Open Source Tool for VH-replacement Products Discovery and Analysis

Adel Gazizova, Andrey Zolotarev
Saint Petersburg State University
Saint-Petersburg, Russia
adelgazizova@yandex.ru, andrewzoldy@gmail.com

Vladislav Myrov, Anastasiya Vinogradova
Saint-Petersburg Academic University
Saint-Petersburg, Russia
myrov_vl@aptu.ru, vinogradova.nastasia@yandex.ru

Aleksandr Cheblokov
Peter the Great St.Petersburg Polytechnic University
Saint-Petersburg, Russia
cheblokov.aa94@gmail.com

Evgeny Bakin
State University of Aerospace Instrumentation
T. Dobzhansky Center for Genome Bioinformatics, SPbU
Saint-Petersburg, Russia
eugene.bakin@gmail.com

Oksana Stanevich
First Pavlov State Medical University
Saint-Petersburg, Russia
oksanaayzsilnieks@gmail.com

***Abstract*—In this project an open source tool for discovery and analysis of abnormal immunoglobulin genes rearrangements was developed. The main goal is in finding so-called ”footprints” - small parts of incompletely rearranged V-segments in IGH genes. Joint usage of publicly available databases, on-line markup IMGT V-Quest and clonal families analyzer Partis allowed to prepare an informative dataset, which was processed via the developed tool. A few dependencies between patient phenotype and footprints statistics were discovered.**

I. Introduction

Immunoinformatics is one of the most challenging areas of the modern medicine [1]. Its significant progress is due to a rapid growth of availability of next generation sequencing instruments (NGS) as well as a continuous development of new algorithms oriented on immunological data processing.

For example, nowadays, repertoire sequencing (RepSeq) is one of the most powerful way for adaptive immune system analysis. Commonly based on Illumina paired-end sequencing technologies, RepSeq allows extracting immunoglobulin genes sequences from a pool of B- or T-lymphocytes with a high precision and coverage [2]. This data can be further processed with various computer tools, e.g. for clonal trees reconstruction, diversity analysis, investigation of fundamental phenomenon in affinity maturation etc. [3], [4], [5].

A prevalence of RepSeq allowed the researchers discovering new interrelations of antibodies features and different diseases. Thus, high-throughput sequencing data is successfully used in development of tumor biomarkers as well as cancer immunotherapy [6]. In [7] the researchers have discovered significant similarity of clones in immune-related lesions and tumor, which indicated that immune response may be elicited against antigens located in tumor and distant organs. Authors of [8] propose a new melanoma therapy, which effectiveness was proved among other things by means of neoantigen-specific T-cells repertoire capturing. A plenty of other examples may be found in [9].

Besides mentioned, RepSeq was applied for discovery and investigation of VH-replacement phenomena (VHR). VHR is an abnormal antibody gene rearrangement in which a rearranged gene segment is not completely removed [10]. The remaining small piece of it (a so-called footprint) may significantly change properties of antibody [11]. In recent papers, authors have shown that VHR plays important role in HIV broadly neutralizing antibodies formation [12] as well as clonal lines shaping in lymphocytic leukemia [13].

However, despite the obvious relevance of VHR investigation, nowadays there is the only one tool, devoted to its analysis - VH Replacement Footprint Analyzer-I (VHRFA-I) [14]. This tool has a wide functionality and usability providing a powerful support for conducting various research concerning footprints. Unfortunately, to the best of our knowledge, VHRFA-I is not available in the public domain.

Thus, the goal of our project is a creation of simple yet useful open source tool, allowing users to perform basic operations with antibody genes sequences: downloading, segmentation, potential footprints extraction and general statistic analysis. This tool may be further used as a component of more complex pipeline, oriented on VHR analysis.

The rest of the paper is organized as follows. Section II provides a brief overview of biological aspects of VH-replacement. In Section III we present chosen tools and databases for pipeline development. Section IV is devoted to an auxiliary program created for automatic interaction with IMGT/V-QUEST (on-line sequence alignment software oriented on immunoglobulin genes [15]). In Sections V and VI a whole pipeline description and its application examples are given respectively. The paper is finalized with Conclusions.

II. Biological background of VH-replacement

A. Phenomenon of VH-replacement

During an adaptive immune response to pathogens, a maturation of highly specialized cells, B-lymphocytes, occurs. This process passes through the stages of pro- and then pre-B lymphocytes, at the end of which a unique pair of the light and heavy chains of the immunoglobulin is expressed on the surface of the cell.

The sequence of immunoglobulin heavy chain is formed by recombination of the V-, D- and J-segments initially located in a germline (see Fig. 1). In order for recombination to occur, the proteins of RAG-1 and RAG-2 must be connected to short signal sequences (recombination signal sequences, RSS) flanking each segment in the germline [11].

