Washington University in St. Louis
Washington University Open Scholarship

Arts & Sciences Electronic Theses and Dissertations | Arts & Sciences

Winter 12-15-2018

Sequence analysis methods for the design of cancer vaccines that target tumor-specific mutant antigens (neoantigens)

Jasreet Hundal
Washington University in St. Louis

Follow this and additional works at: https://openscholarship.wustl.edu/art_sci_etds

Part of the Allergy and Immunology Commons, Bioinformatics Commons, Genetics Commons, Immunology and Infectious Disease Commons, and the Medical Immunology Commons

Recommended Citation
Hundal, Jasreet, "Sequence analysis methods for the design of cancer vaccines that target tumor-specific mutant antigens (neoantigens)" (2018). *Arts & Sciences Electronic Theses and Dissertations*. 1692.
https://openscholarship.wustl.edu/art_sci_etds/1692

This Dissertation is brought to you for free and open access by the Arts & Sciences at Washington University Open Scholarship. It has been accepted for inclusion in Arts & Sciences Electronic Theses and Dissertations by an authorized administrator of Washington University Open Scholarship. For more information, please contact digital@wumail.wustl.edu.

WASHINGTON UNIVERSITY IN ST. LOUIS
Division of Biology and Biomedical Sciences
Human & Statistical Genetics

Dissertation Examination Committee:
Elaine R. Mardis, Chair
Malachi Griffith, Co-Chair
Beatriz M. Carreno
William E. Gillanders
Robert D. Schreiber
S. Joshua Swamidass

Sequence Analysis Methods For The Design Of Cancer Vaccines That Target Tumor-Specific Mutant Antigens (Neoantigens)

by
Jasreet Hundal

A dissertation presented to
The Graduate School
of Washington University in
partial fulfillment of the
requirements for the degree
of Doctor of Philosophy

Dec 2018
St. Louis, Missouri

© 2018, Jasreet Hundal

Table of Contents

List of Figures

List of Tables

Acknowledgments

It seems sort of surreal as I write this section to thank the many people who have supported me and have helped me in this journey. But I'd first like to express my gratitude to the Almighty without whose benevolence I would not have accomplished any of this.

I would like to thank Elaine and Malachi for their mentorship and guidance throughout the degree, and for undoubtedly being the best mentors one could ask for. Elaine, as I look back, I cannot imagine how far we have come together. From being my first "boss" back in the Technology Development to being my thesis mentor, I feel very fortunate indeed to have grown under your tutelage. Working with you on some of the most cutting-edge projects in the field motivated me to pursue research and to go back to graduate school. Leading by example, your enthusiasm and unabashed passion for science and research have always encouraged me to discover and pursue new questions. Also, I feel immensely grateful for all the time that you have spent on me, answering countless questions and editing last-minute manuscript drafts, no matter which time zone you were in. Thank you for empowering me to carve my own niche in the immunogenomics space without forgetting the bigger questions in research. To Malachi, thank you for guiding me through the everyday challenges that come with being a bioinformatics scientist. Having you as my mentor has enabled me to imbibe your keen attention to detail, and develop a sense of perfection in everything I strive to do. I'm grateful that you have given me opportunities to grow as a scientist by trusting me in situations even when the answers may not have seemed too convincing. My heartfelt thanks to the other twin genome, Obi, for being there always and giving me invaluable help with the day-to-day challenges and mentoring me despite not officially being a mentor or thesis committee member.

I would also like to thank my illustrious committee Bob, Will, Beatriz and Josh who made this work possible. Without the questions you had, there would be no answers today. Bob, thank you for introducing me to the wonderful world of neoantigens, and also to the world of injecting them in a mouse! Will and Beatriz, thank you for helping me understand the clinical complexities of this work and in fact helping me see a direct translation of my work in the clinic. Josh, thank you for the relentless questions that have motivated me to think like a scientist as I pursue them.

I am profoundly grateful to the amazing members of the Griffith Lab, Schreiber Lab and McDonnell Genome Institute, who have been my everyday companions and made the working space collaborative and supportive. Special thanks to Susanna Kiwala, Katie Campbel, Zach Skidmore, Ben Ainscough, Connor Liu, Yang-Yang Feng, Huiming Xia, Chris Miller, Josh McMichael, Jason Walker and Matt Gubin and many others for their willingness to always extending a helping hand in times of need.

To my family, they say it takes a village to raise a child. Well, it takes an even larger village to raise a child AND get a PhD. I would like to thank my village, my family, for their unfaltering support throughout my life, and especially for helping me through the last year as I juggled between my roles of being a new mother and a graduate student. I'm extremely grateful to have been blessed with such wonderful parents who have given me both - the roots and the wings, keeping me grounded while supporting me in all my choices. I'm forever indebted to my mother who has been not only a doting mother but also my spiritual guide, enabling me to see the bigger picture in the most testing of times. To my father, for inculcating in me a sense of discipline, perfection and the attitude to accept nothing but the best in whatever I undertake. I also thank my younger sister, Rubina, for being more than just

a sibling and for being my support system through the many hurdles of graduate school and life, and of course, for pestering me incessantly with the dreaded “when will you finish school?” question! Perhaps secretly you wanted us both to be class of 2018.

To my dearest baby Sikander, thank you for being the most adorable and easy going child one could ask for (and thank you more for sleeping through the night so I could work!). You have opened my heart to so much love and happiness that I didn't even know I was capable of receiving or giving. Thank you for your naughty laughter that melted away all my stress as I was finishing through the last pieces of my work. I’m also thankful to my furry kid, Zouk for being with me and sitting peacefully near my feet, without asking any questions, without judging and just being there.

And lastly, but certainly not the least, I’m grateful to my husband, my rock, my lifeline Pawan who has been the backbone to my sanity. Thank you for sticking with me through the thicks and thins of graduate life and otherwise, for supporting me in achieving my dreams, and sometimes my crazy ideas! Thank you for stepping into your shoes as a caring father when I needed to be away for long hours. I feel extremely fortunate and blessed to have a partner like you.

Jasreet Hundal

Washington University in St. Louis

Dec 2018

Dedicated to my parents.

ABSTRACT OF THE DISSERTATION

Sequence analysis methods for the design of cancer vaccines that target tumor-specific mutant antigens (neoantigens)

by

Jasreet Hundal

Doctor of Philosophy in Biology and Biomedical Sciences

Human & Statistical Genetics

Washington University in St. Louis, 2018

Professor Elaine R Mardis, Chair

Professor Malachi Griffith, Co-Chair

The human adaptive immune system is programmed to distinguish between self and non-self proteins and if trained to recognize markers unique to a cancer, it may be possible to stimulate the selective destruction of cancer cells. Therapeutic cancer vaccines aim to boost the immune system by selectively increasing the population of T cells specifically targeted to the tumor-unique antigens, thereby initiating cancer cell death.. In the past, this approach has primarily focused on targeted selection of 'shared' tumor antigens, found across many patients. The advent of massively parallel sequencing and specialized analytical approaches has enabled more efficient characterization of tumor-specific mutant antigens, or neoantigens. Specifically, methods to predict which tumor-specific mutant peptides (neoantigens) can elicit anti-tumor T cell recognition improve predictions of immune checkpoint therapy response and identify one or more neoantigens as targets for personalized vaccines. Selecting the best/most immunogenic neoantigens from a large number of mutations is an important challenge, in particular in cancers with a high mutational load, such as melanomas and

smoker-associated lung cancers. To address such a challenging task, Chapter 1 of this thesis describes a genome-guided *in silico* approach to identifying tumor neoantigens that integrates tumor mutation and expression data (DNA- and RNA-Seq). The cancer vaccine design process, from read alignment to variant calling and neoantigen prediction, typically assumes that the genotype of the Human Reference Genome sequence surrounding each somatic variant is representative of the patient's genome sequence, and does not account for the effect of nearby variants (somatic or germline) in the neoantigenic peptide sequence. Because the accuracy of neoantigen identification has important implications for many clinical trials and studies of basic cancer immunology, Chapter 2 describes and supports the need for patient-specific inclusion of proximal variants to address this previously oversimplified assumption in the identification of neoantigens. The method of neoantigen identification described in Chapter 1 was subsequently extended (Chapter 3) and improved by the addition of a modular workflow that aids in each component of the neoantigen prediction process from neoantigen identification, prioritization, data visualization, and DNA vaccine design. These chapters describe massively parallel sequence analysis methods that will help in the identification and subsequent refinement of patient-specific antigens for use in personalized immunotherapy.

Chapter 1: Introduction & Background

1.1 Introduction

After the landmark discovery of the double helix structure of deoxyribonucleic acid (DNA) in 1953 by Watson and Crick, coupled with several advances in molecular biology techniques, specifically the dideoxynucleotide sequencing techniques of Sanger, the completion of the human reference sequence in 2004, and the development of basic computational DNA analysis and annotation software, heralded a new era for disease research[1]. 'Finishing' the first human reference genome sequence provided a new roadmap to understanding where genes in the genome were located and organized on chromosomes. A few years after the Human Genome Reference sequence was announced, a new paradigm shift occurred in the process of DNA sequencing data production. The new DNA sequencing instruments and methods, known collectively as massively parallel sequencing (MPS), could process millions of sequence reads in parallel in a single instrument run compared to previous Sanger methods, and thereby significantly reduced the time, resources and costs to generate genome sequencing data. These techniques were also called 'next generation' sequencing technologies as they truly spurred a new direction to producing genomic data.[2]

1.2 Massively Parallel Sequencing (MPS) technologies and analysis

Several different technologies and companies emerged around the same timeframe, with different approaches to next generation sequencing[3]. Some of the early commercial contributors were Roche/454, Applied Biosystems SOLiD and Illumina, that differed based

on the sequencing chemistry used- pyrosequencing for Roche/454, ligation-based sequencing for SOLiD and polymerase-based sequencing-by-synthesis for Illumina. Amongst these, Illumina emerged as a market leader in terms of maintaining a reasonable cost of sequencing per Gb of data as well as the accuracy and applicability of the data generated. Currently, Illumina's platform is the most widely used in the field.

1.2.1 Steps to generate MPS data

There are three main steps that are involved in generating MPS data. The first and foremost step is library production. This involves shearing and fragmenting genomic DNA, enzymatically polishing the fragment ends, and ligating known sequence adapters onto the fragment ends. These adapters may also carry bar coded DNA identifiers or indexes to permit downstream pooling of samples. The next step involves attaching these adapter-ligated library sequences onto a solid surface with complementary adapter sequences, and enzymatically amplifying the fragments. Fragment amplification is needed to provide sufficient signal during the sequencing reaction for on-instrument detection. Similarly, libraries for generating RNA-Seq data also can be prepared for MPS, first by conversion of the RNA to DNA using reverse transcriptase, followed by adapter ligation and surface amplification. The data resulting from RNA-Seq can be analyzed to determine digital expression values for genes in a given tissue or tumor. The last step is sequence data generation by reading the signals produced from the stepwise sequencing process that occurs at each amplified fragment population. This process involves detecting the identity of each nucleotide base that is added by DNA polymerase onto the fragments in each amplified library fragment cluster, obtained from differential label detection in certain instruments or by the sequential addition of each nucleotide reagent being coupled with a subsequent detection step that follows each nucleotide addition. Because this sequencing process takes

place on the solid support, for hundreds of millions to billions of fragment clusters being sequenced in parallel, the technology is called 'massively parallel' sequencing. The read length for most commonly used MPS technologies is relatively shorter (100-300bp) than traditional Sanger sequencing read lengths (800bp). This fact, coupled with the high throughput approach to sequencing enabled by MPS, requires analysis of the resulting reads by *in silico* alignment to a fixed reference genome, followed by variant detection. In turn, the requisite sampling level or "coverage" needed to sequence genomes and obtain high confidence variant calling is nearly triple that of Sanger sequencing coverage levels (30-fold or higher coverage by MPS versus 8-10 fold coverage by Sanger). Additionally, the more compute-intensive nature of analyzing MPS data sets for genomes as large as the human (3Bbp), means that these large data sets require carefully constructed and validated analytical pipelines.

MPS can be used to either sequence the entire genome (whole genome sequencing or WGS), or there are methods to select out genomic content from a whole genome library, such as the exons of known protein-coding genes ("exome") or a smaller number of selected genes or regions ("panel"). This is typically accomplished by an approach called "hybridization capture" sequencing or "exome" sequencing (in case of all protein coding genes), whereby synthetic probes designed to hybridize the exon sequences of protein coding genes are designed with covalently attached biotin molecules. By mixing such probes with the library fragments from a whole genome library under appropriate conditions, hybrids form between the synthetic probes and their cognate sequences in the library. These hybrids are subsequently captured by mixture with streptavidin-labeled magnetic beads, due to the binding of biotin by streptavidin, and subsequently the hybrids are isolated by applying a magnetic pull-down to isolate them from the remaining mixture. After isolation, the captured

library fragments are denatured from the synthetic probes and sequenced by MPS. Since human exome sequencing targets only the exons, which make up about 1.5% of the genome, the attendant data production is less expensive, the reads are more readily interpreted for variants, and the depth of coverage obtained can be higher and therefore more sensitive than compared to whole genome sequencing.

1.2.2 Analysis of MPS data

The sequence read alignment step is most often preceded by quality control of raw sequencing data and data preprocessing steps such as trimming of adapter sequence data from the reads. The quality control procedure can inform about any GC bias in the sequencing experiment by analyzing the GC content distribution, as well as any inconsistencies in the experiment by determining the read length distribution.

While smaller, less complex genomes (viral or bacterial) can be assembled from the resulting short read MPS data using specific assembly algorithms, larger complex genomes like the human require alignment of short sequencing reads to the reference genome as a first step toward data analysis. In genomes such as human, mapping the reads to the reference genome is further complicated by not-yet-completed gaps in the reference genome or by differences between the reference and the genome being studied, including structural variants (chromosomal inversions, deletions, and translocations). Additionally, 48% of the human genome consists of repetitive elements that complicate accurate mapping of short reads due to multiple mapping likelihood at many regions in the genome.

The post alignment processing involves data recalibration to mark and/or remove duplicates, and recalculation of quality scores after adjusting for local misalignments. Duplicate

sequencing reads may arise due to preferential enzymatic amplification of DNA fragments by the polymerase, especially when there are differences in the length of library fragments (shorter fragments are amplified preferentially) or differences in G-C content (more skewed A-T/G-C ratios amplify poorly). Such duplicates are reduced to a single read representative by de-duplicating software to avoid misrepresentation of copy number altered regions or to propagate PCR errors. To identify focal structural variants i.e. small insertions or deletions, the read data require a realignment and base quality score recalibration (BQSR) process to enhance their detection by variant calling software programs.

After mapping and post-processing of the aligned reads to the reference genome, variant calling algorithms identify mutations in the sequenced DNA compared to the reference. This includes detection of several different types of variants such as single nucleotide variants (SNVs), focal insertions and deletions (indels), and copy number variants.

One of the major prerequisites for successful detection of high quality variants is adequate sequence read data coverage of the genome or exome, which ensures sufficient depth on a given region of interest has been achieved to give statistical confidence of any variant identified therein. Also, filtering of discordant reads, such as those with multiple mismatches above a set threshold avoids sequencing errors being called as variants. Finally, variants called in aligned reads due to known false positivity contributors are often removed using statistical filters, such as variants found at the ends of reads where data quality is generally lower or variants called with reads only originating from single read orientation.

The variant list is then annotated to determine the protein level effect of the peptide changes and to assess the functional significance of the predicted variants. The pathogenicity of

variants are then evaluated. Some commonly used tools to annotate and predict functional impact include Variant Effect Predictor (VEP)[4,5], SIFT[6,7], CHASM[8], PolyPhen-2[9], MutationAssessor[10], and ParsSNP49[11]. The mutation calls can be classified into different types depending on their effect on the resulting protein sequence such as - a) Missense (nonsynonymous) mutation, where a single amino acid changes to another amino acid b) Nonsense mutation: where the point mutation changes an amino acid to a STOP codon resulting in premature termination of translation c) Silent (synonymous) mutation: where the resulting mutation does not change an amino acid d) Frameshift mutation: these include insertion or deletion of a number of bases such that the frame of translation changes, leading to a completely new amino acid sequence and/or introducing a premature STOP codon, e) stopgain mutation, which eliminates the wildtype stop codon, resulting in a longer protein sequence. Furthermore, these mutations could either lead to "loss of function" of the protein or "gain of function", either of which may be equally important.

1.3 Cancer genome sequencing

1.3.1 Large scale tumor genome sequencing projects

As MPS technologies have attained widespread use and expanding application to multiple experimental aspects of biomedical research, their application to cancer genome characterization and discovery of new cancer genes in the research setting has been profound.

