p47	m4	Treg	33092	17815
p20	m1.1	Treg	14663	8789	p47	m5	Treg	22101	11954
p20	m1.2	Treg	61065	31254	p48	m1.1	CD8	18662	17506
p20	m2.1	Treg	15455	9339	p48	m1.2	CD8	54100	46356
p20	m2.2	Treg	40748	20069	p48	m2.1	CD8	36708	33614
p20	m3.1	Treg	14803	9497	p48	m2.2	CD8	59087	50288
p20	m3.2	Treg	37228	20039	p48	m3.1	CD8	33260	29567
p20	m4.1	Treg	5142	3987	p48	m3.2	CD8	36477	33591
p20	m4.2	Treg	32701	19578	p48	m4.1	CD8	27349	25449
p20	m5.1	Treg	10158	6849	p48	m4.2	CD8	53549	47269
p20	m5.2	Treg	54165	25788	p48	m5.1	CD8	25674	24361
p22	m1	CD8	25365	23648	p48	m5.2	CD8	40514	36387
p22	m2	CD8	24806	22607	p48	m1.1	Th	12334	10478
p22	m3	CD8	16925	15798	p48	m1.2	Th	31958	25243
p22	m4	CD8	23478	21881	p48	m2.1	Th	24765	21157
p22	m5	CD8	21666	19755	p48	m2.2	Th	51064	39059
p22	m1	Th	30285	23033	p48	m3.1	Th	18690	16226
p22	m2	Th	25439	19205	p48	m3.2	Th	39081	31867
p22	m3	Th	48149	32532	p48	m4.1	Th	18822	15785
p22	m4	Th	42423	30153	p48	m4.2	Th	51180	41463
p22	m5	Th	31373	23021	p48	m5.1	Th	29306	25057
p22	m1	Treg	30963	15568	p48	m5.2	Th	41416	32058
p22	m2	Treg	21881	11376	p48	m1.1	Treg	21043	12666
p22	m3	Treg	20903	11120	p48	m1.2	Treg	40476	19323
p22	m4	Treg	15881	8540	p48	m2.1	Treg	25550	14523
p22	m5	Treg	9491	5688	p48	m2.2	Treg	40706	21106
p25	m1	CD8	32543	29877	p48	m3.1	Treg	24747	13655
p25	m2	CD8	39416	35935	p48	m3.2	Treg	41160	21316