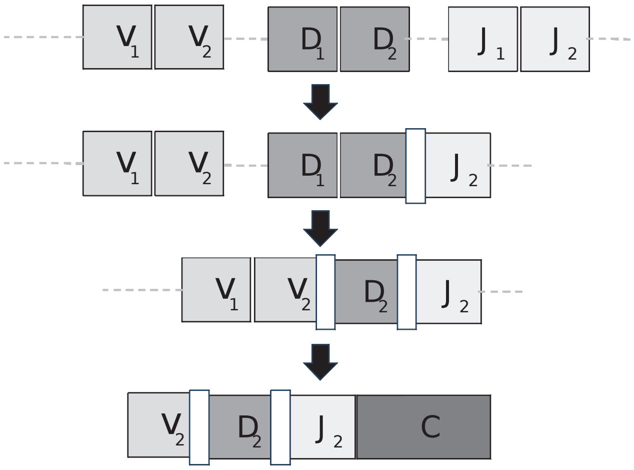

Fig. 1. The process of classical V-, D- and J- recombination

However, in spite of the huge variety of sequences provided by this mechanism, the inaccuracies of the segment rearrangement generate a large number of nonfunctional sequences, two-thirds of which are outside the reading frame [11].

In model mice it was noticed that some of the already formed IgH sequences soon after the first recombination undergoes a second one, that edits the existing unproductive sequence [16], [17]. The additional recombination is called VH-replacement and occurs in at least 5% of the human immunoglobulin repertoire [18]. The same frequency was calculated for the phenomenon of VHR in wild-type mice, but in mice with an autoreactive-prone phenotype and patients suffering from autoimmune diseases, this frequency was significantly higher [19], [20].

Since the discovery of this phenomenon, the replacement of VH has been actively studied in transgenic mouse models, including mice that were genetically engineered to contain two non-productively rearranged heavy chain alleles [21], as well as in human cell lines and human blood. Together with the error correction mechanism, it makes an additional contribution to the diversity of the Ig repertoire.

B. Disrupting of 12/23 rule during VH-replacement

The original RSSs are heptamer (CACTGTG) and nonamer (GGTTTTTGT), which are separated by a spacer with a size of 12 bp or 23 bp [10]. According to a rule 12/23, classical recombination occurs between gene segments flanked by RSSs with different lengths. In the future, if the sequence after rearrangement requires additional recombination, rule 12/23 is disrupted via the previous excision of all remaining D-genes and its flanking RSSs with 12-bp spacers. For this reason, a new recombination occurs between the upstream V-segment flanked by 23-spacer RSS and highly conserved cryptic RSS (cRSS; TACTGTG) located closer to 3 end of the sequence. As a result, the footprint of a preexisting V-gene remains in the sequence after the replacement (see Fig. 2).

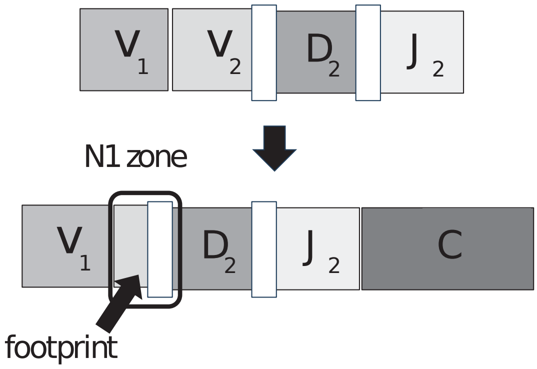

Fig. 2. An upstream heavy chain variable gene segment (VH) invades a preexisting rearrangement (VDJ)

C. VH-replacement in Central and Peripheral Organs of Immunopoiesis

There are several hypotheses whether the VH-replacement occurs in pre- and pro-B lymphocytes in the bone marrow, or it can be detected in B-lymphocytes in organs of peripheral immunopoiesis.

The theory that the VHR occurs in bone marrow is supported by the evidence of N-additions in junction sequences of IgH knock-in mice that have undergone VHR [22]. An addition of random nucleotides (N-nucleotides) by the TdT enzyme, which is maximally expressed in stage of pro- and pre-B lymphocytes, indirectly indicates that VHR occurs at or near the time of conventional IgH gene rearrangement.

D. Criteria for footprint search and list of footprint sequences

According to mentioned patterns and peculiarities of V-segments location in a germline, researchers proposed several criteria that make it possible to distinguish a footprint from a random sequence [10].