The first MPS-based cancer genome was sequenced and reported by Ley et al[12] in 2008. Single patient cancer genome sequencing studies such as this study also highlighted the importance of sequencing the patient's normal genome in addition to the tumor genome[13], as a means of distinguishing inherited ("germline") alterations from those acquired in and

thereby specific to, the cancer genome ("somatic"). These initial small studies were accomplished by WGS and were quite expensive at the time, but as the cost of generating MPS data decreased, large-scale projects to discover and catalogue cancers were initiated and resulted in data generation and analysis from thousands of adult and pediatric cancers. For example, large scale sequencing efforts have been completed by several international consortia, such as The Cancer Genome Atlas (TCGA)[14], International Cancer Genome Consortium (ICGC)[15], Pediatric Cancer Genome Project (PCGP) and Pan Cancer Analysis of Whole Genomes (PCAWG). Due to the advances in MPS technologies and analytical approaches, it was possible to sequence and identify the full range of somatic alterations in the genome, including single nucleotide variations, insertions and deletions, copy number variations, and large genomic rearrangements such as translocations, inversions, and other complex structural rearrangements[16].

In addition to defining the genomic landscape of thousands of cancer genomes across many disease types, these projects revealed the heterogeneity of mutational burden across cancer types (Figure 1.1). A broad spectrum of mutational frequencies is observed between, for example, carcinogen driven tumors such as tobacco smoking-associated lung cancers and UV radiation-driven melanomas which show the highest mutational burden, versus hematological tumors that have the lowest mutational load.

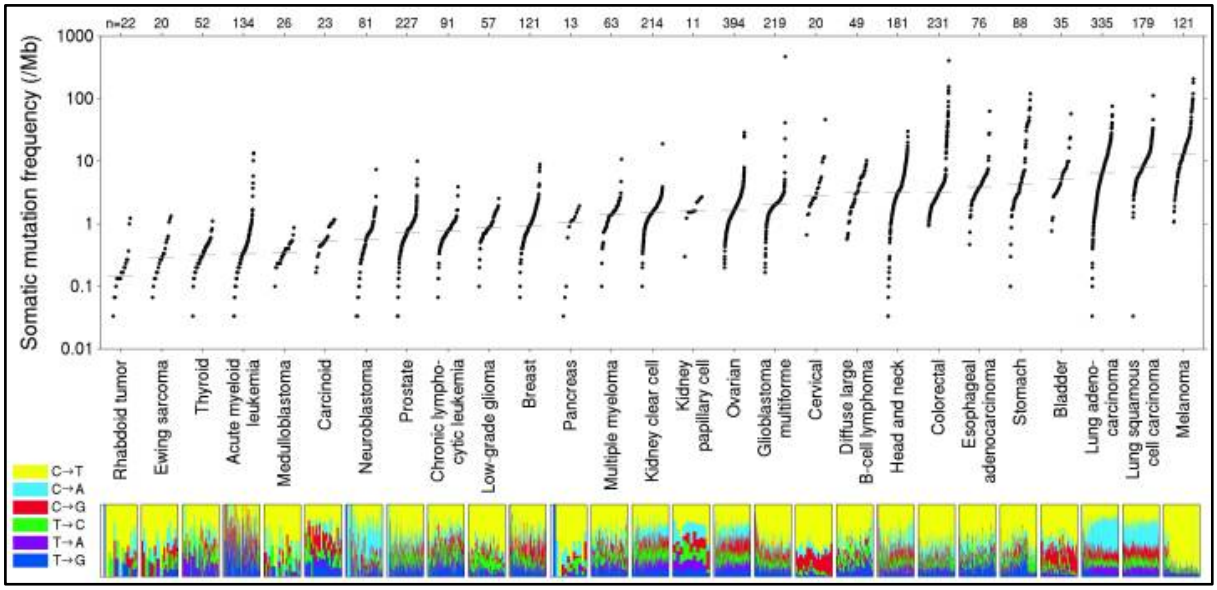


Figure 1.1 Mutational heterogeneity of cancer. Figure originally published in Lawrence et. al.[17] *"Each dot corresponds to a tumor–normal pair, with vertical position indicating the total frequency of somatic mutations in the exome. Tumor types are ordered by their median somatic mutation frequency, with the lowest frequencies (left) found in hematological and pediatric tumors, and the highest (right) in tumors induced by carcinogens such as tobacco smoke and ultraviolet light. Mutation frequencies vary more than 1,000-fold between lowest and highest across different cancers and also within several tumor types. The bottom panel shows the relative proportions of the six different possible base-pair substitutions, as indicated in the legend on the left."* [17]

Also, the large scale detailed analysis of the patient cohorts has led to findings about driver genes, driver mutations and passenger mutations, including the identification of significantly mutated genes (SMGs) occurring within specific tissue sites, as well as across multiple cancer types. These characterizations of genetic alterations have helped in substantially advancing our understanding of cancer genomes, especially as the data are integrated with RNAseq, methylation and other omics data types produced concurrently.

By capturing even low abundance aberrations in clonal populations using MPS, these studies have also highlighted that even individual tumor types can have substantial genomic heterogeneity. It was found that even though there were SMGs seen across patient cohorts, at an individual patient level, the patterns and combinations of mutations that led to

development of cancer were different. This complexity further underscored the need to sequence each patient's genome, due to the level of detail needed to discern and characterize the individual disease.

1.3.2 Cancer genomics based treatments

Coincident with our ability to sequence cancer DNA has been the development of new classes of cancer therapies that are developed to address specific cancer driver gene alterations. Initially, these were therapies such as Imatinib that targeted the protein fusion product of the BCR-ABL gene fusion, or Herceptin that targeted HER2 amplified breast cancers, both of which are interpretable without the need for DNA sequencing. Over time due to enhanced discovery offered by MPS-based analysis, we have now ascertained that cancer genes can be mutated in different tissue sites, such as BRAF mutations in melanomas, lung cancers, and brain cancers. This new realization has shifted diagnostic applications of MPS[18] from single gene mutation testing (often accomplished by PCR and Sanger sequencing) to multi-gene panel assays capable of surveying the many possible genes that may be mutated and may offer treatment decision-supporting evidence to physicians. Gene panel testing is also favored because it is cost effective and less time consuming to analyze the resulting data. It also offers a high rate of sensitivity to detect mutations in tumor DNA, which often has different amounts of normal cells interspersed that can decrease sensitivity, since it is possible to achieve fairly high coverage (300-500x) at the regions of interest.

Due to unprecedented leaps in our understanding of cancer development powered by large scale sequencing of cancer samples, genomics based treatments have allowed the field of therapeutics move away from a 'one size fits all' approach to one that is more personalized to

the patient's genome. Concurrently, advances in molecular drug development and targeted therapies have enabled the transition of genomic assays into clinical use in patients with cancer[19].

Cancer genomics based assays have introduced a new era of molecular pathology and personalized, or "precision" medicine. By interpreting genomic information in light of targeted therapies, it is possible to identify additional cancer patients who might benefit.

An emerging use of cancer genomics data is in predicting responses from new classes of cancer drugs known broadly as immunotherapy or immune checkpoint blockade therapy. In particular, cancer patients who have responses to immune checkpoint blockade therapies have been characterized by MPS as having an elevated mutational load or tumor mutation burden (TMB). By coupling our understanding of the impact of high TMB, namely the encoding of new protein sequences, that are unique to the cancer cells, with the potential impact on the adaptive immune system that would identify these novel proteins/peptides as 'non-self', one can predict that the response to immune checkpoint therapy is due to high numbers of tumor-specific mutant antigens (neoantigens) being presented to the immune system on these cancer cells.

1.4 Cancer Immunotherapy

1.4.1 Tumor antigens

In the late twentieth century, the immune-focused therapies for cancer that were postulated and tested did not deliver much clinical benefit, resulting in a diminution of enthusiasm for such therapeutic approaches. These studies were based on the hypothesis that due to the

aberrations in the cancer genomes, unique peptides are presented on the cancer cells by the major histocompatibility complex (MHC) molecules which are then recognized by the Cytotoxic T lymphocytes (CTL)[20,21]. The altered peptides either could be normal differentiation antigens or aberrantly expressed normal proteins that are overexpressed in tumor cells versus the normal cells (tumor-associated antigens; TAAs), or those resulting from either oncogenic viral proteins or mutations unique to the tumor (tumor-specific antigens/neoantigens; TSAs).

Tumor Associated antigens arise from non-mutated self proteins and are aberrantly expressed in tumor cells. One such type of self-differentiation antigens are the cancer–testis antigens (CTAs) that are present in immune-privileged normal cells but also are selectively presented to the immune system by tumor cells. These are not expressed in normal tissues except for testis, fetal ovaries, and trophoblasts but are also expressed in varying tumor types[22]. Melanoma Associated Antigen-1 (MAGE-1) was the first identified CTA which was discovered in 1991[23] using autologous cytotoxic T lymphocytes and autologous tumor mRNA. Another type of self-antigen is derived from non-mutated melanocyte lineage-specific antigens, which include Melan-A/MART-1, gp100, and tyrosinase. These antigens usually have tissue specific expression and have been found to exhibit some level of enriched expression in various cancers. Since TAAs are also present in the normal cells, there is a risk of autoimmunity and the use of tumor-associated antigens as therapeutic targets can lead to normal tissue destruction and toxicity.

TSAs are antigens arising from somatic changes in the tumor acquired during cancer initiation and progression, resulting in unique peptide sequences and are not seen in the normal cells. Since these peptides are specifically presented in the tumor cells, they are likely

to be less susceptible to mechanisms of immunological tolerance. Early works on mouse models from the lab of Thierry Boon[24] identified the first neoantigen resulting from a point mutation in the P91A gene in a mutagenized mouse tumor. A few years later, Hans Schreiber and colleagues[25,26] demonstrated that tumor-specific mutations could result in immunogenic tumor-specific neoantigens, and demonstrated in vivo tumor rejection based on T cells targeting highly immunogenic, primary UV–induced mouse tumors. Another group demonstrated an autologous antibody-based method to clone and identify different human TSAs[27]. All of these efforts paved the way for harnessing the power of antigens - TAAs or TSAs and showed an increasing evidence that the tumor antigenome was associated with a combination of these antigens. However, historically targeting TSAs was more laborious and challenging as the therapeutic strategies are specific to individual patients. As a result, the focus was primarily on targeting the TAAs or shared antigens that are not tumor-specific but expressed on large groups of cancers.

1.4.2 Pre-MPS identification of tumor antigens

Several approaches were developed to detect potential antigens that could be used for cancer immunotherapy[28,29]. Traditionally, reverse-transcription-polymerase chain reaction (RT-PCR) and real-time PCR (RQ-PCR) were used to detect expression of TAAs for a range of solid and haematological malignancies but these assays were used to limited to tumor antigens which had already been discovered. Another approach called Representational Difference Analysis (RDA) was developed to identify the CTAs such as the MAGE family of antigens[30]. Briefly, this technology uses subtractive hybridization to PCR-mediated kinetic enrichment to detect the differences between RNA sequences from the normal tissue (driver) and a tumor sample (tester).

In 1995, Sahin et al[27] developed a method of *Se*rological analysis of tumor antigens by *Re*combinant cDNA *Ex*pression cloning (SEREX). This method provided a base for high throughput screening of several TAAs, and was used widely to screen for plethora of antigens (> 2000) (http://www.licr.org/D_programs/d4a1i_SEREX.php) across different types of solid tumors as well as haematological malignancies. One example is NY-ESO-1, which is probably one of the most immunogenic CTAs discovered to date. It elicits specific CD4+ as well as CD8+ T-cell mediated immune responses in patients with solid tumors. SEREX starts with the construction of a cDNA library from freshly isolated tumor cells, which is expressed recombinantly. Thereafter, recombinant proteins are transferred onto membranes and screened with diluted serum of the same patient. TAAs are identified by their reactivity with IgG antibodies present in the patient's serum. Though several antigens can be screened together, one of the most laborious step in the SEREX approach is the construction expression libraries. This approach also does not detect post-translational modifications.

Another approach known as *s*erological *p*roteome *a*nalysis (SERPA)[31] combined proteomics approaches for the purpose of separating proteins and the serological screening with human serum antibodies. It involves using the proteomics workflow for an effective separation on 2-DE gels for proteins that were extracted from primary tumors or cell lines, followed by an identification by Mass Spectrometry (MS). Unlike SEREX, this does not depend on recombinant expression of proteins, and thus can also screen for tumor specific post-translational modifications of proteins.

Besides using the Mass Spectrometry (MS) based approaches, cDNA microarrays also have been used to compare the differential expression of tumor antigens from normal tissues and cancer tissues. Most of these approaches focused on the identification and validation of

TAAs. Although there were attempts in the mid 1980's to identify tumor-specific mutant antigens (TSMAs) from cancers, these efforts were painstaking and not scalable using Sanger sequencing approaches.

1.4.3 Neoantigens

Many early studies demonstrated the importance of TSMAs or neoantigens as significant targets of antitumor immune responses[32,33,34]. However, screening for neoantigens before the sequencing of the human genome and the advent of MPS, was a daunting task with very little gain since it required the identification of mutations and HLA typing to be done for each patient. Thus, much of the focus for cancer immunotherapy based treatments in the early 2000s was on targeting and evaluation of TAAs and CTAs, as described.

With the completion of the first sequence of the human reference genome in 2004, early efforts to design PCR primers and amplify and Sanger sequence DNA from cancers to identify somatic mutations emerged, making it clear that cancer cells, especially in some tissue sites, carry multiple mutations in their genes. From these studies, a visionary suggestion emerged[35]. In 2008, Allison and Vogelstein[36] proposed that all cancers have mutations that could form neoantigens by conducting an *in silico* analyses of exome-sequencing data from breast and colorectal cancers. While the study lacked any experimental validation of the predicted neoantigens, they profiled several mutations that were predicted to form tumor-specific mutant antigens for CD8+ T cells, exhibiting the ability of the cancer genome to form epitopes recognizable by the immune system due to accumulated genetic changes.

In 2012, two independent groups demonstrated neoantigen prediction approaches based on MPS somatic variant identification, including how the identified variants, when considered in the context of MHC binding affinity, could predict tumor specific neoantigens in murine sarcoma models[37,38]. By using the resulting set of potential neoantigens to query T cell reactivity, the predictions were validated. It was further demonstrated that these validated neoantigens were the same epitopes recognized by anti-PD1 and anti-CTLA4[39] immune checkpoint blockade therapies and that peptide vaccines comprised of the peptide neoantigens provided either prophylactic or therapeutic efficacy. Several other studies have also characterized neoantigens as being derived from somatically mutated genes[40–43], and have shown that they can be recognized by T-cells.