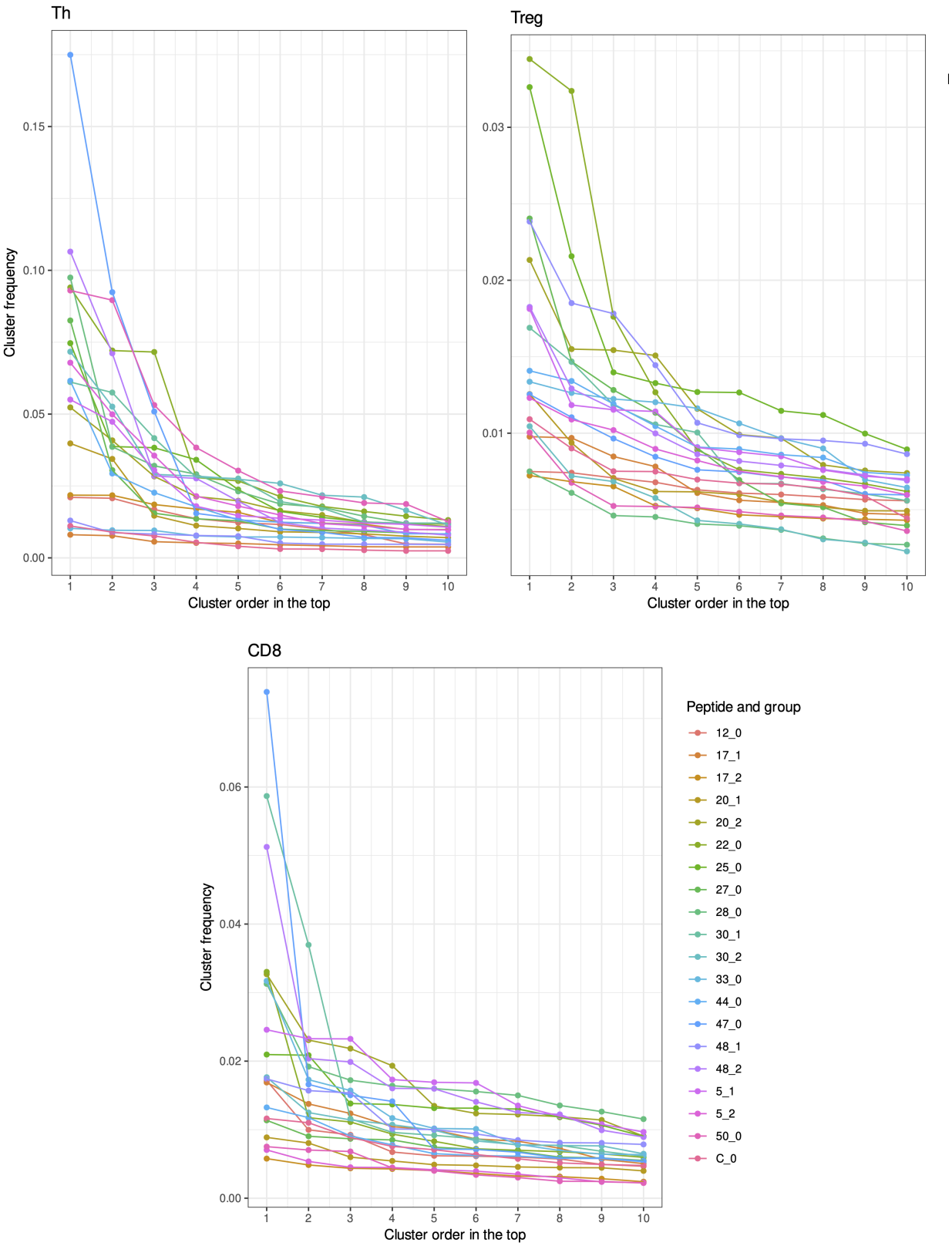
p25	m3	CD8	28505	24076	p48	m4.1	Treg	23800	12844
p25	m4	CD8	42283	38388	p48	m4.2	Treg	38984	18243
p25	m5	CD8	42584	38911	p48	m5.1	Treg	9244	6018
p25	m1	Th	18580	15241	p48	m5.2	Treg	40552	20259
p25	m2	Th	21468	17699	p5	m1.1	CD8	12958	12232
p25	m3	Th	24454	16726	p5	m1.2	CD8	71796	61626
p25	m4	Th	30228	23684	p5	m2.1	CD8	17968	15563
p25	m5	Th	49198	38714	p5	m2.2	CD8	69709	59993
p25	m1	Treg	36133	18630	p5	m3.1	CD8	19230	18052
p25	m2	Treg	46987	23304	p5	m3.2	CD8	53255	47770
p25	m3	Treg	28441	12010	p5	m4.1	CD8	18557	17322
p25	m4	Treg	54019	21761	p5	m4.2	CD8	39686	36551
p25	m5	Treg	21024	11448	p5	m5.1	CD8	15672	13659
p27	m1	CD8	22214	19998	p5	m5.2	CD8	73491	64616
p27	m2	CD8	29744	26290	p5	m1.1	Th	18594	13889
p27	m3	CD8	28942	26046	p5	m1.2	Th	42431	34284
p27	m4	CD8	23422	21113	p5	m2.1	Th	36425	20797
p27	m5	CD8	26080	22943	p5	m2.2	Th	47665	37784
p27	m1	Th	51337	36055	p5	m3.1	Th	23971	20278
p27	m2	Th	47952	38991	p5	m3.2	Th	40763	26554
p27	m3	Th	47155	35184	p5	m4.1	Th	36903	25648
p27	m4	Th	44123	34650	p5	m4.2	Th	40930	32119
p27	m5	Th	37700	30405	p5	m5.1	Th	7602	6605
p27	m1	Treg	21366	11917	p5	m5.2	Th	37357	26057
p27	m2	Treg	14839	9037	p5	m1.1	Treg	21729	10443
p27	m3	Treg	10369	6333	p5	m1.1	Treg	33682	17640
p27	m4	Treg	18859	10233	p5	m2.1	Treg	20612	10191
p27	m5	Treg	8379	5552	p5	m2.2	Treg	28552	16172
p28	m1	CD8	54538	45814	p5	m3.1	Treg	35411	15647
p28	m2	CD8	46397	38849	p5	m3.2	Treg	30826	17340
p28	m3	CD8	41525	35710	p5	m4.1	Treg	18963	8694
p28	m4	CD8	58987	52740	p5	m4.2	Treg	40987	21288
p28	m5	CD8	68029	59114	p5	m5.1	Treg	18844	8437
p28	m1	Th	39208	27503	p5	m5.2	Treg	58592	29467
p28	m2	Th	36596	22514	p50	m1	CD8	20543	16660
p28	m3	Th	35370	25183	p50	m2	CD8	12776	11968
p28	m4	Th	72550	53657	p50	m3	CD8	27498	24647