Criteria 1. Footprint should be in N1 zone (zone formed after random N-nucleotide addition and located between V- and D-genes in rearranged sequence).

Criteria 2. Footprint V-gene should be located before sequence V-gene in the locus.

Criteria 3. Footprint can not appear from VH6.1 gene.

Further in our research we used these criteria for footprint search. The sequences of footprint was taken from supplementary material of the paper [20].

III. Components of developed tool

Currently a lot of data is available in on-line immunological databases. It provides a possibility to make analysis which was not available before. These databases contains data about B- and T-cell epitopes [23], raw sequences [24], HIV sequences [25] and many other. However, the raw data in most cases cannot be applied directly and have to be analyzed for a meaningful information extraction. Most of the biological data is unstructured (DNA or protein sequences, mutations and e.t.c.) and complex algorithms are required to work with it. Some of the core algorithms are implemented in online tools that do not require installation or a lot of prerequisite steps. Some of these tools will be described below.

One of the common tasks in immunoinformatics is V-D-J rearrangement analysis. The goal is to identify V-D-J genes and find their positions. The most common tools for this task are V-Quest [26] and IgBlast [27]. These tools are based on alignment of known sequences of genes and provide quite accurate results. Also these tools have an option to provide scores for each found gene and the researcher may use this information to decide to accept the result or to reject it.

However, these tools have some limitations: the alignment-based algorithm may give the wrong results if genes in sequences have mutations or are not in tool database.

To avoid the limitation of alignment-based algorithms the family of stochastic algorithms was created [28], [29]. These algorithms are based of Hidden Markov Models (HMM) and can identify V-D-J genes correctly even if these genes have mutations. This approach has limitations too: it requires the HMM model to run and the results depend from the train dataset.

In our research we used alignment based tools because they are more reliable and are widely used. To compare and choose between IgBlast and V-quest tools we performed a series tests of a short dataset [30]. We chose this dataset because only individuals homozygous for the most common genotypes, IGHV3-23*01 and IGHJ6*02, were included in this study [30]. The results of IgBlast and V-quest were quite similar but V-quest is possible to detect the N1-zone position and characterize the sequence as productive or unproductive. Due to this reason we chose V-quest as a component of our tool.

Despite the easiness of use of the chosen web-based online tool, it has the crucial limitation for researches: most of the tools has limitations for the number of data for single run (V-Quest can accept only 50 sequences at once). In our research we had more than 32 thousand of sequences for analysis and manual running of the tool could be too time consuming. To solve this problem weve developed a web-bot, described in the next section.

For analysis of VH-replacement process we used the data from NCBI nucleotide database [31]. To find the relevant data we used the search query ”(immunoglobulin heavy chain) AND ”Homo Sapiens” NOT ”pseudogene” and chose the next filters: Molecule Type - ”Genomic DNA/RNA”, Source databases - ”INSDC (GeneBank)”, Sequence length - ”up to 500 bp”. Thus, we retrieved 32301 sequences totally.

The general structure of developed tool is shown in Fig. 3 and will be discussed in further sections.

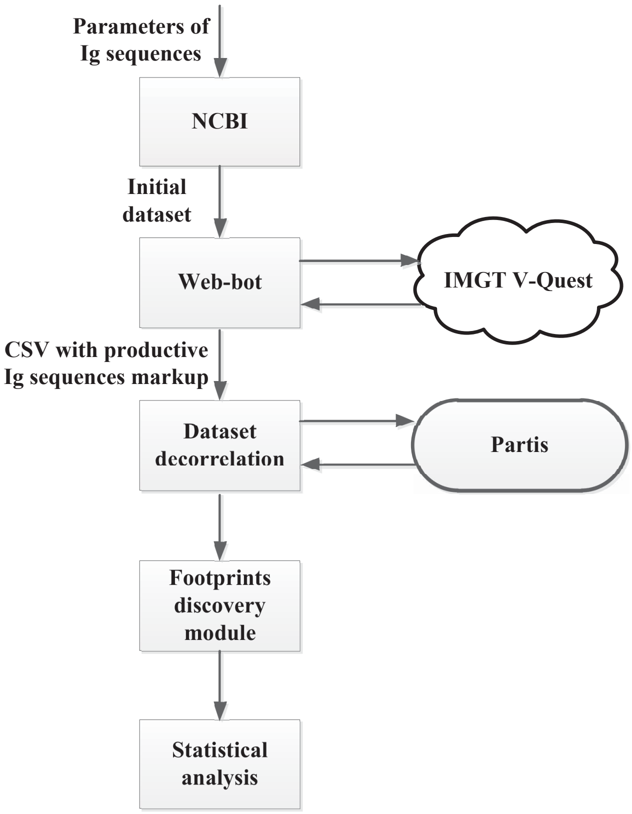
Parameters of Ig sequences
NCBI
Initial dataset
Web-bot
IMGT V-Quest
CSV with productive Ig sequences markup
Dataset decorrelation
Partis
Footprints discovery module
Statistical analysis

Fig. 3. General structure of developed tool

IV. Web-bot for automated requests generation

A. Summary of solution

It was necessary to solve the problem with the limitations of the IMGT V-quest due to the large amounts of experimental data. Routine parameter input for every fifty sequences makes any statistical processing extremely difficult and long.