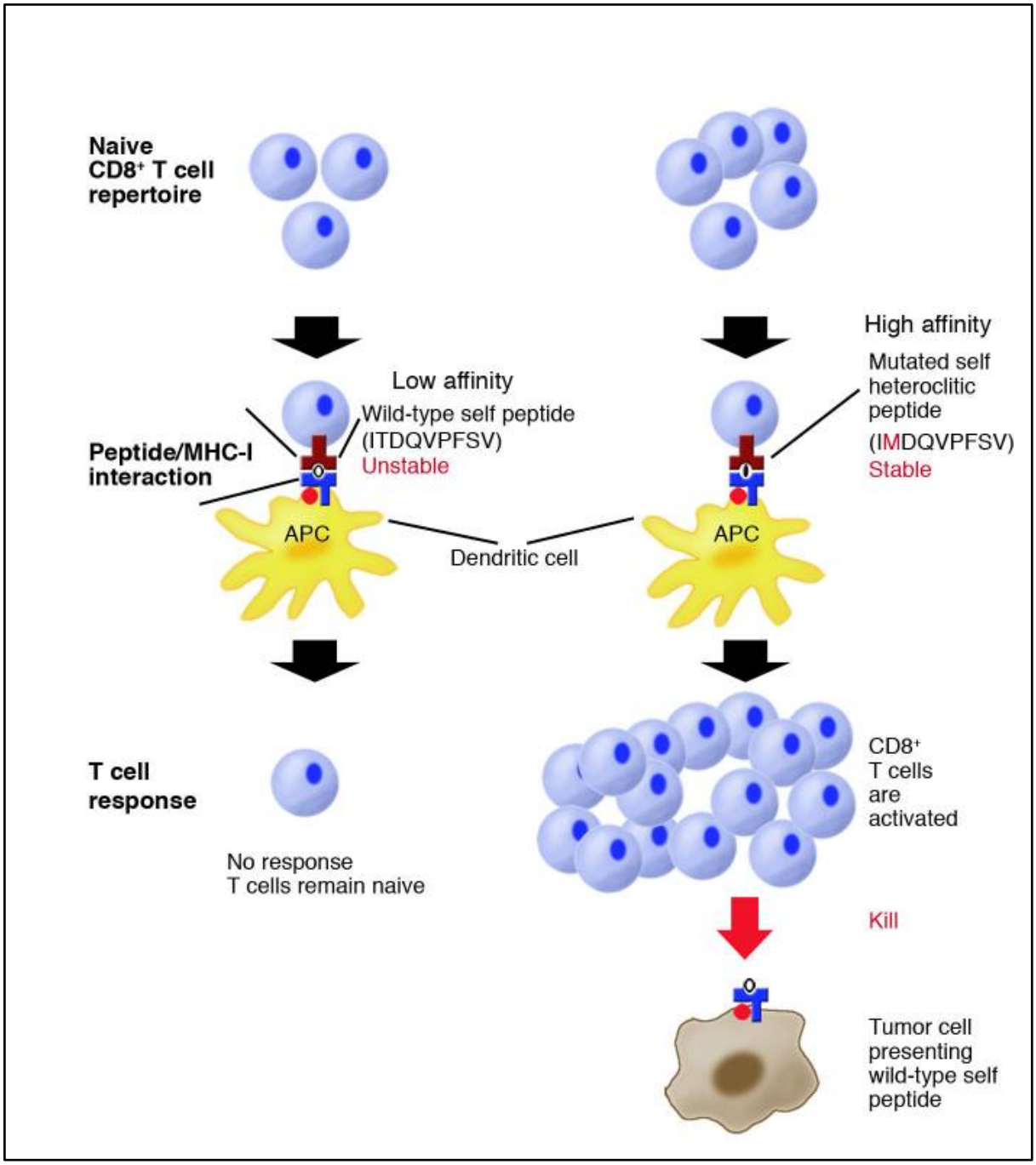


Figure 1.2: An overview of self vs non-self paradigm for neoantigen vaccines: *Figure originally published in Houghton and Guevara-Patiño[44] - "Initial studies demonstrated that host CD8+ T cells respond to a self peptide presented by MHC-I molecules on tumor cells. The wild-type self peptide ITDQVPFSV is unable to trigger an immune response against the tumor cells either due to being a weak binder or being eliminated due to the process of self-tolerance. The mutated peptide ITDQVPFSV results in a neoantigen which is highly stable and activates the host T cells. Once the CD8+ T cells are activated, they are competent to recognize and kill host tumor cells presenting the non-mutated self peptide."*

These landmark studies by Matsushita et al[37] and Castle et al[38] introduced "reverse immunology approaches" for identification of neoantigens and opened new avenues to tailor personalized treatments using TSMAs by exploiting the full mutanome of a patient. Briefly, in these studies the analysis pipelines combined genomic, bioinformatic and immunological methods[45] to determine the set of most immunogenic neoantigens (Figure 1.2). Starting with the set of somatic variants identified using MPS, the amino acid changes are translated, and representative short peptides are generated that window through the variant amino acid, to place it at locations from beginning to end in the peptide, and its ability to bind to the MHC is determined using epitope prediction algorithms. The normal exome data are analyzed through a specific algorithm to identify the HLA haplotypes used for this evaluation. The two most critical components of this approach are the identification of high quality somatic mutations using MPS data as well as accurate prediction of the peptide binding to the MHC alleles. As mentioned in previous sections, depth of coverage by MPS sequencing reads from the tumor is an important factor for detection of high confidence somatic variants. The coverage is also needed to correctly infer clonality of the mutation, based on the reads that identify the variant and their representation at the variant site of germline vs. mutant. This type of analysis is needed to partition variants in the founder clone (those present in every cell) from those in subclonal populations (not present in every cell, by inference). From an immunotherapy perspective, targeting neoantigens that associate with the founder clone produces an immune response that targets all cancer cells rather than only selected cells (i.e. those carrying subclonal mutations/neoantigens). Another aspect that determines the immunogenicity of the selected neoantigen is its ability to bind to the MHC molecule. Typical, peptides of lengths 8–11 amino acids bind to MHC Class I molecules that present them to cytotoxic CD8+ T cells and peptides of longer lengths, typically between 11–30 amino acids, bind to MHC Class II molecules and are presented to CD4+ T cells. Predicting the binding affinity of a

large number of peptides arising from somatic mutations to the MHC alleles is onerous, and therefore has now been streamlined by various computational pipelines.

1.4.4 Computational methods for peptide binding prediction to MHC alleles

One critical component for the correct prediction of antigen binding to MHC alleles is obtaining the high-resolution HLA allelotype of the patient. Traditionally, PCR and Sanger sequencing-based clinical assays have been used to derive patient-specific HLA haplotypes. More recently, several *in silico* based tools have been developed that can predict the correct HLA haplotypes at up to a 99% accuracy for four-digit resolution[46]. These tools either use an alignment-based approach or an assembly-based approach from MPS read data. Alignment-based HLA typing softwares such as PolySolver[47] and Optitype[48] align the read to the reference HLA sequences (genomic, exomic or transcriptomic) and use probabilistic modeling to infer correct HLA types. Assembly-based approaches such as ATHLATES[49] and HLAminer[50] involve *de novo* assembly of MPS read data into contigs which are then aligned to the reference sequences of known HLA alleles to resolve the haplotypes.

Experimental screening of large numbers of predicted neoantigen candidates is an expensive and time-consuming endeavor. As MPS technologies have become more widely used for cancer immunogenomics evaluations[51], computational methods to determine accurate binding prediction of the peptides to the MHC molecules (Class I or Class II) have also evolved[52].

Several computational tools of this type, including BIMAS[53], SYFPEITHI[54], and RANKPEP[55], rely on position-specific scoring matrices (PSSMs). These softwares make predictions of MHC-peptide binding based on the position of amino acids in the putative neo-

peptide as it sits in the binding cleft of the protein, such as at the anchor position, unusual anchor position, and auxiliary anchor position. The binding potential of different peptides is based on identification of allele-specific motifs and by taking into account the positions of the amino acids - optimally preferred positions of amino acids scored higher over the undesirably placed amino acids in calculating the binding affinity.

After the widespread adoption of MPS, the number of MHC alleles (specifically HLA alleles) that were identified also increased exponentially, and it became necessary to improve MHC binding prediction methods. One major shortcoming of the PSSM based approaches was the inability to correctly account for the interrelationships between different MHC cleft binding positions. This was overcome by using more complex, machine learning-based approaches to establish models for the different types of binding site interactions, and thereby determine patterns of binding based on existing datasets. These machine learning-based methods are broadly divided as follows:

a) *Artificial neural networks (ANN)*: These methods are trained using a set of experimentally derived binding affinity measurements for different class I and II HLA haplotypes. One of the first software programs developed using ANN-based modeling was NetMHC[56,57]. The accuracy of this method largely depends on the quality and size of the training set, and hence it is more accurate for predicting MHC-peptide binding potential for the more common MHC alleles due to the relative abundance of binding data for common haplotypes. Since rare HLA alleles are not as well supported by experimental binding data, more accurate algorithms were developed to calculate the peptide binding affinity potential to such alleles. NetMHCpan[57,58] and Pickpocket[59] are two such pan-specific algorithms that were trained by expanding the original training set to include data from other species, as

well as extrapolating the affinities for rare alleles from known binding affinities. Incorporating these additional parameters subsequently led to improved accuracy for rare MHC alleles. Another ANN-based approach, applied in NetMHCCons[60], combines multiple tools to obtain more reliable binding affinity predictions and uses consensus to compensate for the strengths and weaknesses of different algorithms. Further improvements to neural network-based methods have resulted in tools such as MHCflurry[61] and MHCnuggets[62].

b) *Stabilized Matrix Method (SMM)*: These methods use protein position-weight matrices to model the binding process. Examples include SMM[63] and SMMPMBEC[64] that use quantitative matrices not only for predicting binding of peptide to MHC class I molecules, but also for predicting both the transport of the peptide by transporter associated with antigen processing (TAP), and proteasomal cleavage of protein sequence.

Though most of the efforts have been focused on developing methods for binding predictions to MHC Class I molecules, some of these softwares such as NetMHCIIpan[65] and MHCnuggets[62] , extend support for Class II predictions as well.

Once a list of strong binding peptide candidates is determined using the aforementioned methods, the peptides are validated *in vitro* to determine their binding potential and/or their T cell based immune response using different assays. These include MHC Peptide binding assays that assess peptides based on their competitive binding to the MHC molecules, ELISPOT/ELISA that measures IFNγ production, flow cytometry-based Peptide–MHC tetramer/dextramer assays, or combinations of any of these.

The selected neoantigen peptides could be further employed to study different aspects of cancer immunology and immunotherapy. One type of cancer immunotherapy, called 'immune checkpoint blockade' acts by removing immune suppression by binding to cell surface markers that induce T cell exhaustion. Several groups have studied the correlation between tumor mutation burden (TMB) or neoantigen load as a predictor of immune checkpoint blockade response likelihood[66,67], and have shown a strong association between somatic mutation burden and clinical response in patients treated with anti-CTLA4 or anti-PD1. While checkpoint blockade therapies predominately provide a tumor-specific immune response, there could be often adverse reactions because these therapies target native immune molecules such as CTLA-4, PD-1 and PD-L1. Another type of immunotherapy-based approach called 'personalized vaccines' is more tumor specific since these therapies target neoantigens or TSMAs, that result from somatic alterations. As these therapies are tissue specific, they yield fewer off-target effects and hence, lowered likelihood of severe adverse events. To administer such vaccines, different types of vaccine platforms have been tested and employed in different settings, as explained in the next section.

1.4.5 Vaccine platforms for neoantigen-based therapy

Once a list of strong binding neoantigens has been identified and validated *in vitro*, a vaccine platform is required for therapeutic administration into the patient. Multiple vaccine platforms are being explored for personalized cancer vaccines, some of which are elucidated below[68]:

a) *DNA vaccines:* This vaccine platform involves constructing a circularized DNA insert carrying one or more encoded neoantigenic peptide sequences and inserting the construct into a DNA vector. One important consideration in the design of these

vaccines is the order of neoantigenic sequences along with spacer sequences such that no new strong binding junctional epitope is encoded by the resulting sequence. The DNA vaccine platform is relatively inexpensive to produce as a GMP-grade construct because DNA synthesis is rapid, automated and inexpensive, and DNA sequencing can be used to verify the final sequence of the construct prior to administration. These attributes also would enable rapid scalability of DNA vaccines for precision medicine applications in cancer therapy, if warranted.

b) *Peptide-based vaccines*: This vaccine platform involves the direct administration (by intramuscular injection) of synthetic neoantigenic peptide cocktails suspended in administrative adjuvant solution. A recent publication described the use of this vaccine platform using synthetic long 15–30-mer peptides suspended in poly-ICLC (Hiltonol)[69] wherein six melanoma patients were immunized with up to 20 distinct neoantigenic peptides predicted to bind to MHC class I molecules. Peptide vaccines are relatively more expensive to generate due to the requirement of synthesizing the peptides under GMP conditions, and don't always solubilize well with one another, so often several cocktails must be made in the adjuvant and administered separately. However, peptide based vaccines are easy to characterize, and straightforward to synthesize, making the platform relatively scalable.

c) *RNA-based vaccines:* These vaccines are conceptually similar to DNA and peptide vaccines wherein synthetic RNAs encode the various predicted neoantigenic peptides evaluated from the patient's tumor. One of the main advantages of using this approach is that by using mRNA as the vaccine, it can be readily translated once in the cell in order to induce antigen-specific T cell immune responses. However, stability of RNA

molecules and appropriate processing and presentation by antigen presenting cells remains one of the major challenges of this vaccination strategy. Sahin et al[70] used this vaccine platform to immunize 13 patients with advanced cutaneous melanoma. Synthetic RNAs encoding five linker-connected 27-mer peptides were formulated based on neoantigen predictions of both MHC Class I and Class II binders. In this study, 60% of the candidate neoepitopes elicited some sort of immune response, and the majority of T cell responses were CD4+, directed against MHC class II-restricted antigens.

d) *Dendritic cell (DC) based vaccines*: These vaccines involve collection of circulating blood, from which isolated dendritic cells are matured *ex vivo* in the presence of neoantigenic peptides. The matured dendritic cells are infused into the same patient. A recent melanoma trial by Carreno et al[71] used this strategy in three melanoma patients, who had been pretreated with ipilimumab. These patients were vaccinated with DC loaded with about 7 neoantigenic derived from each patient's tumor. Only a subset of the candidates showed existing or *de novo* immune responses, thus underscoring the difficulties of predicting peptides that will be processed and presented.

Using tumor specific neoantigens, these vaccine platforms and the associated trials have shown that there are few, if any, severe adverse events and little cross-reactivity to the wild type (normal) peptide. These initial clinical trials[69,70,71] show promising results about the safety and tolerance of vaccines that employ neoantigens for cancer therapy. Two of these studies further treated patients that had a recurrent/progressive disease with checkpoint blockade therapy (anti-PD1)[72] and clinically significant disease regression was reported,

showing value in the use of combinations such as neoantigen-based vaccines with immune checkpoint blockade.

Another type of therapy that employs the use of neoantigens, though not directly in the form of vaccines, is adoptive T cell transfer therapy (ACT). This type of therapy pioneered by Rosenberg et al. involves removing tumor-specific T cells, either from peripheral blood or the resected tumor, and expanding them *ex vivo* which allows them to regain their cytotoxic function, such that when they are transferred back into the patient, they can drive tumor elimination. ACT, when used on patients with solid tumors[41,73] showed measurable T-cell specific immune responses against neoantigens. Following treatment with ACT in pilot trials, it has been shown[74] that 40% of treated patients experienced complete regressions of all measurable lesions for at least five years. While ACT provides another tumor-specific approach for treatment with fewer side effects, these therapies are costly due to the need for *ex vivo* expansion of patient-specific T cells for each patient.

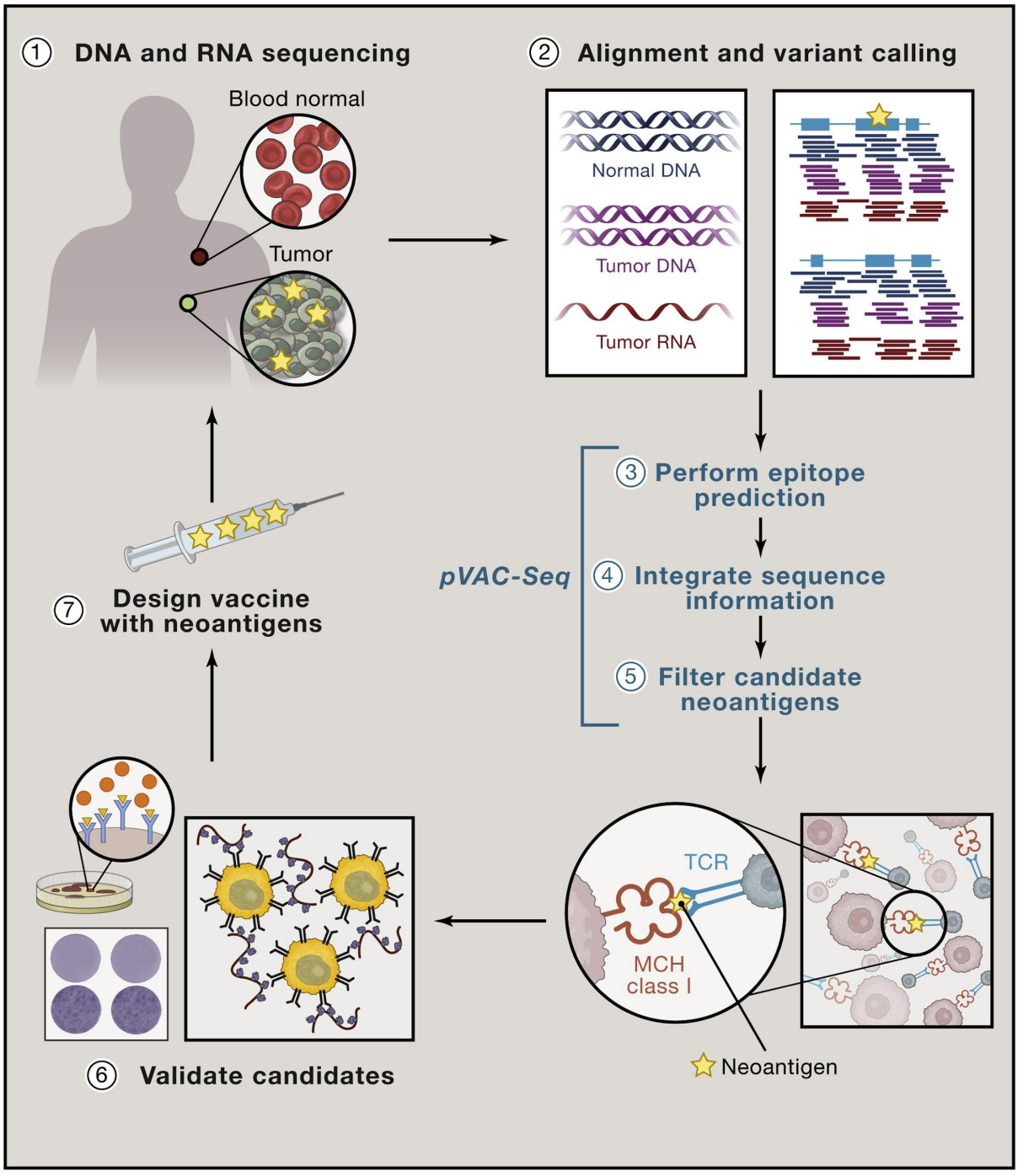


Figure 1.3: Overview of personalized vaccine design process: *Figure originally published in Liu and Mardis[45]. Tumor tissue and blood normal samples are extracted from the patient and exome and RNA sequencing is performed, followed by read alignment and variant calling algorithms to identify and confirm expression of somatic mutations. This input is used to determine strong binding neoantigens by using MHC class I & class II epitope prediction algorithms. The selected set of high quality neoantigens are further prioritized by ELISPOT and peptide binding assays, before being incorporated in a vaccine.*