p28	m5	Th	41709	32809	p50	m4	CD8	14910	14091
p28	m1	Treg	4760	3316	p50	m5	CD8	3272	2897
p28	m2	Treg	4336	2793	p50	m1	Th	35051	19014
p28	m3	Treg	15885	7555	p50	m2	Th	23841	18568
p28	m4	Treg	8248	5243	p50	m3	Th	43587	29085
p28	m5	Treg	4833	3272	p50	m4	Th	27159	21604
p30	m1.1	CD8	25255	22870	p50	m5	Th	5008	4156
p30	m1.2	CD8	41821	37788	p50	m1	Treg	20872	9002
p30	m2.1	CD8	23215	21726	p50	m2	Treg	7752	4385
p30	m2.2	CD8	36449	31444	p50	m3	Treg	13953	7600
p30	m3.1	CD8	17293	16391	p50	m4	Treg	10777	5905
p30	m3.2	CD8	34831	30367	p50	m5	Treg	688	599

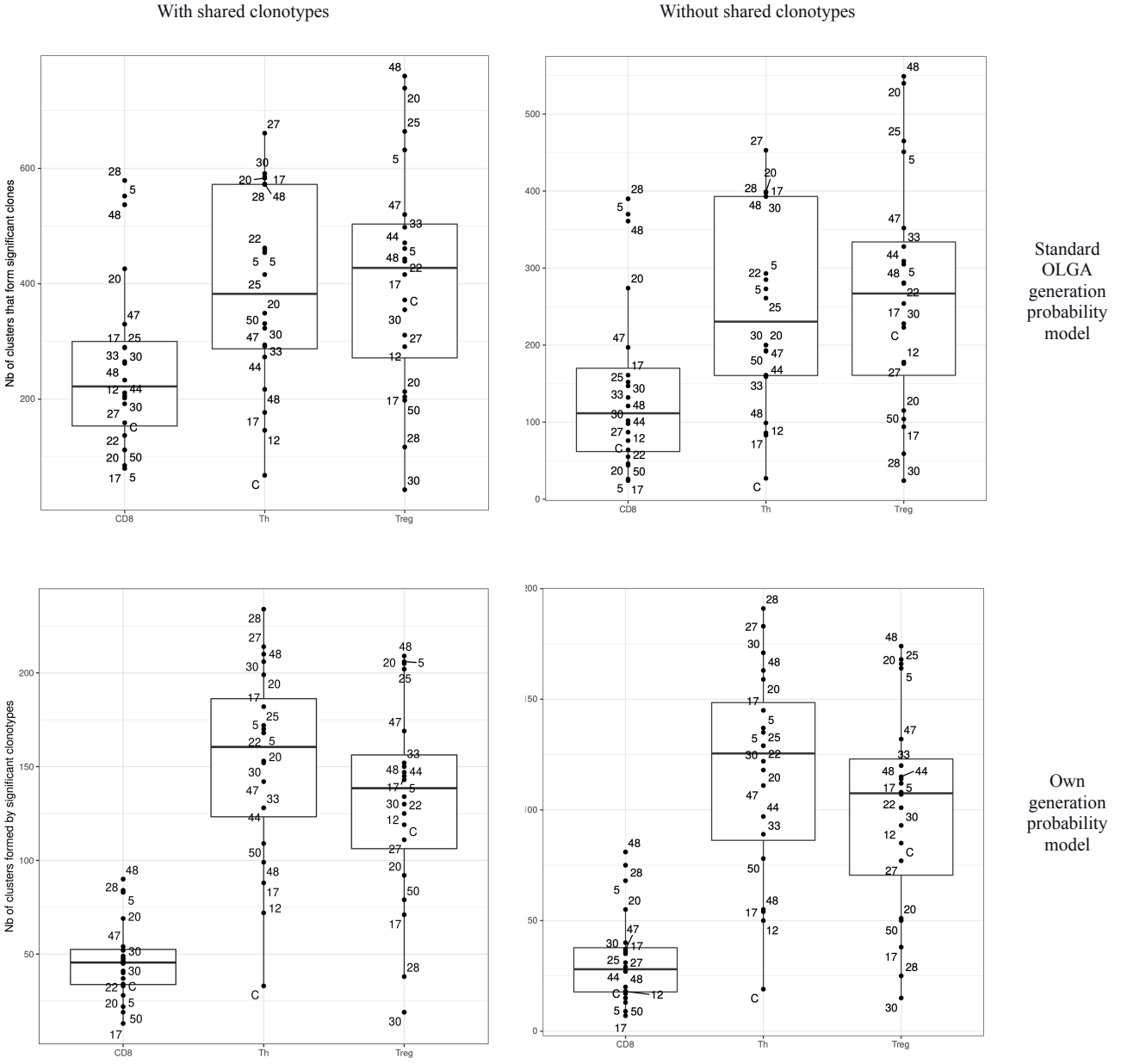
Appendix Table 3. Number if identified clusters of significant TCRs with frequency higher than 0.25%