The solution we developed is a web-bot that allows the researcher to automate the processing of large amounts of data subject to the limitation described above. We came to the conclusion that a suitable basis for our objective is Selenium Web-Driver [32]. Selenium is a software library with the open source code, which is widely represented for a number of the most popular web browsers and compatible with such programming languages as C#, Python, JavaScript and others. This module emulates user behavior on the site, what allows to set parameters once and then implement the repetition of their setting by Selenium API. Also, the reason for choosing this software was a number of performance advantages over other similar tools [33].

As a result of our work we present the program that automates requests to IMGT for huge datasets and includes an interface for configuring the search parameters that are specific to a particular task. Result of program execution is a table in CSV-format that contains data required for the researcher.

In addition to Selenium, we used the following libraries and modules.

TABLE I. LIBRARIES USED IN THE TOOL

Library (function)	Description
SeqIO (from BioPython)	Module allows to parse the data stored in fasta-files.
CSV	We decided to record the data from the V-quest in .csv format due to its popularity and usability.
Re (Regular Expression)	For parse output of the V-quest we used regular expression operations.
Argparse	Provides an opportunity to use command-line arguments when starting the program, such as a path to the data file, choosing species and receptor type or locus.

B. Output processing

The developed program initially opens a browser window where the address of the V-quest website is transferred, then the algorithm gets lists of available species and receptor types or loci. After this operation, the user is asked to select the settings that are mentioned above. Chosen settings will be used for next iterations (with this data) as a preset which frees user up from the necessity of selection them every fifty sequences.

Through the use of the temporary file, the entire data file will be loaded into the V-quest system by fifty sequences at a time. Each operation of this cycle will end with output parsing and recording of received data to the CSV table. Specifically, the analysis of V-quest output was carried out by the following regular expression (see Table II).

TABLE II. REGULAR EXPRESSION FOR OUTPUT PARSING

RegEx part	Description
`0: ID_csv = re.compile(r"""`	# beginning of the regular expression
`1: >(?P<ID>AB[\d]6\.[\d]1)`	# searching for ID of each record and save it as a group
`2: [\s\S]*?`	# passing all symbols to next pattern, using lazy quantifier
`3: Result[\s]summary:[\s]; [\s]*(?P<Functionality> (Productive)|(Unproductive)| (Unknown)|(No[\s]results)| (Productive[\s]with[\s] comments)|())`	# searching information about the functionality of the V-domain
`4: [\s\S]*?`	# passing all symbols to next pattern, using lazy quantifier
`5: V-GENE\sand\sallele; Homsap\s(?P<position> [A-Z\d\s]*`	# searching for position of a gene in the germline
`6: [\s\S]*?`	# passing all symbols to next pattern, using lazy quantifier
`7: CDR3-IMGT[\s]*`	# searching for CDR3-IMGT header
`8: (?P<CD_start>[\d]+)\.\.`	# searching for numeric, that inform in which position CDR3 region started and save it as a group
`9: (?P<CD_stop>[\d]+)`	# searching for numeric, that inform in which position CDR3 region ended and save it as a group
`10: [\s \S]*?`	# passing all symbols to next pattern, using lazy quantifier
`11: N1-REGION[\s]*`	# searching for N1-REGION header
`12: (?P<N1_start>[\d]+)`	# searching for numeric, that inform in which position N1 region started and save it as a group
`13: \.\.(?P<N1_stop>[\d]+)`	# searching for numeric, that inform in which position N1 region ended and save it as a group
`14: [\s \S]*?`	# passing all symbols to next pattern, using lazy quantifier
`15: /nucleotide\ssequence[\s] *(?P<N1_seq>[agtcAGTC]*)`	searching for a sequence of N1-region
`16: [\s \S]*?`	# passing all symbols to next pattern, using lazy quantifier
`17: /translation`	
`18: """, re.VERBOSE)`	# ending of the regular expression. re.VERBOSE - is a flag that makes expression more readable

As a result, the user receives a table with the markup of only productive sequences. By launching web-bot for test and working datasets weve got the following performance:

- 67 minutes for processing test dataset (6k sequences)
- 269 minutes for processing our working dataset (32k sequences)

V. PIPELINE FOR VH-DISCOVERY AND ANALYSIS

In order to estimate the frequency of VH-replacement occurrence, we have created a pipeline that allows to detect (corresponding to the above-mentioned criteria) footprints in the N1 zone of immunoglobulin H sequences.