1.5 Development and improvement of sequencing methods for neoantigen characterization

Methods to more accurately identify which tumor-specific mutant peptides (neoantigens) can elicit anti-tumor T cell immunity are needed if the pursuit of personalized vaccine therapy and adoptive T cell therapy continues to advance in medical care of cancer patients. Although the cost-related barriers to producing somatic mutations from MPS data sets have fallen away with new instrumentation and analytical improvements, barriers to facile use of computational analyses that yield neoantigen predictions still remain and, if not addressed, diminish the widespread adoption of this approach to cancer treatment. Since the mutational repertoire of a typical cancer genome can vary anywhere from 50 to 500+ mutations, there was a need for a strategy that integrates somatic mutation calls from MPS data, identifies the neoantigens in the context of the patient's HLA alleles, and parses out a list of optimal peptides for downstream testing, as illustrated in Figure 1.3. Chapter 2 describes such a genome-guided *in silico* approach to identifying tumor neoantigens that starts with a simple list of somatic nonsynonymous point mutations from which it predicts high affinity neoantigens refined by sequencing coverage and gene expression data. Such a pipeline has been utilized in several first-in-human clinical trials of patient-specific vaccines using different vaccine platforms, including DNA-based and dendritic cell-based vaccines. We have further refined the approach to neoantigen prediction from that presented in Chapter 2 by testing the impact of regional-specific adjustments to peptide sequences entering consideration for MHC binding affinity predictions. Namely, this approach accounts for the individual-specific variants in the germline, as well as the adjacent somatic alterations that occur most frequently in high mutation load tumors, proximal to the variant being evaluated for neoantigen binding. This correction in germline proximal variants is often not done

because there is an assumption that the human reference genome is representative of the patient's inherited variants, some of which may change amino acid sequence in the resulting peptide. All of these assumptions can lead to incorrect determination of neoantigenic peptide sequences and to incorrect binding affinities as a consequence. Chapter 3 describes our results from studying the effect of proximal sequence correction on the calculated binding affinities and suggests a corrective approach that, in some cases, dramatically impacts the calculated binding affinity.

We also recognized that neoantigen vaccine therapies may be applied in cancers without high mutational loads, and that the challenges of identifying a sufficient number of neoantigens to construct a vaccine might be challenging in this circumstance. Hence, we sought to increase the numbers of variants that could be evaluated for MHC binding by including rarer types of variants, including frameshift indel and gene fusions, into neoantigen prediction. In addition, since there is evidence that MHC Class II binding neoantigens are appropriate to include in vaccines[70], we also wanted to include these predictions into our established pipeline described in Chapter 2. Chapter 4 describes a modular computational framework to facilitate each of the critical stages of neoantigen characterization. It integrates the information from Chapters 2 and 3 into a toolkit that enables neoantigen prediction from the breadth of somatic alterations including point mutations, insertions, deletions, and gene fusions. Furthermore, prioritization and selection of these neoantigens for non-informatics savvy users is provided through a graphical web-based interface. Lastly, this revised approach provides a functionality to determine the optimal order of selected neoantigen candidates in a DNA vector-based vaccine.

Chapter 2: pVAC-Seq: A genome-guided *in silico* approach to identifying tumor neoantigens

Hundal, Jasreet et al. pVAC-Seq: A genome-guided *in silico* approach to identifying tumor neoantigens Genome Med. 8, 11 doi:10.1186/s13073-016-0264-5 (2016).

2.1 Introduction

Cancer immunotherapy has gained significant momentum from recent clinical successes of checkpoint blockade inhibition. Massively parallel sequence analysis suggests a connection between mutational load and response to this class of therapy. Methods to identify which tumor-specific mutant peptides (neoantigens) can elicit anti-tumor T cell immunity are needed to improve predictions of checkpoint therapy response and to identify targets for vaccines and adoptive T cell therapies. Here, we present a flexible, streamlined computational workflow for identification of personalized Variant Antigens by Cancer Sequencing (pVAC-Seq) that integrates tumor mutation and expression data (DNA- and RNA-Seq). pVAC-Seq is available at https://github.com/griffithlab/pVAC-Seq.

2.2 Background

Boon et al. were the first to demonstrate that cancer-specific peptide/MHC class 1 complexes could be recognized by CD8+ T cells present in cancer patients[75]. Substantial evidence now suggests that anti-tumor T cells recognize tumor somatic mutations, translated as single amino acid substitutions, as 'neoantigens'. These unique antigenic markers arise from numerous genetic changes, acquired somatically that are present exclusively in tumor (mutant) and not in normal (wild-type (WT)) cells[76]. Recent preclinical data indicate that these mutated proteins, upon processing and presentation in the context of MHC molecules expressed by antigen-presenting cells, can be recognized as 'non-self' by the immune system. Our previous work in murine sarcoma models was one of the first demonstrations of how somatic cancer mutations could be identified from massively parallel sequencing, and when considered in the context of MHC binding affinity, can predict tumor specific neoantigens[38]. A subsequent study further demonstrated that these neoantigens were the same epitopes

recognized by anti-PD1 and anti-CTLA4 checkpoint blockade therapies and that peptide vaccines comprising neoantigens could provide prophylactic effects. Several other studies have also characterized these neoantigens as being derived from somatically mutated genes in mouse [39] as well as in humans[40–43], and have shown that they can be recognized by T cells.

While checkpoint blockade therapies have achieved tremendous success in the clinic, patient-specific vaccines still meet a clinical need in those patients that either do not respond, develop resistance, or cannot tolerate the associated side effects of checkpoint blockade drugs. The main paradigm behind the development of cancer vaccines rests on the assumption that if the immune system is stimulated to recognize neoantigens, it may be possible to elicit the selective destruction of tumor cells. Vaccines incorporate these neoantigen peptides with the aim of enhancing the immune system's anti-tumor activity by selectively increasing the frequency of specific CD8+ T cells, and hence expanding the immune system's ability to recognize and destroy cancerous cells. This process is dependent on the ability of these peptides to bind and be presented by HLA class I molecules, a critical step to inducing an immune response and activating CD8+ T cells[77].

As we move from vaccines targeting 'shared' tumor antigens to a more 'personalized' medicine approach, *in silico* strategies are needed to first identify, then determine which somatic alterations provide the optimal neoantigens for the vaccine design. Ideally, an optimal strategy would intake mutation calls from massively parallel sequencing data comparisons of tumor to normal DNA, identify the neoantigens in the context of the patient's HLA alleles, and parse out a list of optimal peptides for downstream testing. At present, elements of this ideal strategy exist, but are not available as open source code to permit others

to adopt these methods into cancer care strategies. This manuscript describes one such approach, and provides a link to open source code for end users.

For example, to optimize identification and selection of vaccine neoantigens, several *in silico* epitope binding prediction methods have been developed[57,78–81]. These methods employ various computational approaches such as Artificial Neural Networks (ANN) and Support Vector Machines (SVM) and are trained on binding to different HLA class I alleles to effectively identify putative T cell epitopes.

There are also existing software tools (IEDB[82], EpiBot[83], EpiToolKit[84]) that compile the results generated from individual epitope prediction algorithms to improve the prediction accuracy with consensus methods or a unified final ranking. The current implementation of EpiToolKit (v2.0) also has the added functionality of incorporating sequencing variants in its Galaxy-like epitope prediction workflow (via its Polymorphic Epitope Prediction plugin). However, it does not incorporate sequence read coverage or gene expression information available from massively parallel sequencing datasets, nor can it compare the binding affinity of the peptide in the normal sample (WT) versus the tumor (mutant). Another multi-step workflow Epi-Seq[85] uses only raw RNA-Seq tumor sample reads for variant calling and predicting tumor-specific expressed epitopes.

We report herein an open source method called pVAC-Seq that we developed to address the critical need for a workflow that assimilates and leverages massively parallel DNA and RNA sequencing data to systematically identify and shortlist candidate neoantigen peptides from a tumor's mutational repertoire that could potentially be used in a personalized vaccine after immunological screening. This automated analysis offers the functionality to compare and

differentiate the epitopes found in normal cells against the neoepitopes specifically present in tumor cells for use in personalized cancer vaccines, and the flexibility to work with any user-specified list of somatic variants. Preliminary versions of this pipeline were applied in mouse models of cancer to identify expressed mutations in cancer cells and characterize tumor-specific mutant peptides that drive T cell-mediated tumor rejection in mice with MCA-induced sarcomas[38,39] . More recently, we used this pipeline in a proof-of-concept trial in melanoma patients, to identify the neoantigen peptides for use in dendritic cell-based personalized vaccines[71].

2.3 Methods

Our *in silico* automated pipeline for neoantigen prediction (pVAC-Seq) requires several types of data input from next-generation sequencing assays. First, the pVAC-Seq pipeline requires a list of non-synonymous mutations, identified by a somatic variant-calling pipeline. Second, this variant list must be annotated with amino acid changes and transcript sequences. Third, the pipeline requires the HLA haplotypes of the patient, which can be derived through clinical genotyping assays or *in silico* approaches. Having the above-mentioned required input data in-hand, pVAC-Seq implements three steps: performing epitope prediction, integrating sequencing-based information, and, lastly, filtering neoantigen candidates. The following paragraphs describe the analysis methodology from preparation of inputs to the selection of neoantigen vaccine candidates via pVAC-Seq (**Figure 2.1**).

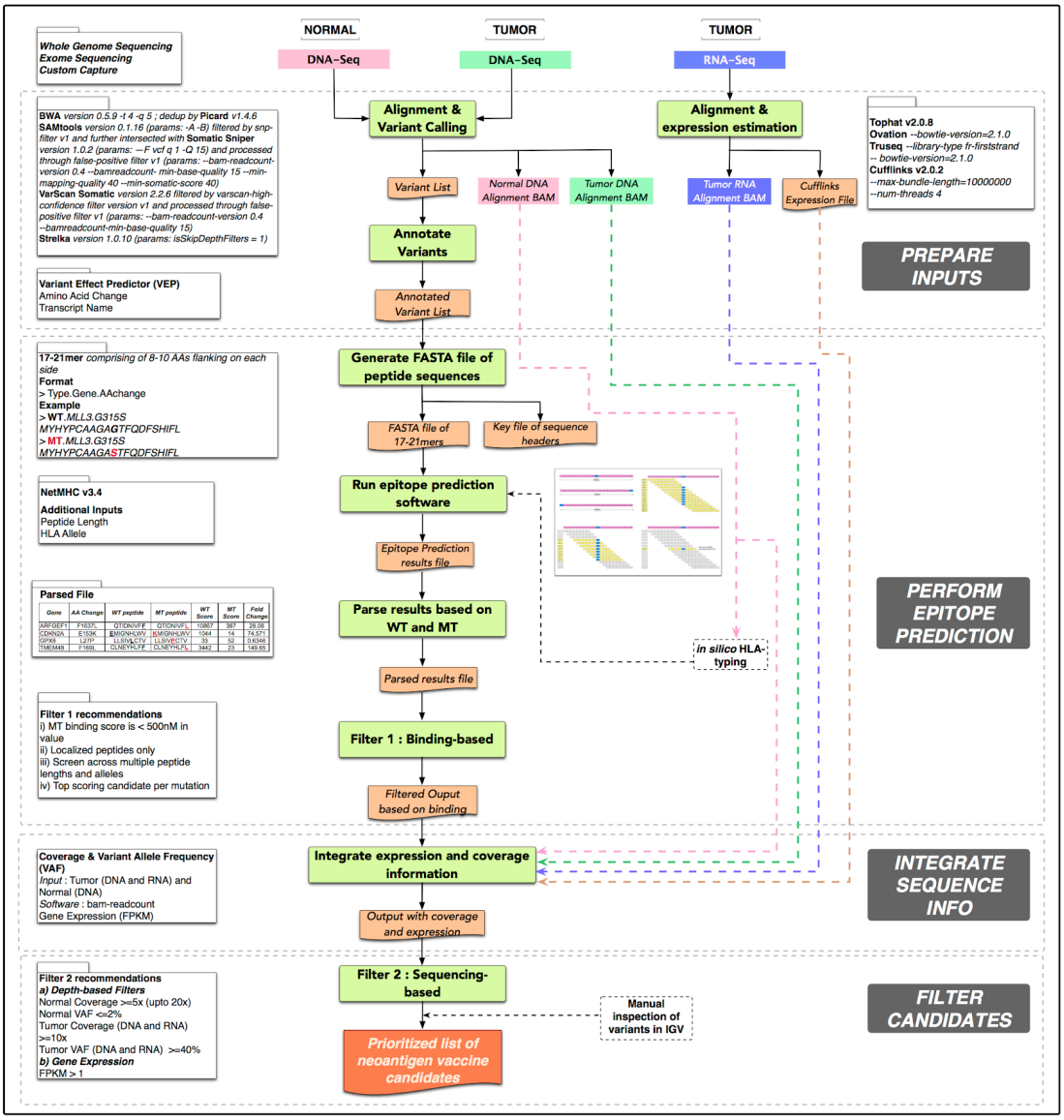


Figure 2.1. Overview of the pipeline pVAC-Seq*: This figure illustrates the methodological framework behind the pVAC-Seq pipeline. Starting with preparation of inputs, it consists of three main steps - epitope prediction, integration of sequencing information, and filtered candidate selection.*

2.3.1 Prepare input data: HLA typing, alignment, variant detection, and annotation

As described above, pVAC-Seq relies on input generated from the analysis of massively parallel sequencing data that includes annotated nonsynonymous somatic variants that have

been ‘translated’ into mutant amino acid changes, as well as patient-specific HLA alleles. Importantly, these data can be obtained from any appropriate variant calling, and annotation pipeline and HLA typing approach. Here, we outline our preparatory steps to generate these input data[71]. Somatic variant analysis of exome sequencing datasets was performed using the Genome Modeling System (GMS)[86] for alignment and variant calling. In brief, BWA (version 0.5.9)[87] was used for alignment with default parameters, except that the number of threads was set to 4 (-t 4) for faster processing, and the quality threshold for read trimming to 5 (-q 5). The resulting alignments were de-duplicated via Picard MarkDuplicates (version 1.46) [88].

In cases where clinically genotyped HLA haplotyping calls were not available, we used *in silico* HLA typing by HLAminer (version 1)[50] or by Athlates[49] . HLA typing was performed on the normal (peripheral blood mononuclear cells), rather than the tumor sample. Though the two software tools were >85% concordant in our test data (unpublished data), it is helpful to use both algorithms in order to break ties reported by HLAminer (see below).

I. HLAminer for *in silico* HLA-typing using Whole Genome Sequencing (WGS) data: When predicting HLA class I alleles from WGS data, we used HLAminer in *de novo* sequence alignment mode[89] by running the script HPTASRwgs_classI.sh, provided in the HLAminer download, with default parameters. (The download includes detailed instructions for customizing this script, and the scripts on which it depends, for the user’s computing environment.) For each of the three HLA loci, HLAminer reports predictions ranked in decreasing order by score, where ‘Prediction #1’ and ‘Prediction #2’ are the most likely alleles for a given locus. When ties were present for Prediction 1 or Prediction 2, we used all tied predictions for downstream neoepitope prediction. However, it should be noted that most epitope prediction algorithms, including

NetMHC[57,80] only work with an algorithm-specific subset of HLA alleles, so we are constrained to the set of NetMHC-compatible alleles. The current version NetMHC v3.4 supports 78 human alleles.