Peptide	**Type**	**Replica**	**Nb of identified clusters**	**Peptide**	**Type**	**Replica**	**Nb of identified clusters**
12	CD8	0	0	30	Treg	1	4
12	Th	0	9	30	Treg	2	1
12	Treg	0	1	33	CD8	0	0
17	CD8	1	0	33	Th	0	3
17	Th	1	1	33	Treg	0	3
17	Th	2	11	44	CD8	0	0
17	Treg	1	3	44	Th	0	7
17	Treg	2	0	44	Treg	0	4
20	CD8	1	0	47	CD8	0	0
20	CD8	2	1	47	Th	0	10
20	Th	1	4	47	Treg	0	1
20	Th	2	16	48	CD8	1	2
20	Treg	1	1	48	CD8	2	2
20	Treg	2	5	48	Th	1	4
22	CD8	0	2	48	Th	2	13
22	Th	0	6	48	Treg	1	4
22	Treg	0	5	48	Treg	2	5
25	CD8	0	0	5	CD8	1	1
25	Th	0	27	5	CD8	2	1
25	Treg	0	3	5	Th	1	6

27	CD8	0	1	5	Th	2	13
27	Th	0	12	5	Treg	1	2
27	Treg	0	0	5	Treg	2	2
28	CD8	0	1	50	CD8	0	1
28	Th	0	16	50	Th	0	13
28	Treg	0	2	50	Treg	0	1
30	CD8	1	0	C	CD8	0	0
30	CD8	2	0	C	Th	0	1
30	Th	1	14	C	Treg	0	0
30	Th	2	17				



Appendix Fig. 1. Cluster frequency distribution of 10 the most abundant clusters of significant TCRs. Clonotype was identified as significant by TCRgrapher if the adjusted p-value was lower than 0.05. Standard OLGA generation probability model was used for TCRgapher analysis. Each line corresponds to one group.



Appendix Fig. 2. Distribution of the number of clusters formed by significant clonotypes between T cell subsets. Clonotype was identified as significant by TCRgrapher if the adjusted p-value was lower than 0.05. There were 14 groups immunised with different B16F10 melanoma peptides and control groups. Five peptides have two experimental groups. Each dot corresponds to one group. There were five mice per group, except control groups and one p17 group.