Antibodies are produced by B-cells of the immune system. During the penetration of some foreign object into the organism, immune system makes an accurate selection of antibodies for a particular pathogen [34]. After finding a specific antigen, this B-cell receives a signal for reproduction, it undergoes a clonal expansion, as a result of which the organism contains a population of a single B-cell clones that produces the same antibodies containing the same footprint (see Fig. 4).

Therefore, in order to avoid a dependence of dataset items, the first step of our work was an application of Partis tool to separate our sequences into the clonal families and form a clonal-independent sample. Partis is a HMM-based framework and it can do sequence annotation, simulation, clonal family, and germline inference, but we have used only separation into the clonal families [28]. Due to the fact that there were no more than eight sequences in each clonal family, we have selected one sequence from each clonal family randomly. The second step of our work was the use of a tool named V-quest, based on alignment on the germline (various V, D, J genes), for marking N1 zones between V and D genes, inside the CDR3 zone [26]. According to the literature data, footprints must be located in the N1 zone, and we should look for them in this part of antibody (see Fig. 5).

The third stage was the search of footprints in N1 zones of clonal-independent antibody sequences. For this purpose, we created a script that performs exact and inexact search for footprints. The exact search is to find the substring (footprint sequence) in the string (N1 zone sequence). However, according to the literature data in the peripheral blood (outside the bone marrow) there is a process called somatic hypermutagenesis [34]. Somatic hypermutation is a point nucleotide replacement in the antibody sequence (see Fig. 6).

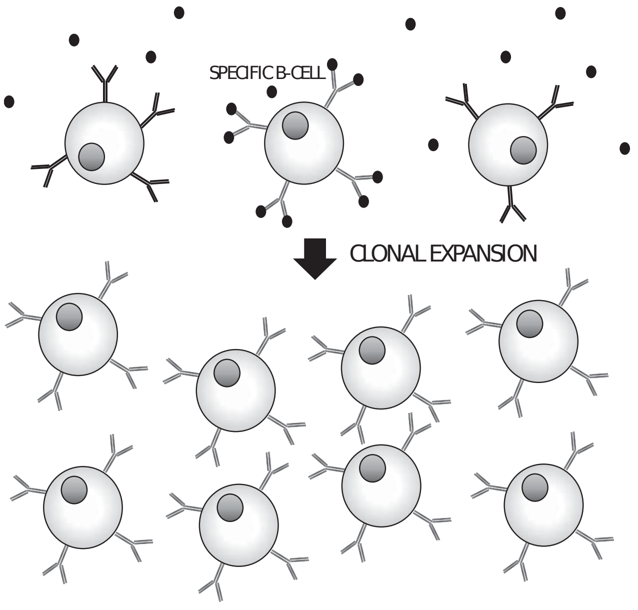

Fig. 4. Clonal expansion of a specific B-cell

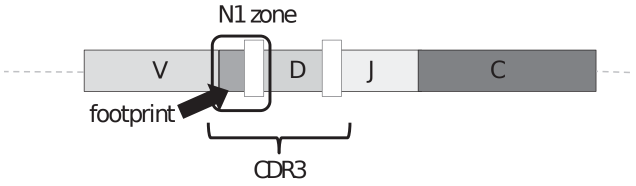

Fig. 5. N1 zone location in IGH sequence

This process helps further to adjust the antibody specificity to the antigen and presumably can mask footprints. Consequently, detected frequency of VH-replacement can be incorrect. Therefore, we applied a local alignment algorithm with a single permissible mismatch.

VI. Numerical example

As a part of our study, we have downloaded data from GenBank that included immunoglobulin H sequences. After that, we have ejected article titles for each group of sequences and divided our data into the phenotypic groups. Phenotypic groups included healthy people, people with autoimmune diseases, lymphomas, leukemia, chronic and acute infections (see Table III).