II. Athlates for *in silico* HLA-typing using exome sequence data: We diverged from the recommended Athlates protocol at two points: (1) We performed the alignment step, in which exome sequence data from the normal tissue sample are aligned against reference HLA allele sequences present in the IMGT/HLA database[90], using BWA with zero mismatches (params : bwa aln -e 0 -o 0 -n 0) instead of NovoAlign[91] with one mismatch. (2) In the subsequent step, sequence reads that matched, for example, any HLA-A sequence from the database were extracted from the alignment using bedtools[92] instead of Picard. This procedure is resource-intensive, and may require careful resource management. Athlates reports alleles that have a Hamming distance of at most 2 and meet several coverage requirements. Additionally, it reports 'inferred allelic pairs', which are identified by comparing each possible allelic pair to a longer list of candidate alleles using a Hamming distance-based score. We typically used the inferred allelic pair as input to subsequent steps in the neoepitope prediction pipeline.

After alignments (and optional HLA typing) were completed, somatic mutation detection was performed using the following series of steps (1) *Samtools* [93,94] mpileup v0.1.16 was run with parameters '-A -B' with default setting for the other parameters. These calls were filtered based on GMS 'snp-filter v1' and were retained if they met all of the following rules: (a) Site is greater than 10 bp from a predicted indel of quality 50 or greater; (b) The maximum mapping quality at the site is ≥40; (c) Fewer than three single-nucleotide variants (SNV) calls are present in a 10 bp window around the site; (d) The site is covered by at least three reads

and less than 1 × 109 reads; and (e) Consensus and SNP quality is ≥20. The filtered Samtools variant calls were intersected with those from *Somatic Sniper*[95] version 1.0.2 (params: -F vcf q 1 -Q 15), and were further processed through the GMS 'false-positive filter v1' (params: --bam-readcount-version 0.4 --bamreadcount-min-base-quality 15 --min-mapping-quality 40 --min-somatic-score 40). This filter used the following criteria for retaining variants: (a) ≥1% of variant allele support must come from reads sequenced on each strand; (b) variants must have ≥5% Variant Allele Fraction (VAF); (c) more than four reads must support the variant; (d) the average relative distance of the variant from the start/end of reads must be greater than 0.1; (e) the difference in mismatch quality sum between variant and reference reads must be less than 50; (f) the difference in mapping quality between variant and reference reads must be less than 30; (g) the difference in average supporting read length between variant and reference reads must be less than 25; (h) the average relative distance to the effective 3' end of variant supporting reads must be at least 0.2; and (i) the variant must not be adjacent to five or more bases of the same nucleotide identity (for example, a homopolymer run of the same base). (2) *VarScan Somatic* version 2.2.6[96,97] was run with default parameters and the variant calls were filtered by GMS filter 'varscan-high-confidence filter version v1'. The 'varscan-high-confidence v1' filter employed the following rules to filter out variants: (a) P value (reported by Varscan) is greater than 0.07; (b) Normal VAF is greater than 5%; (c) Tumor VAF is less than 10%; or (d) less than two reads support the variant. The remaining variant calls were then processed through false-positive filter v1 (params: --bam-readcount-version 0.4 --bamreadcount-min-base-quality 15) as described above. (3) *Strelka* version 1.0.10 [98](params: isSkipDepthFilters = 1).

Our GMS pipeline expects a matched normal sample for filtering out potentially rare germline variants. However, in the absence of a matched normal tissue, the dbSNP and 1000 Genome databases could be used for filtering these variants.

The consolidated list of somatic mutations identified from these different variant-callers was then annotated using our internal annotator as part of the GMS pipeline. This annotator leverages the functionality of the Ensembl database[99] and Variant Effect Predictor (VEP)[5].

We wish to emphasize that any properly formatted list of annotated variants can be used as input to subsequent steps in the pipeline. From the annotated variants, there are two critical components that are needed for pVAC-Seq: amino acid change and transcript sequence. Even a single amino acid change in the transcript arising from missense mutations can alter the binding affinity of the resulting peptide with the HLA class I molecule and/or recognition by the T cell receptor. Larger insertions and deletions like those arising from frameshift and truncating mutations, splicing aberrations, gene fusions, and so on may also result in potential neoantigens. However, for this initial version of pVAC-Seq, we chose to focus our analysis on only missense mutations.

One of the key features of our pipeline is the ability to compare the differences between the tumor and the normal peptides in terms of the peptide binding affinity. Additionally, it leverages RNA-Seq data to incorporate isoform-level expression information and to quickly cull variants that are not expressed in the tumor. To easily integrate RNA-Seq data, both transcript ID as well as the entire WT transcript amino acid sequence is needed as part of the annotated variant file.

2.3.2 Perform epitope prediction

One of the key components of pVAC-Seq is predicting epitopes that result from mutations by calculating their binding affinity against the HLA class I molecule. This process involves the following steps for effectively preparing the input data as well as parsing the output (**Figure 2.2**).

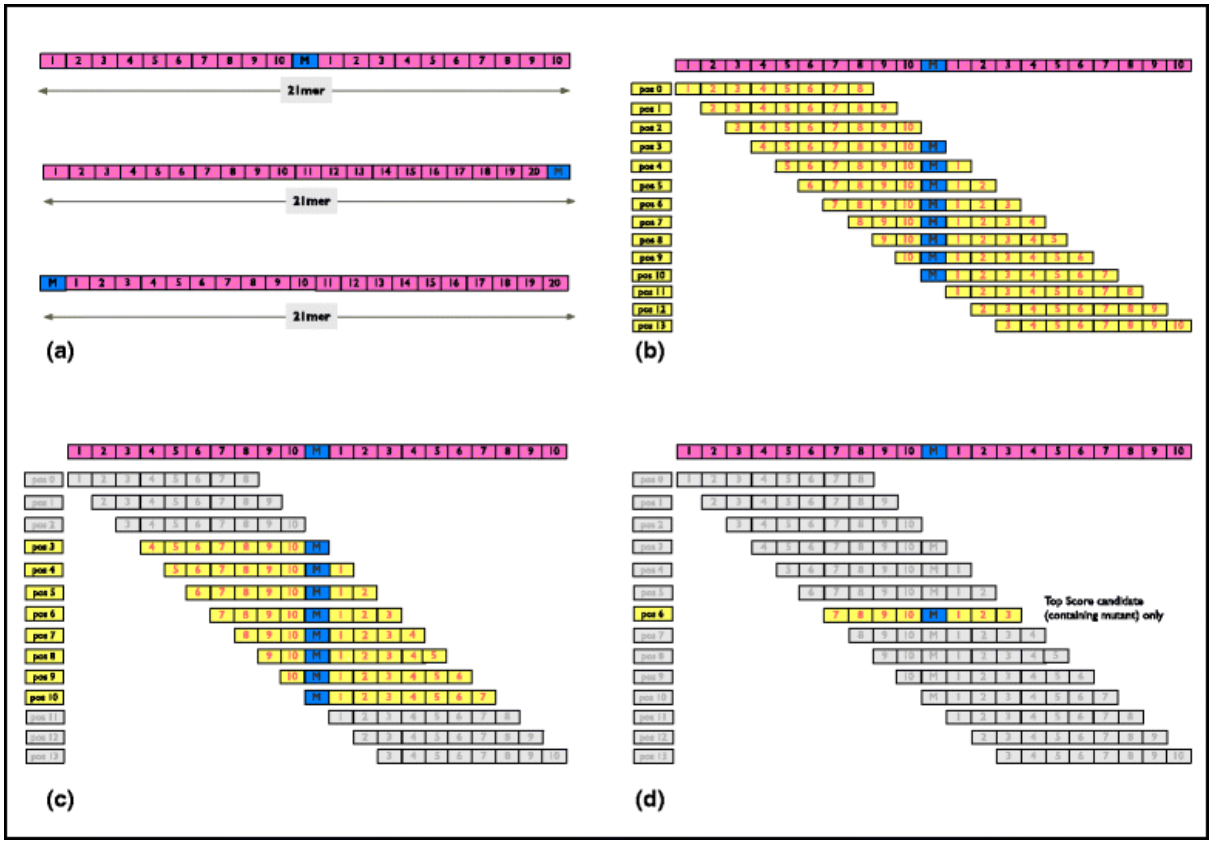


Figure 2.2 Generation of peptide sequences and filtering predicted epitope candidates. *a Amino acid FASTA sequence is built using 10 flanking amino acids on each side of the mutated amino acid. The preceding or succeeding 20 amino acids are taken if the mutation lies near the end or beginning of the transcript, respectively. b All predicted candidate peptides from epitope prediction software based on selected k-mer window size. c Only localized peptides (those containing the mutant amino acid) are considered to compare to WT counterpart. d The 'best candidate' (lowest MT binding score) per mutation is chosen across all specified k-mers and between all independent HLA allele types that were used as input*

Generate FASTA file of peptide sequences

Peptide sequences are a key input to the MHC binding prediction tool, and the existing process to efficiently compare the germline normal with the tumor is very onerous. To streamline the comparison, we first build a FASTA file that consists of two amino acid sequences per variant site: WT (normal) and mutant (tumor). The FASTA sequence is built using approximately eight to 10 flanking amino acids on each side of the mutated amino acid. However, if the mutation is towards the end or beginning of the transcript, then the preceding or succeeding 16 to 20 amino acids are taken, respectively, as needed, to build the FASTA sequence. Subsequently, a key file is created with the header (name and type of variant) and order of each FASTA sequence in the file. This is done to correlate the output with the name of the variant protein, as subsequent epitope prediction software strips off each FASTA header.

Run epitope prediction software

Previous studies[100,101] have shown that allele-specific epitope prediction software, such as NetMHC, perform slightly better when compared to pan-specific methods such as NetMHCpan[58,65,102] in case of well-characterized alleles due to availability of large amounts of training data. However, pan-specific methods could be beneficial in cases where there is limited peptide binding data for training, for arbitrary HLA molecules, or when predicting epitopes for non-human species. We do anticipate adding this support for additional softwares in upcoming versions of pVAC-Seq. To predict high affinity peptides that bind to the HLA class I molecule, currently only the standalone version of NetMHC v3.4 is supported. The input to this software is the HLA class I haplotype of the patient, determined via genotyping or using *in silico* methods, as well as the FASTA file generated in the previous step comprising mutated and WT 17-21-mer sequences. Typically, antigenic epitopes presented by HLA class I molecules can vary in length and are in the range of eight

to 11 amino acids (aa). Hence, we recommend specifying the same range when running epitope prediction software.

Parse and filter the output

Starting with the output list of all possible epitopes from the epitope prediction software, we apply specific filters to choose the best candidate mutant peptides. First, we restrict further consideration to strong- to intermediate-binding peptides by focusing on candidates with a mutant (MT) binding score of less than 500 nM. Second, epitope binding calls are evaluated only for those peptides that contain the mutant amino acid (localized peptides). This filter eliminates any WT peptides that may overlap between the two FASTA sequences. Our workflow enables screening across multiple lengths and multiple alleles very efficiently. If predictions are run to assess multiple epitope lengths (for example, 9-mer, 10-mer, and so on), and/or to evaluate all patient's HLA-A, -B, and -C alleles, we review all localized peptides and choose the single best binding value representative across lengths (9 aa, 10 aa, and so on) based on lowest binding score for MT sequence. Furthermore, we choose the 'best candidate' (lowest MT binding score) per mutation between all independent HLA alleles that were used as input. Additionally, in the output file, the WT peptide binding score is provided. Although this score may not directly affect candidate choice or immunogenicity, end users may find this comparative information useful.

2.3.3 Integrate expression and coverage information

We subsequently apply several filters to ensure we are predicting neoantigens that are expressed as RNA variants, and that have been predicted correctly based on coverage depth in the normal and tumor tissue datasets. We have found that gene expression levels from RNA-Seq data, measured as fragments per kilobase of exon per million reads mapped

(FPKM), provide a good method to filter only the expressed transcripts. We used the tuxedo suite - Tophat[103,104] and Cufflinks[105] - as part of the GMS to align RNA-Seq data and subsequently infer gene expression for our in-house sequencing data. Depending on the type of RNA prep kit, Ovation® RNA-Seq System V2 (NuGEN Technologies, Inc., San Carlos, CA, USA) or TruSeq Stranded Total RNA Sample Prep kit (Illumina, Inc., San Diego, CA, USA) used, Tophat was run with the following parameters: Tophat v2.0.8 '--bowtie-version = 2.1.0' for Ovation, and '--library-type fr-firststrand --bowtie-version = 2.1.0' for Truseq. For Ovation data, prior to alignment, paired 2 × 100bp sequence reads were trimmed with Flexbar version 2.21[106] (params: --adapter CTTTGTGTTTGA --adapter-trim-end LEFT --nono-length-dist --threads 4 --adapter-min-overlap 7 --maxuncalled 150 --min-readlength 25) to remove single primer isothermal amplification adapter sequences. Expression levels (FPKM) were calculated with Cufflinks v2.0.2 (params: --max-bundle-length = 10000000 --num-threads 4).

For selecting unique vaccine candidates, targeting the best 'quality' mutations is an important factor for prioritizing peptides. Sequencing depth as well as the fraction of reads containing the variant allele (VAF) are used as criteria to filter or prioritize mutations. This information was added in our pipeline via bam-readcount[107]. Both tumor (from DNA as well as RNA) and normal coverage are calculated along with the VAF from corresponding DNA and RNA-Seq alignments.

2.3.4 Filter neoepitope candidates

Since manufacturing antigenic peptides is one of the most expensive steps in vaccine development and efficacy depends on selection of the best neoantigens, we filter the list of predicted high binding peptides to the most highly confident set, primarily with expression

and coverage based filters. The pVAC-Seq pipeline permits user-specified filters, and we encourage new users to experiment with these cutoffs in order to tailor the pipeline to their input data and analysis needs. We employ the following filters: (a) Depth based filters: We filter out any variants with normal coverage <=5× and normal VAF of >=2%. The normal coverage cutoff can be increased up to 20× to eliminate occasional misclassification of germline variants as somatic. Similarly, the normal VAF cutoff can be increased based on suspected level of contamination by tumor cells in the normal sample.

For tumor coverage from DNA and/or RNA, a cutoff is placed at >=10× with a VAF of >=40%. This ensures that neoantigens from the founder clone in the tumor are included, but the tumor VAF can be lowered to capture more variants, which are less likely to be present in all tumor cells. Alternatively, if the patients are selected based on a pre-existing disease-associated mutation such as BRAF V600E in the case of melanoma, the VAF of the specific presumed driver mutation can be used as a guide for assessing clonality of other mutations. Also, other known driver mutations such as KRAS G12/G13 or NRAS Q61 may be used to determine purity, and to subsequently adjust the VAF filters to target founder clone mutations. (b) Expression based filters: As a standard, genes with FPKM values greater than zero are considered to be expressed. We slightly increase this threshold to 1, to eliminate noise. Alternatively, we analyze the FPKM distribution (and the corresponding standard deviation) over the entire sample, to determine the sample-specific cutoffs for gene expression. Spike-in controls may also be added to the RNA-Seq experiment to assess quality of the sequencing library and to normalize gene expression data. Since alternative splicing can give rise to multiple transcripts that encompass the variant residue, optionally, all these transcripts could be included in analysis during the annotation step. However, one should be careful as this could potentially give rise to transcripts that do not include the variant. Also,

long transcripts or transcripts with high G/C content might show some bias if RNA-CapSeq is used but in our experience are generally well represented. The primary goal of using RNA-(Cap)Seq data in our method is to address to questions of primary importance: (1) is the gene expressed at a reasonably high level (for example, FPKM >1); and (2) is the variant allele expressed in the RNA-seq fragment population.

This filtered list of mutations is manually reviewed via visual inspection of aligned reads in a genome viewer like IGV[108],[109] to reduce the retention of obvious false positive mutations.

2.3.5 Dataset

To demonstrate the workings of our *in silico* pVAC-Seq pipeline, we applied it to four metastatic melanoma patients, the clinical results for three of whom were described previously[71]. In brief, there were three patients (MEL21, MEL38, MEL218) with stage III resected cutaneous melanoma, all of whom had received prior treatment with ipilimumab, and one patient (MEL69) with stage IV cutaneous melanoma. All four patients were enrolled in a phase 1 vaccine clinical trial (NCT00683670, BB-IND 13590) employing autologous, functionally mature, interleukin (IL)-12p70-producing dendritic cells (DC). Informed consent for genome sequencing and data sharing was obtained for all patients on a protocol approved by the Institutional Review Board of Washington University. We performed genomic analysis of their surgically excised tumors to select candidates for the personalized DC vaccine. Three of these patients (MEL21, MEL38, MEL69) had multiple metachronous tumors. Exome sequencing as well as RNA-CapSeq was performed for each of these tumors, and their corresponding matched normal tissue. The raw exome and transcriptome sequence data are available on the Sequence Read Archive database: Bioproject PRJNA278450, and corresponding dbGaP accession: phs001005.

2.4 Results and Discussion

Since melanoma patients harbor hundreds of mutations, it can be challenging to filter down and target the best set of potentially immunogenic neoantigens for vaccine design. For each of the four metastatic melanoma patients, we used the annotated list of SNVs generated using the GMS strategy described above, and analyzed them via our pVAC-Seq pipeline. As mentioned earlier, for the demonstration of this workflow, amino acid changes resulting from only missense mutations were considered for analysis. **Table 2.1** shows the breakdown of these SNVs described previously[71] and the data generated in subsequent steps through our workflow, leading to a high-confidence list of neoepitopes. As part of our local workflow, NetMHC v3.4 was used as the epitope prediction software to generate HLA class I restricted epitopes.

	MEL21			**MEL38**			**MEL218**	**MEL69**	
	LN (2011)	**Skin (2012)**	**Skin (2013)**	**Axilla (2012)**	**Breast (2013)**	**AbW all (2013)**	**LN (2005)**	**Skin / Limb (2013)**	**Skin / Scalp (2013)**
Total SNVs	702	838	1099	359	402	385	695	256	282
Missense SNVs	443	515	598	219	247	238	437	141	162
21-mer FASTA entries *(WT & MT)*	856	1004	1002	424	482	462	850	272	314

Raw NETMHC output *(9-mers)*	11152 *2 (HLA-A02:01, HLA-A01:01)	13072 *2 (HLA-A02:01, HLA-A01:01)	13044 *2 (HLA-A02:01, HLA-A01:01)	5512*3 (HLA-A02:01, HLA-A31:01, HLA-B07:02)	6270*3 (HLA-A02:01, HLA-A31:01, HLA-B07:02)	6010 *3 (HLA-A02:01, HLA-A31:01, HLA-B07:02)	11050*3 (HLA-A02:01, HLA-A03:01, HLA-B44:02	3542 *2 (HLA-A02:01, HLA-A11:01)	4088 *2 (HLA-A02:01, HLA-A11:01)
Parsed NetMHC output *(comparing WT with MT)*	3796 *2 (HLA-A02:01, HLA-A01:01)	4465 *2 (HLA-A02:01, HLA-A01:01)	4458 * 2 (HLA-A02:01, HLA-A01:01)	1871*3 (HLA-A02:01, HLA-A31:01, HLA-B07:02)	2131*3 (HLA-A02:01, HLA-A31:01, HLA-B07:02)	2042 *3 (HLA-A02:01, HLA-A31:01, HLA-B07:02)	3770 *3 (HLA-A02:01, HLA-A03:01, HLA-B44:02	1217 *2 (HLA-A02:01, HLA-A11:01)	1395 *2 (HLA-A02:01, HLA-A11:01)
Filter 1: Binding based	110	121	144	103	112	111	161	50	65
HLA-A02:01 candidates only	79	96	111	52	48	46	93	25	34
Filter 2: Manually reviewed HLA-A02:01 candidates (*Exome plus RNA-Seq*)	11	11	12	14	16	16	24	6	12
Filter 3: Experimentally tested	16			14			18	12	
Filter 4: Vaccine tested	7			7			7	7	
Immunogenicity	3			3			3	3	

Table 2.1 Summary of predicted epitope candidates through pVAC-Seq pipeline

The table illustrates the number of raw candidates predicted by NetMHC, and the parsing and filtering strategies applied thereafter to the final list of neoantigen candidates. These candidates were then communicated to our vaccine design collaborators who evaluated this list by patient-specific immunological assays (Filters 3 & 4)[71]

As is evident from **Table 2.1**, there were multitudes of epitopes reported by NetMHC v3.4 in its raw format. This number increased even further with the addition of each HLA class I allele. Using pVAC-Seq, and its recommended thresholds for filtering (binding and coverage-based), we were able to produce a more reasonable list of high affinity HLA class I binding neoantigen candidates for experimental validation.

These candidate neoantigens were experimentally tested in binding assays and those with confirmed binding to HLA class I restricting molecules were incorporated in the vaccine formulation[71]. Since all of these patients harbor the BRAF V600E mutation, we used its VAF in each sample as a comparative control of tumor purity and clonality. Integration of variant coverage information from Exome and RNA-Seq (VAF), as well as mutant expression information (FPKM), provided additional information needed to make an informed decision on the number and identity of peptides to include in each patient-specific vaccine (**Figure 2.3**, **Figure 2.4, Figure 2.5,** and **Figure 2.6**).

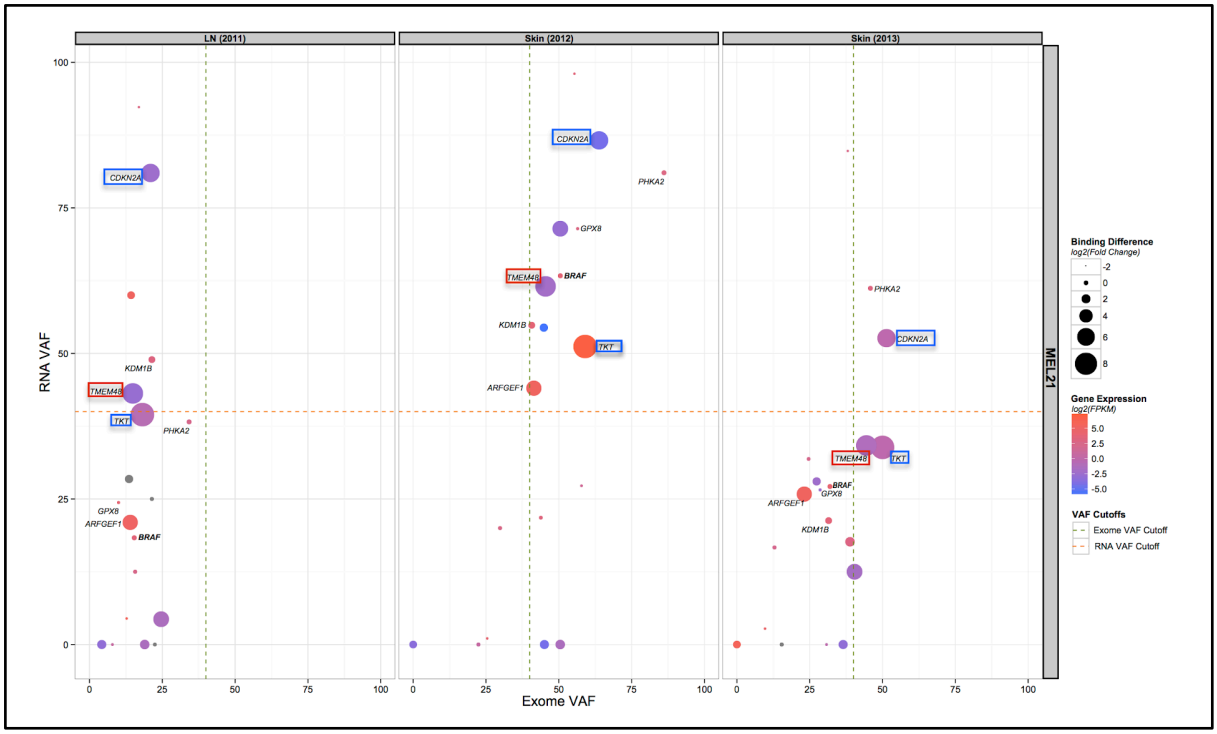


Figure 2.3 Landscape of filtered neoantigen candidates in MEL21. *This figure illustrates the landscape of neoantigen vaccine candidates in patient MEL21 after being prioritized using the pVAC-Seq pipeline. The points represent the overall sequencing information: exome and RNA VAFs, gene expression in terms of log2 FPKM value, as well log2 fold change, calculated as the ratio of WT binding affinity over mutant binding affinity. Recommended exome and RNA VAF cutoffs are also indicated. Candidates that were incorporated in the vaccine are labeled based on the genes containing these somatic mutations. Red boxes depict naturally occurring (that is, pre-existing T cell response) and blue boxes denote vaccine-induced neoantigens that were recognized by T cells. Since BRAF was used as a guide for assessing clonality of other mutations, it is also shown in each of three metachronous tumors (from the same patient)*

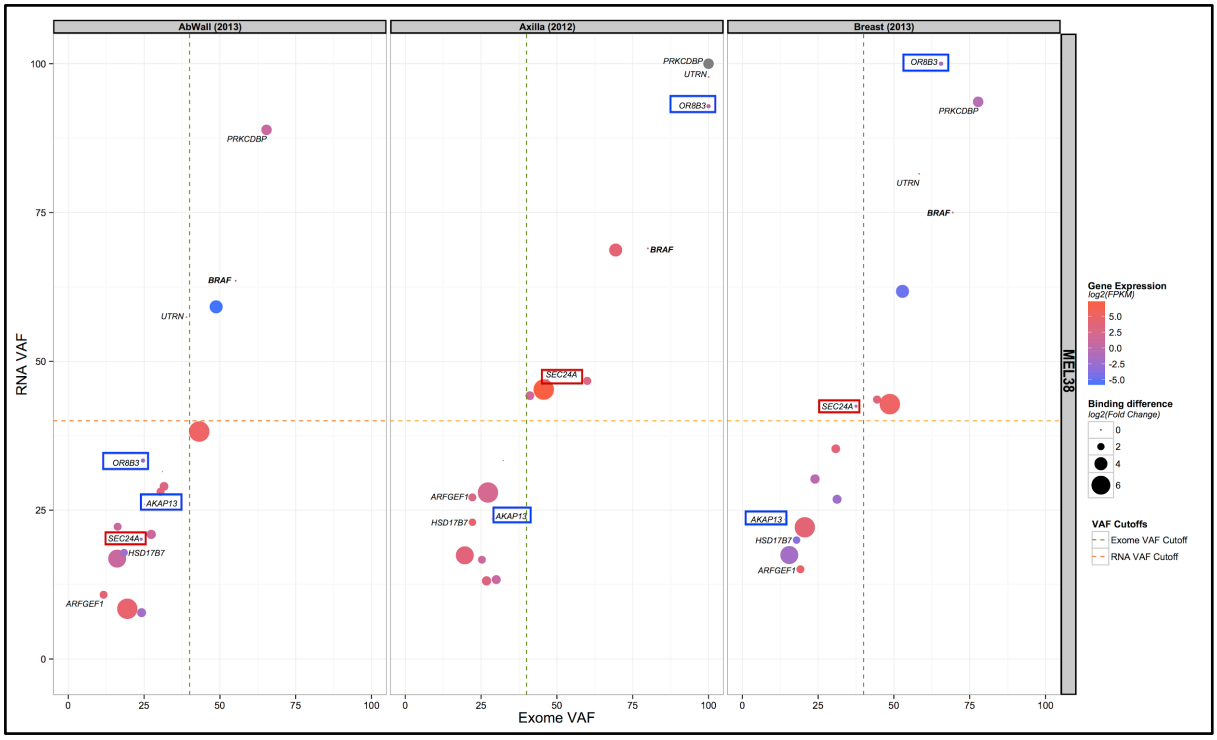


Figure 2.4 Landscape of filtered neoantigen candidates in MEL38 *This figure illustrates the landscape of neoantigen vaccine candidates in patient MEL38 after being prioritized using the pVAC-Seq pipeline. The points represent the overall sequencing information: exome and RNA VAFs, gene expression in terms of log2 FPKM value, as well log2 fold change, calculated as the ratio of WT binding affinity over mutant binding affinity. Recommended exome and RNA VAF cutoffs are also indicated. Candidates that were incorporated in the vaccine are labeled based on the genes containing these somatic mutations. Red boxes depict naturally occurring (that is, pre-existing T cell response) and blue boxes denote vaccine-induced neoantigens that were recognized by T cells. Since BRAF was used as a guide for assessing clonality of other mutations, it is also shown in each of three metachronous tumors (from the same patient).*

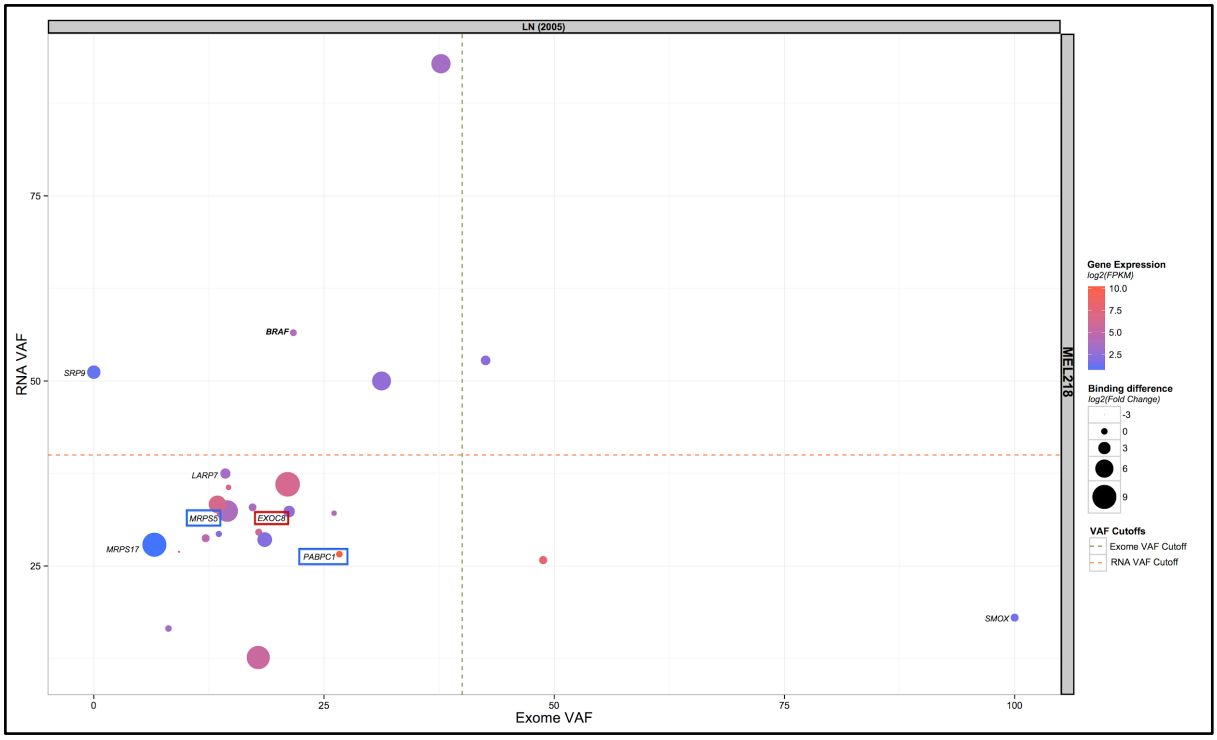


Figure 2.5 Landscape of filtered neoantigen candidates in MEL218 *This figure illustrates the landscape of neoantigen vaccine candidates in patient MEL218 after being prioritized using the pVAC-Seq pipeline. The points represent the overall sequencing information: exome and RNA VAFs, gene expression in terms of log2 FPKM value, as well log2 fold change, calculated as the ratio of WT binding affinity over mutant binding affinity. Recommended exome and RNA VAF cutoffs are also indicated. Candidates that were incorporated in the vaccine are labeled based on the genes containing these somatic mutations. Red boxes depict naturally occurring (that is, pre-existing T cell response) and blue boxes denote vaccine-induced neoantigens that were recognized by T cells. Since BRAF was used as a guide for assessing clonality of other mutations, it is also shown.*