TABLE III. Phenotypic groups of antibody sequences

Phenotype	Number of sequences
Healthy people	2742
Systemic lupus erythematosus	1257
Multiple sclerosis	202
X-linked hyper-IgM syndrome	994
Wegeners granuloma	160
Chronic lymphocytic leukemia	741
Hodgkin lymphoma	98
Non-Hodgkin lymphomas	328
Hepatitis C	317
HIV-1	80
Infectious mononucleosis	130
Acute viral and bacterial infections	249
Pneumococcal vaccine	97

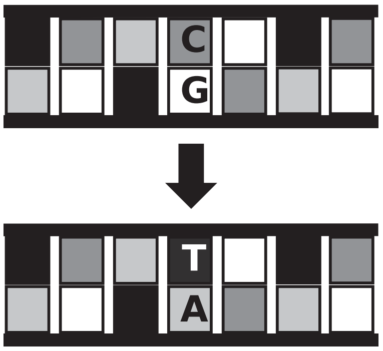

Fig. 6. Somatic hypermutation

Then we have applied the search described above to establish the frequency of VH-replacement in people with different phenotypes and obtained the following results:

1) *Exact search*: During exact search we have found that VH-replacement frequency significantly increases for subjects infected with HIV-1, as well as for ones vaccinated against pneumococcus (see Fig. 7).
2) *Inexact search*: During inexact search we obtained a similar trend, but we see a significant increase of VH-replacement frequency for subjects with multiple sclerosis, X-linked hyper-IgM syndrome, HIV-1, infectious mononucleosis and people vaccinated against pneumococcus. However, for subjects with systemic lupus erythematosus VH-replacement frequency decreases (see Fig. 8).

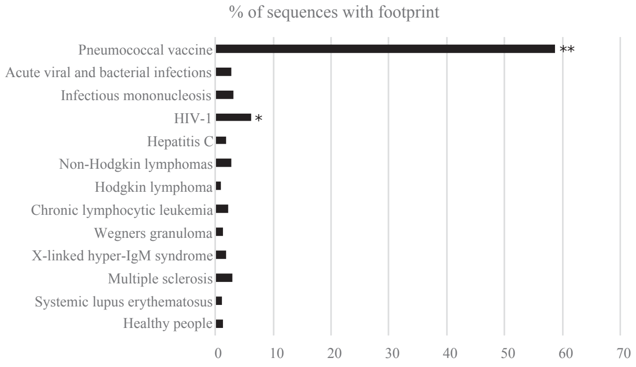

Fig. 7. Percent of sequences with footprint (exact search)

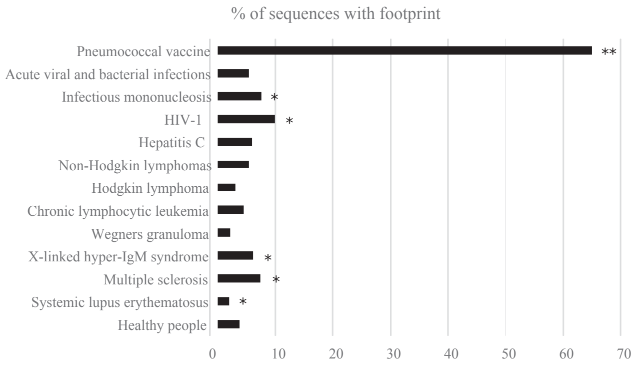

Fig. 8. Percent of sequences with footprint (inexact search)

The results obtained are consistent with the previous research previously described in [20].

VII. Conclusion

In this project a flexible tool for immunoglobulins footprints analysis was developed. Application of this tool allowed to discover a few statistically significant dependencies between VH-replacement frequency and phenotype. For example pneumococcal vaccination significantly increases its rate.

The further extension of this project may be as follows:

- perform relevant biological analysis of found dependencies;
- extend statistical analysis towards footprints positions;
- make a detailed research of footprints statistics on high-quality problem-oriented datasets, e.g. comparing B-cells rearrangements before and after vaccination.

The developed tool is available at [35].

Acknowledgment

We would like to thank Andrey Slabodkin and Maria Chernigovskaya for their essential and inspiring participation in discussions devoted to this project.

References

[1] R. K. De and N. Tomar, Immunoinformatics, Second edition. New York: Humana Press.

[2] J. Benichou, R. Ben-Hamo, Y. Louzoun, and S. Efroni, Rep-Seq: uncovering the immunological repertoire through next-generation sequencing., Immunology, vol. 135, no. 3, pp. 183191, Mar. 2012.

[3] H. X. Dang, B. S. White, S. M. Foltz, C. A. Miller, J. Luo, R. C. Fields, and C. A. Maher, ClonEvol: clonal ordering and visualization in cancer sequencing., Ann. Oncol., vol. 28, no. 12, pp. 30763082, Dec. 2017.

[4] N. Niknafs, V. Beleva-Guthrie, D. Q. Naiman, and R. Karchin, SubClonal Hierarchy Inference from Somatic Mutations: Automatic Reconstruction of Cancer Evolutionary Trees from Multi-region Next Generation Sequencing., PLoS Comput. Biol., vol. 11, no. 10, p. e1004416, Oct. 2015.