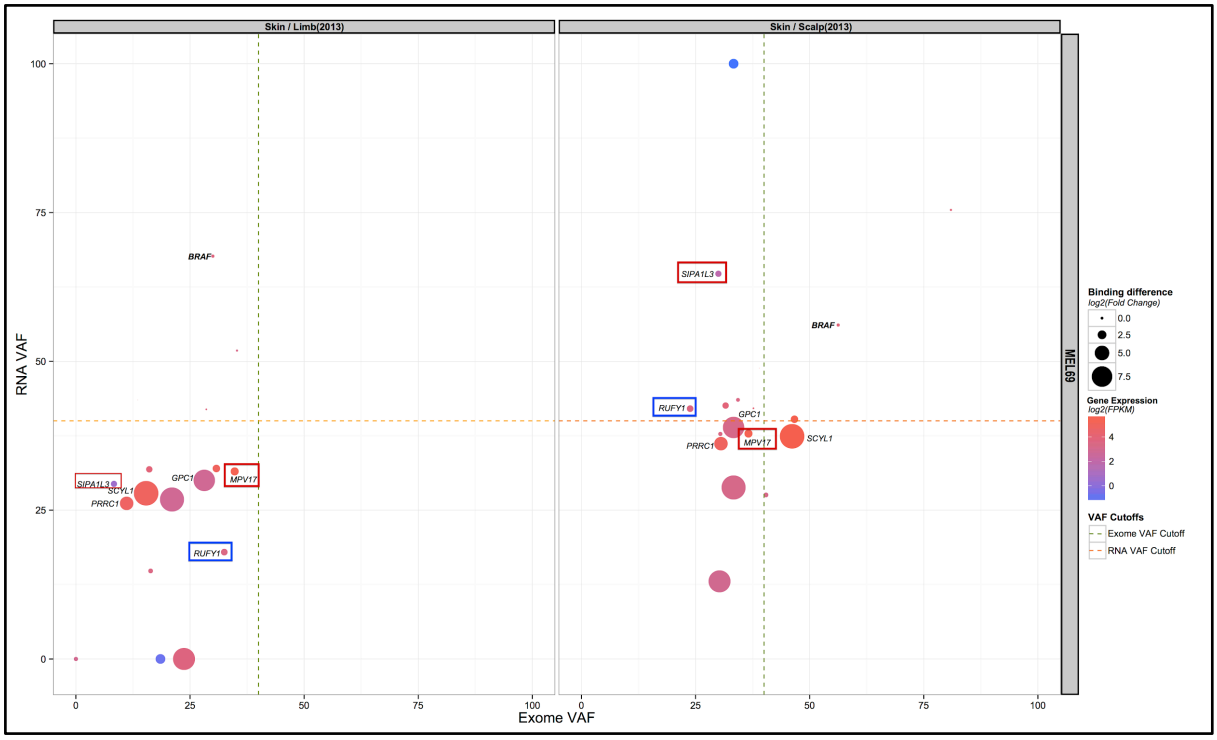


Figure 2.6 Landscape of filtered neoantigen candidates in MEL69. *This figure illustrates the landscape of neoantigen vaccine candidates in patient MEL69 after being prioritized using the pVAC-Seq pipeline. The points represent the overall sequencing information: exome and RNA VAF cutoffs, gene expression in terms of log2 FPKM value, as well log2 fold change, calculated as the ratio of WT binding affinity over mutant binding affinity. Recommended exome and RNA VAFs are also indicated. Candidates that were incorporated in the vaccine are labeled based on the genes containing these somatic mutations. Red boxes depict naturally occurring (that is, pre-existing T cell response) and blue boxes denote vaccine-induced neoantigens that were recognized by T cells. Since BRAF was used as a guide for assessing clonality of other mutations, it is also shown in both the metachronous tumors*

As shown, if existing epitope prediction software tools were solely used to generate neoantigen predictions in these patients, it would have been challenging to integrate the filters as well as the important digital sequencing metrics that ultimately determined the 'quality' of these candidates. By implementing the novel methods reported in this manuscript, we were able to rapidly streamline the screening and identification of a smaller number of potentially immunogenic neoepitopes within the landscape of all neoepitopes. This method can be further extended to include other genomic alterations such as frame-shift insertions and deletions, splicing aberrations, and gene fusions, which may in some cases cause larger

changes in epitope binding affinities. We are currently testing approaches to include binding predictions from frame-shift insertions and deletions by incorporating VEP annotation, and once tested, will be adding this functionality to the github repository for pVAC-Seq. By expanding the focus from just somatic point mutations to the entire neoantigen landscape, it may also be possible to better assess whether neoantigen load itself can serve as a biomarker for prediction of checkpoint blockade response.

2.5 Conclusions

The current regimen for predicting and screening neoantigens from sequencing data is laborious and involves a large number of intermediate steps such as creating FASTA files, running the prediction algorithms (most of the time online), and filtering output for high binding affinity candidates. Our flexible, automated *in silico* workflow, pVAC-Seq, provides higher efficiency and faster turnaround by automating many of these steps. This approach should help to evaluate tumor-specific neoepitopes in a much-reduced time, thereby increasing its applicability for clinical use. As we learn from ongoing early mouse and human trials, the methods developed will help optimize the composition of personalized cancer vaccines with high precision and will expedite vaccine design to address growing clinical demand.

2.6 Authors and Contributions

Jasreet Hundal[1], Beatriz M. Carreno[2], Allegra A. Petti[1], Gerald P. Linette[2] , Obi L. Griffith[1,2,4,5] , Elaine R. Mardis[1,3,4,5,6*], Malachi Griffith[1,4,5]

[1]McDonnell Genome Institute, Washington University School of Medicine, St. Louis, MO;

[2]Department of Medicine, Division of Oncology, Washington University School of Medicine, St. Louis, MO;

[3]Department of Medicine, Division of Genomics and Bioinformatics, Washington University School of Medicine, St. Louis, MO;

[4]Department of Genetics, Washington University School of Medicine, St. Louis, MO;

[5]Siteman Cancer Center, Washington University School of Medicine, St. Louis, MO;

[6]Department of Molecular Microbiology, Washington University School of Medicine, St. Louis, MO;

*Corresponding author: emardis@wustl.edu

JH was involved in all aspects of this study including designing and developing the methodology, analyzing and interpreting data, and writing the manuscript, with input from MG, OLG, GPL, and ERM. BMC participated in the design of the study, interpreting the data, and performing immunological and vaccine experiments, and participated in writing the manuscript. AAP tested and developed HLA-typing methods and was involved in writing the manuscript. GPL, ERM, and MG oversaw all the work performed and planned experiments. All authors read and approved the final manuscript.

2.7 Acknowledgements

We are grateful for creative and computational input from Zachary L. Skidmore, Susanna Siebert, Todd N. Wylie, Jason R. Walker, and Chris A. Miller. We thank Dr. Robert D. Schreiber for his expertise and guidance on foundational mouse models work. Dr. William E.

Gillanders provided important scientific input to the pipeline development work. MG was supported by the National Human Genome Research Institute (K99 HG007940). OLG was supported by the National Cancer Institute (K22 CA188163). BC, GPL, and JH were supported by the National Cancer Institute (R21 CA179695). ERM was supported by the National Cancer Institute (R21 CA179695) and the National Human Genome Research Institute (NIH NHGRI U54 HG003079). AAP was supported by the National Human Genome Research Institute (NHGRI U54 HG003079).

Chapter 3: Accounting for proximal variants improves neoantigen prediction

Hundal, J. et al. Accounting for proximal variants improves neoantigen prediction. Nat. Genet. doi:10.1038/s41588-018-0283-9 (2018).

3.1 Introduction

Over the past two decades, approaches to identify and screen for antigens, both self and non-self, have evolved rapidly[32,110]. This is due in part to advances in sequencing technologies, in the accuracy of algorithmic identification of somatic variants, and in computational modeling to predict the binding affinity of the resulting novel, tumor-specific peptides to major histocompatibility complex (MHC) molecules[45]. Thus, current immunogenomic approaches can identify somatic variants that give rise to tumor-specific mutant antigens or 'neo'-antigens and evaluate their ability to bind to MHC Class I and Class II molecules[45].

Typically, to evaluate strong-binding neoantigens from genomic sequencing data, the raw sequencing reads from tumor and normal DNA libraries are aligned to the human reference genome, and somatic variants are identified by comparison of tumor to normal read alignments. The resulting somatic variants of interest (SVOI) are then annotated to predict protein sequence changes and to infer possible neoantigenic peptides. Individual neoantigenic peptides are selected by sliding an amino acid window (usually 8-11-mers) across the variant position to consider each possible 'register'. These peptides are assessed using various algorithms to predict binding affinity to MHC and determine the strongest binding epitopes. These predicted neoantigenic peptides are prioritized as we have previously described[111]. The cancer vaccine design process, from read alignment to variant calling and neoantigen prediction typically assumes the reference genome sequence surrounding each somatic variant is representative of the patient's genome sequence.

However, any sequence variant proximal to an SVOI in the patient's genome that differs from the human reference may alter the amino acid sequence of the resulting peptide (note,

proximal is defined here as 'situated close to' or 'nearby', not the classic genetics meaning of 'closer to the centromere'). Existing pipelines that are used for computational prediction of neoantigens from sequencing data, such as MuPeXI[112] and pVAC-Seq[111], do not explicitly incorporate patient-specific nearby germline or somatic variants (collectively referred to as 'proximal variants' hereafter) into the peptide sequence considered in neoantigen prediction. Some pipelines such as Vaxrank[113] infer the coding sequence from assembly of tumor RNA reads, thus accounting for both somatic and germline variants implicitly, but this is largely dependent on the availability of RNA-Seq data. Failing to account for patient-specific nearby germline or somatic variants (i.e. proximal variants) could impact the efficacy of a vaccine, possibly resulting in immunization with incorrect peptides or failure to identify highly neoantigenic peptides.

To investigate these possibilities, we identified somatic and germline variants proximal to SVOIs in a data set of tumor sequencing studies representing different tissue sites and mutational loads. For this analysis, given that the upper-bound for the length of MHC-binding peptides (accounting for both Class I and Class II) is typically considered to be 30 amino acids[114,115] we chose a nucleotide window of 89 bp upstream and downstream of each SVOI in which to identify relevant proximal variants (**Methods**). We limited our analysis to only include missense proximal variants and SVOIs. We then incorporated these proximal variants in the final peptide sequences (proximal variant correction; PVC) and re-evaluated the resulting peptide set using our neoantigen prediction pipeline (pVAC-Seq)[111]. Our results suggest that taking individual proximal variation into account can have a significant effect on the accuracy of neoantigen selection, resulting in a more personalized vaccine design.

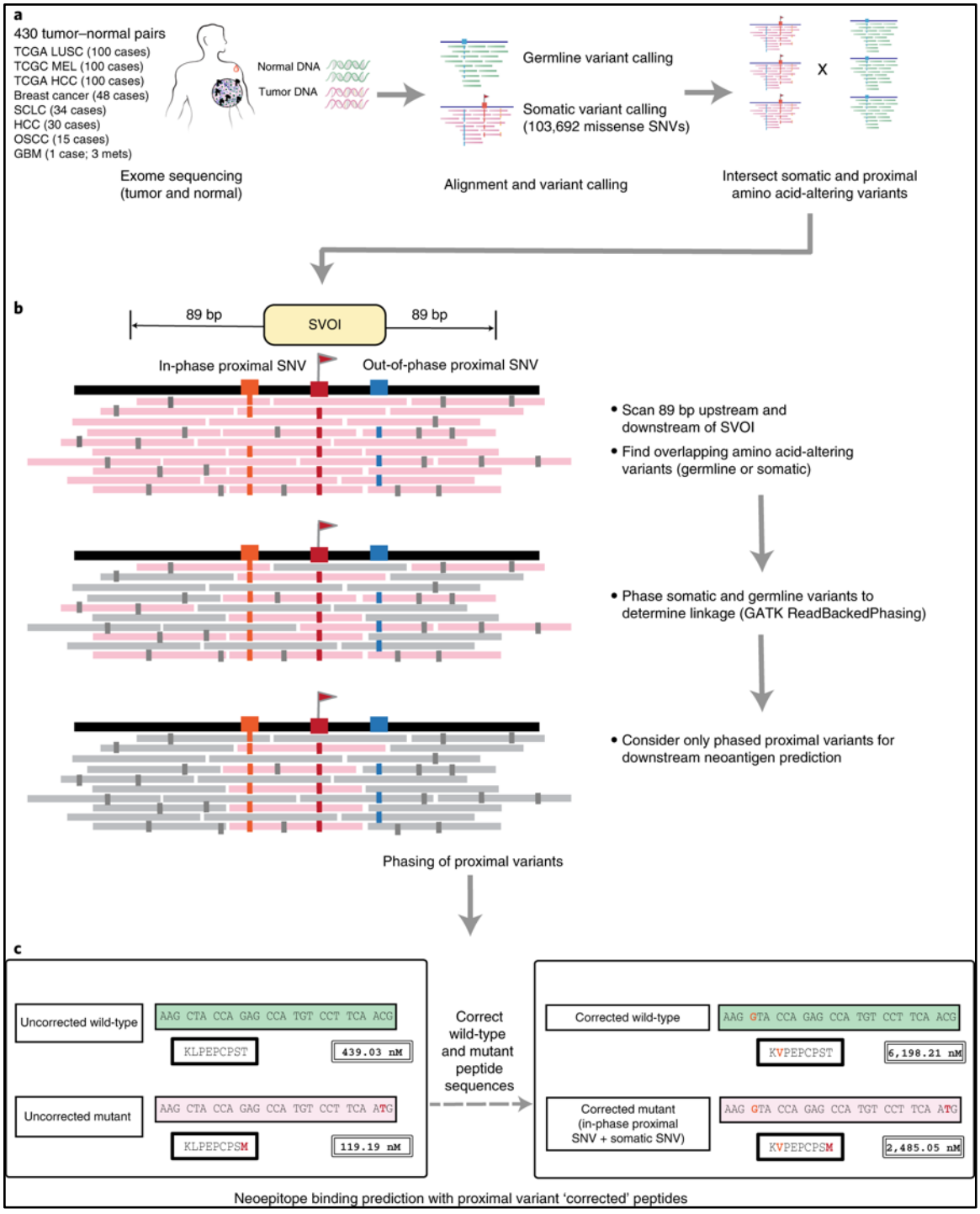


Figure 3.1: Overview of the pipeline for proximal variant correction *The steps required for incorporating and assessing the impact of proximal variants on neoantigen binding prediction are depicted as a flow diagram. There are three main steps. (A) Alignment and variant calling of matched tumor (pink) and normal (green) sequencing data. (B) Phasing of proximal somatic and germline variants: The pink bars represent the tumor sequence reads, with mismatches/sequencing errors shown in small gray rectangles. For a somatic variant of interest (SVOI; labeled with a red flag), we scan 89 bp on either side to assess for proximal germline or somatic SNVs (labeled with blue and orange boxes). These proximal variants are then phased together to determine linkage. Only proximal variants that are in phase (orange*

box) with the SVOI (red box) are considered for downstream neoantigen analysis. Other (out-of-phase) proximal variants (blue box) are ignored. (C) Neoantigen binding predictions are then assessed after performing proximal variant correction (PVC). The left panel shows the 'uncorrected' wildtype and mutant peptides along with their respective binding scores for a single SVOI example. The right panel shows PVC ('corrected') peptides and scores for this SVOI.

3.2 Methods

3.2.1 Sequence data alignment and variant calling

To investigate the prevalence of proximal variants (germline SNPs or somatic variants), we analyzed publicly available sequencing data from the TCGA as well as datasets generated in-house, altogether representing seven different tissue sites. These data sets were chosen to adequately represent low, medium and high mutational burden tumors.

Analysis of in-house whole genome/exome sequencing datasets was performed as previously described[111,86,116]. Briefly, raw sequencing reads from both the tumor and normal were aligned to the human reference genome sequence (either GRCh37 or GRCh38) using BWA[117], then merged and deduplicated using Picard (see URLs). A combination of three or four different variant callers was used to identify somatic variants by comparison of tumor and normal variant calls: Samtools[93], Sniper[95], Strelka[98], and VarScan[96,97]. These variants were filtered as previously described[118,119] and then manually reviewed using IGV per the standard operating procedures[120] to obtain a list of high confidence variant calls. On average, 80% of the filtered variants passed manual review. Germline variant analysis of the normal samples was performed using Samtools.

For the TCGA data, aligned tumor and normal BAMs from BWA (version 0.7.12-r1039) as well as somatic variant calls from VarScan2 (in VCF format) were downloaded from the Genomic Data Commons (GDC). We restricted our analysis to only consider 'PASS' variants in these VCFs that are higher confidence than the raw set. Since TCGA does not provide germline variants, we used GATK's HaplotypeCaller to perform germline variant calling using default parameters. These calls were refined using VariantRecalibrator in accordance with GATK Best Practices[121].

For this study, we restricted the variant calls to only include missense SNVs, in both- TCGA as well as in-house datasets.

3.2.2 Phasing of variants to assess linkage

Somatic and germline missense variant calls from each sample were combined using GATK's CombineVariants, and the variants were subsequently phased using GATK's ReadBackedPhasing algorithm.

In silico HLA-typing

OptiType[48] was used to perform *in silico* HLA typing for the in-house samples. For the datasets downloaded from TCGA, existing *in silico* HLA typing information was obtained from The Cancer Immunome Atlas (TCIA[122]) database.