[5] M. El-Kebir, L. Oesper, H. Acheson-Field, and B. J. Raphael, Reconstruction of clonal trees and tumor composition from multi-sample sequencing data., Bioinformatics, vol. 31, no. 12, pp. i62-70, Jun. 2015.

[6] B. Ye, D. Smerin, Q. Gao, C. Kang, and X. Xiong, High-throughput sequencing of the immune repertoire in oncology: Applications for clinical diagnosis, monitoring, and immunotherapies., Cancer Lett., vol. 416, pp. 4256, Mar. 2018.

[7] H. L?ubli, V. H. Koelzer, M. S. Matter, P. Herzig, B. Dolder Schlienger, M. N. Wiese, D. Lardinois, K. D. Mertz, and A. Zippelius, The T cell repertoire in tumors overlaps with pulmonary inflammatory lesions in patients treated with checkpoint inhibitors., Oncoimmunology, vol. 7, no. 2, p. e1386362, 2018.

[8] P. A. Ott, Z. Hu, D. B. Keskin, S. A. Shukla, J. Sun, D. J. Bozym, W. Zhang, A. Luoma, A. Giobbie-Hurder, L. Peter, C. Chen, O. Olive, T. A. Carter, S. Li, D. J. Lieb, T. Eisenhaure, E. Gjini, J. Stevens, W. J. Lane, I. Javeri, and C. J. Wu, An immunogenic personal neoantigen vaccine for patients with melanoma., Nature, vol. 547, no. 7662, pp. 217221, Jul. 2017.

[9] N. Jiang, Immune engineering: from systems immunology to engineering immunity., Current Opinion in Biomedical Engineering, vol. 1, pp. 5462, Mar. 2017.

[10] W. Meng, S. Jayaraman, B. Zhang, G. W. Schwartz, R. D. Daber, U. Hershberg, A. L. Garfall, C. S. Carlson, and E. T. Luning Prak, Trials and Tribulations with VH Replacement., Front. Immunol., vol. 5, p. 10, Jan. 2014.

[11] A. Sun, T. I. Novobrantseva, M. Coffre, S. L. Hewitt, K. Jensen, J. A. Skok, K. Rajewsky, and S. B. Koralov, VH replacement in primary immunoglobulin repertoire diversification., Proc Natl Acad Sci USA, vol. 112, no. 5, pp. E458-66, Feb. 2015.

[12] K. O. Saunders, L. K. Verkoczy, C. Jiang, J. Zhang, R. Parks, H. Chen, M. Housman, H. Bouton-Verville, X. Shen, A. M. Trama, R. Scearce, L. Sutherland, S. Santra, A. Newman, A. Eaton, K. Xu, I. S. Georgiev, M. G. Joyce, G. D. Tomaras, M. Bonsignori, and B. F. Haynes, Vaccine Induction of Heterologous Tier 2 HIV-1 Neutralizing Antibodies in Animal Models., Cell Rep., vol. 21, no. 13, pp. 36813690, Dec. 2017.

[13] S. Trudel, H. Ghamlouch, J. Dremaux, C. Delette, V. Harrivel, J.-P. Marolleau, and B. Gubler, The Importance of an In-depth Study of Immunoglobulin Gene Rearrangements When Ascertaining the Clonal Relationship between Concomitant Chronic Lymphocytic Leukemia and Multiple Myeloma., Front. Immunol., vol. 7, p. 625, Dec. 2016.

[14] L. Huang, M. D. Lange, and Z. Zhang, VH Replacement Footprint Analyzer-I, a Java-Based Computer Program for Analyses of Immunoglobulin Heavy Chain Genes and Potential VH Replacement Products in Human and Mouse., Front. Immunol., vol. 5, p. 40, Feb. 2014.

[15] V. Giudicelli, X. Brochet, and M.-P. Lefranc, IMGT/V-QUEST: IMGT standardized analysis of the immunoglobulin (IG) and T cell receptor (TR) nucleotide sequences., Cold Spring Harb. Protoc., vol. 2011, no. 6, pp. 695715, Jun. 2011.

[16] R. Kleinfield, R. R. Hardy, D. Tarlinton, J. Dangl, L. A. Herzenberg, and M. Weigert, Recombination between an expressed immunoglobulin heavy-chain gene and a germline variable gene segment in a Ly 1+ B-cell lymphoma., Nature, vol. 322, no. 6082, pp. 843846, 1986.

[17] M. Reth, P. Gehrmann, E. Petrac, and P. Wiese, A novel VH to VHDJH joining mechanism in heavy-chain-negative (null) pre-B cells results in heavy-chain production., Nature, vol. 322, no. 6082, pp. 840842, 1986.

[18] Z. Zhang, M. Zemlin, Y.-H. Wang, D. Munfus, L. E. Huye, H. W. Findley, S. L. Bridges, D. B. Roth, P. D. Burrows, and M. D. Cooper, Contribution of Vh gene replacement to the primary B cell repertoire., Immunity, vol. 19, no. 1, pp. 2131, Jul. 2003.

[19] L. Huang, M. D. Lange, Y. Yu, S. Li, K. Su, and Z. Zhang, Contribution of V(H) replacement products in mouse antibody repertoire., PLoS ONE, vol. 8, no. 2, p. e57877, Feb. 2013.

[20] M. D. Lange, L. Huang, Y. Yu, S. Li, H. Liao, M. Zemlin, K. Su, and Z. Zhang, Accumulation of VH Replacement Products in IgH Genes Derived from Autoimmune Diseases and Anti-Viral Responses in Human., Front. Immunol., vol. 5, p. 345, Jul. 2014.

[21] J. Lutz, W. M?ller, and H.-M. J?ck, VH replacement rescues progenitor B cells with two nonproductive VDJ alleles., J. Immunol., vol. 177, no. 10, pp. 70077014, Nov. 2006.

[22] C. Chen, Z. Nagy, E. L. Prak, and M. Weigert, Immunoglobulin heavy chain gene replacement: a mechanism of receptor editing., Immunity, vol. 3, no. 6, pp. 747755, Dec. 1995.

[23] R. Vita, L. Zarebski, J. A. Greenbaum, H. Emami, I. Hoof, N. Salimi, R. Damle, A. Sette, and B. Peters, The immune epitope database 2.0., Nucleic Acids Res., vol. 38, no. Database issue, pp. D854-62, Jan. 2010.

[24] R. Leinonen, H. Sugawara, M. Shumway, and International Nucleotide Sequence Database Collaboration, The sequence read archive., Nucleic Acids Res., vol. 39, no. Database issue, pp. D19-21, Jan. 2011.

[25] C. Kuiken, B. Korber, and R. W. Shafer, HIV sequence databases., AIDS Rev., vol. 5, no. 1, pp. 5261, Mar. 2003.

[26] V. Giudicelli, D. Chaume, and M.-P. Lefranc, IMGT/V-QUEST, an integrated software program for immunoglobulin and T cell receptor V-J and V-D-J rearrangement analysis., Nucleic Acids Res., vol. 32, no. Web Server issue, pp. W435-40, Jul. 2004.

[27] J. Ye, N. Ma, T. L. Madden, and J. M. Ostell, IgBLAST: an immunoglobulin variable domain sequence analysis tool., Nucleic Acids Res., vol. 41, no. Web Server issue, pp. W34-40, Jul. 2013.

[28] D. K. Ralph and F. A. Matsen, Consistency of VDJ rearrangement and substitution parameters enables accurate B cell receptor sequence annotation., PLoS Comput. Biol., vol. 12, no. 1, p. e1004409, Jan. 2016.

[29] S. Munshaw and T. B. Kepler, SoDA2: a Hidden Markov Model approach for identification of immunoglobulin rearrangements., Bioinformatics, vol. 26, no. 7, pp. 867872, Apr. 2010.

[30] L. Ohm-Laursen, M. Nielsen, S. R. Larsen, and T. Barington, No evidence for the use of DIR, D-D fusions, chromosome 15 open reading frames or VH replacement in the peripheral repertoire was found on application of an improved algorithm, JointML, to 6329 human immunoglobulin H rearrangements., Immunology, vol. 119, no. 2, pp. 265277, Oct. 2006.

[31] Nucleotide - NCBI. [Online]. Available: https://www.ncbi.nlm.nih.gov/nucleotide?cmd=search. [Accessed: 29-Jan-2018].

[32] Selenium software official website, Web-Driver project page. [Online]. Available: http://www.seleniumhq.org/projects/webdriver/. [Accessed: 28-Jan-2018].

[33] I. Singh and B. Tarika, Comparative Analysis of Open Source Automated Software Testing Tools: Selenium, Sikuli and Watir, International Journal of Information & Computation Technology, vol. 4, no. 15, pp. 15071518, Jun. 2014.

[34] C. Janeway, Immunobiology 5: the immune system in health and disease. New York: Garland Pub., 2001.

[35] GitHub repository. [Online]. Available: https://github.com/evgeny-bakin/VH-replacement-Products-Discovery-and-Anlalysis [Accessed: 01-March-2018].