3.2.3 Choosing an appropriate window for neoantigen analysis

Due to the absence of patient-specific HLA Class II typing information, we limited our neoantigen binding prediction analysis to MHC Class I, though we believe that the Class II peptides are also important in contributing to immunogenicity. Hence, our nucleotide window

was chosen such that it encompasses both Class I and Class II MHC peptide lengths, to demonstrate the prevalence of proximal variants within that genomic region. Most strong-binding Class I MHC peptides are around 8-11 amino acids in length. There is no length restriction on Class II MHC peptides due to an open binding groove, and longer peptide lengths are much more common, typically 13-25-mers[123] but peptides as long as 30-mer have been reported[114,115]. The majority (99.2%) of human linear T-cell epitopes with MHC class II restriction currently reported in IEDB [82]are 8-30-mers. To identify the best binding 30-mer around a missense variant of interest, one would ideally scan 29 amino acids upstream and downstream of the mutant (MT) amino acid, hence a window of 59 amino acids. At the nucleotide level, this corresponds to 87 nucleotides. Given that the frame of the missense mutation is not always known, we allow for 2 extra bases leading to a window size of 89 nucleotides on each side of the SVOI.

The appropriate nucleotide window for any peptide length can be calculated using this formula: ((peptide length -1)*3)+2.

3.2.4 Corrected neoantigen binding prediction using pVACtools

For each sample, the phased variant calls as well as the somatic variant calls were annotated using Variant Effect Predictor (VEP[4]), specifically using the Downstream plugin as well as the custom Wildtype plugin, available via pVACtools (see URLs) . To evaluate the effect of relevant nearby variants on neoantigen identification, we re-assessed the binding affinities of the neoantigens with the corrected mutant peptide sequence (**Figure 3.1C**), using NetMHCv4.0[56,57] via an updated version of the pVACtools software. This version takes as input the VEP-annotated phased VCF file of somatic and germline variants, in addition to the existing VEP-annotated somatic VCF.

3.2.5 Calculating False Discovery and False Negative Rates

To calculate FNR and FDR, we first determined the number of weak binders before PVC that were falsely omitted (false negatives (FN)), as well as the number of peptides that were identified as strong binders before PVC, but whose sequence (MTpeptide) was altered due to a proximal variant and which were thus incorrectly considered during neoantigen selection (false positives (FP)). We also calculated the number of peptides that were identified strong binders before correction and remained unaltered by proximal variants (true positives (TP)).

$$FN: (MTscore_{uncorrected} > 500\,nM) \wedge (MTscore_{corrected} < 500\,nM)$$
$$FP: (MTscore_{uncorrected} < 500\,nM) \wedge (MTpeptide\_corrected \neq MTpeptide\_uncorrected)$$
$$TP: (MTscore_{uncorrected} < 500nM) \wedge (MTpeptide\_corrected \equiv MTpeptide\_uncorrected)$$

$$FNR = \frac{FN}{FN + TP}$$

$$FDR = \frac{FP}{FP + TP}$$

The FNR is then defined as the number of false negatives divided by the number of false negatives plus the number of true positives. The FDR is defined as the number of false positives divided by the number of all positive calls, including both true positives and false positives.

3.3 Results

To determine how frequently proximal variants occur within the vicinity of an SVOI, we assessed 430 tumors with varying mutational loads identified from whole genome/exome sequence data of matched normal and tumor tissue (**Figure 3.1, Methods**). Specifically, data from 100 cases each of melanoma, hepatocellular carcinoma and lung squamous cell

carcinoma were obtained from TCGA. We also evaluated data from 48 cases of HER2+ breast cancer**,** 34 cases of small cell lung cancer, 30 cases of hepatocellular carcinoma, 15 cases of oral squamous cell carcinoma, and one hypermutated glioblastoma (one primary and two metastatic samples) from in-house studies. After performing alignment and variant calling, we confirmed the linkage of SVOIs and proximal (somatic or germline) variants by phasing the variants using GATK[124] [**Figure 3.1B**] (**Methods**). Then, the list of SVOIs from each of the samples was intersected with the respective lists of in-phase amino acid-altering proximal variants to assess their presence within the chosen nucleotide window.

3.3.1 Missense variants overlap with missense proximal variants

Out of 430 tumor samples analyzed, 380 samples (88.3%) had at least one (range: 1 to 377) missense SVOI in phase with a proximal missense variant. Of a total of 103,673 missense variants identified in these tumors, there were 7,783 SVOIs (7.5%) with a proximal missense variant (somatic or germline) within 89 nucleotides on either side. 5,344 of these missense SVOIs (5.1%) were also in phase with their respective proximal variants. In most cases (93.8%), SVOIs had a single proximal germline or somatic variant in phase, but occasionally multiple (range: two to six) variants were proximal to the SVOI. An average of 241 missense somatic variants were analyzed per sample. Per patient, an average of 6.5% of SVOIs had a proximal missense variant, and 5% had one or more proximal missense variants in phase with the SVOI. On average, 62.2% of these proximal variants were germline missense variants and 37.7% were somatic missense variants. The majority (68.0%) of proximal somatic variants were contributed by Dinucleotide Polymorphisms (DNPs). Most variant callers (including those used for the harmonized analysis of TCGA data in the Genomic Data Commons) report DNPs as two separate SNVs. Excluding the DNPs, on average, 88.4% of the proximal variants were germline missense SNPs, and 11.6% were somatic missense SNVs.

3.3.2 Predicted binding affinity changes with PVC

To identify neoantigens capable of eliciting an effective anti-tumor T-cell response, it is critical to both determine the correct tumor specific peptide sequence and assess its ability to bind MHC[125]. First, we sought to assess how accounting for proximal variants in the neoantigen peptide sequence may influence binding affinity to MHC. In order to evaluate this, we quantified the impact of missing or incorrectly selecting strong-binding neoantigens when ignoring proximal variants. We compared binding affinity scores before and after PVC for each patient's peptides against their respective MHC Class I alleles.

A typical Class I neoantigen binding evaluation and screening is carried out by sliding over shorter sub-peptide registers[111]. To evaluate strong-binding Class I neoantigens of lengths 8-11-mers, we ideally scan 7-10 amino acids on each side of the mutated amino acid resulting from the SVOI. Even if a proximal variant alters an amino acid in the full peptide window, it may not be included in every register we consider as a candidate neoantigen (**Figure 3.2**).

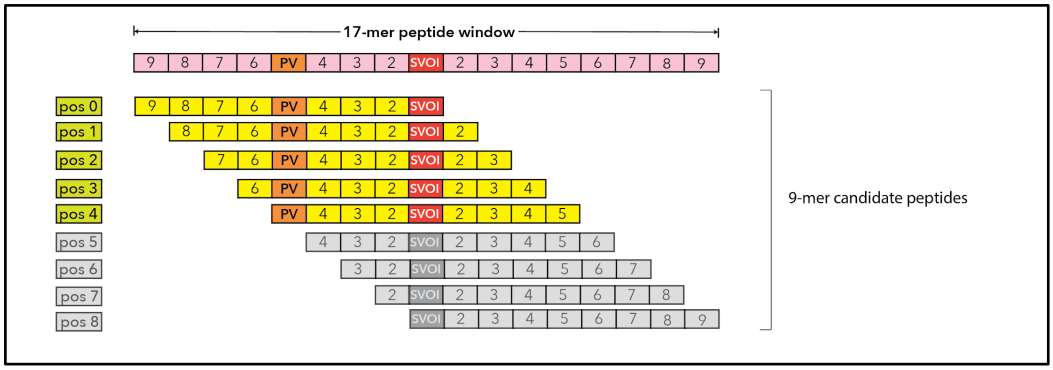


Figure 3.2 Example of candidate neoantigen evaluation *This figure shows the possible sub-peptide registers for selection of a candidate neoantigen of length 9. The 17-mer peptide window for a 9-mer candidate is selected by scanning 8 amino acids on each side of the mutated amino acid resulting from the SVOI (red box). Only those registers that contain amino acid changes resulting from both - the proximal variant (PV; orange box), as well as the SVOI (red box) were considered for this analysis (five peptides shown in yellow for this*

example). The remaining registers shown (grey boxes) contain the SVOI but are not affected by the proximal variant.

In some rare cases, a proximal variant may translate to the same amino acid sequence as the SVOI, or the SVOI and proximal variant both lead to amino acid changes if considered in isolation, but if they are in phase and considered together, they result in no change to the amino acid sequence. To take into account these cases and accurately assess the effect of amino-acid changes due to proximal variants on binding predictions, we only considered those registers that contained both the proximal variant and the SVOI amino acid changes, when translated together. Across 8-11-mers, on average 45.95% of all neoantigen peptide registers contained both. **Figure 3.3** summarizes the effect of proximal variants on neoantigen binding affinity. Although the effect is less pronounced for 8-mers, the smallest length we examined, we see drastic changes in binding affinity due to PVC across all four peptide lengths (represented as log10 of mutant (MT) epitope fold change ($MT_{uncorrected}/MT_{corrected}$), with ranges spanning from -3.0 to 3.1 for 8-11-mers (**Figure 3.3A**). **Figures 3.3C-D** show the distribution of log(MT fold change) scores for 9-mer and 10-mer peptides, respectively. For both peptide lengths, most weak binders stay within the same range before and after PVC but very few strong binders remain unchanged, after PVC. We chose 500 nM as the binding affinity cutoff for a potential binder, as most known T-cell epitopes have an affinity value of less than 500 nM[126]. For the binding prediction changes, we only considered a call as erroneous if PVC yielded at least a 10% change in predicted binding affinity.

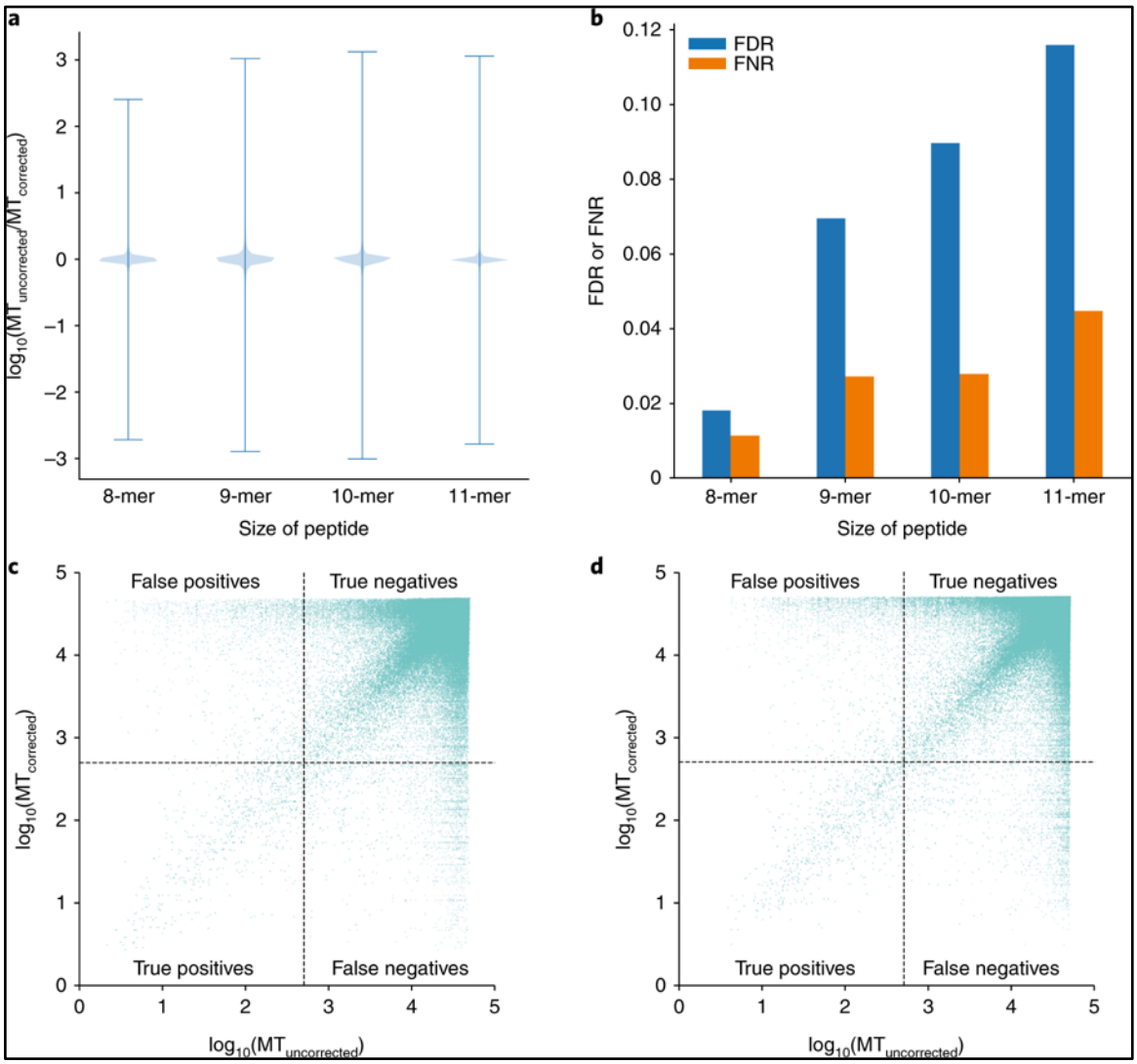


Figure 3.3 Mischaracterization of neoantigens before proximal variant correction
The effect of accounting for proximal variants in neoantigen selection is summarized in several ways (n=380 biologically independent samples with at least one proximal variant). (A) Violin plot (distribution of all data in blue and whiskers indicating max/min values) showing the change in uncorrected neoantigen binding using the existing approach ($MT_{uncorrected}$) versus PVC ($MT_{corrected}$), represented as log10 MT fold change ($MT_{uncorrected}$ / $MT_{corrected}$) across 8-11-mers for all variants in phase with the somatic variant of interest. (B) For 8-11-mer peptides, the False Negative Rate (FNR) (shown as orange bars) represents the number of instances when a truly strong-binding peptide was mistaken as a weak-binding peptide ($MT_{uncorrected}$ > 500 nM, and MT fold-change < 1.1). The False Discovery Rate (FDR) (shown in blue bars) represents the number of instances where a strong-binder before PVC ($MT_{uncorrected}$ < 500nM) is determined to have an incorrect peptide sequence as a result of a proximal variant. (C) Log10 scaled comparison of corrected versus uncorrected binding scores for 9-mer peptides considering patient-specific MHC Class I alleles. Dotted lines demarcate the binding affinity threshold of 500 nM. (D) Log10 scaled comparison of corrected versus uncorrected binding scores for 10-mer peptides considering patient-specific MHC Class I alleles.

3.3.3 Impact of PVC on False Discovery and False Negative Rates

In addition to the effect a proximal amino acid substitution may have on a neoantigen's binding potential, it is also important to consider whether the peptide sequence of the selected neoantigen is correct and representative of the sequence in the tumor. Failure to do so may affect the immunogenic potential of the neoantigen being selected, as the uncorrected neoantigen will not produce tumor-specific T-cells, even if it binds well and is presented by the MHC.

To determine how many neoantigens were being erroneously predicted, and the effect that mischaracterization of neoantigens due to proximal variants would have on candidate selection, we calculated the False Negative Rate (FNR) and False Discovery Rate (FDR) after applying PVC. The FNR and FDR represent probabilities of potential MHC binders (binding affinity < 500 nM) being discarded (false negatives) and of erroneous peptides being mistaken for potential binders (false positives), respectively.

An average of 9 SVOI and 10 neoantigenic peptides were mischaracterized per case. As a consequence, 1,165 potential binders ($MT_{corrected}$ < 500 nM) were erroneously rejected, and 3,305 peptides which were strong binders before PVC were misidentified across all 430 patients investigated here. Overall, FNR and FDR across lengths 8-11 were 0.026 and 0.069, respectively (**Figure 3.3B**).

As a representative example, **Figure 3.4** illustrates data from one of the TCGA melanoma samples with a heterozygous missense SNV in the reverse strand gene *MARCH10* that overlaps an in-phase heterozygous germline single nucleotide polymorphism (SNP), 21

nucleotides upstream. When translated, this germline SNP results in S357F (NP_001275708.1:p.Phe357Ser) alteration that is 7 amino acids downstream to the missense somatic variant F350S (NP_001275708.1:p.Ser350Phe). This variant directly affects the final neoantigen sequence for a peptide of any length (> 8-mer). To evaluate the effect of this germline SNP on the binding affinity of the neoantigen peptide, we calculated the binding affinity of the uncorrected versus the PVC neoantigenic peptides. The binding affinity of the best register for a 10-mer peptide using the uncorrected approach ($MT_{uncorrected}$ = 55.44 nM) is within the range for a good binder (< 500 nM). However, after including this patient's proximal germline variant, the binding affinity for the same register decreases almost 70-fold ($MT_{corrected}$ = 3766.72 nM), thus predicting a very weak binder. Using the uncorrected analysis approach, one might have selected this neoantigenic peptide for a vaccine but after PVC, the candidate peptide is unsuitable. This result illustrates the importance of using the individual variation of the germline genome while selecting and designing neoantigens for personalized immunotherapy.

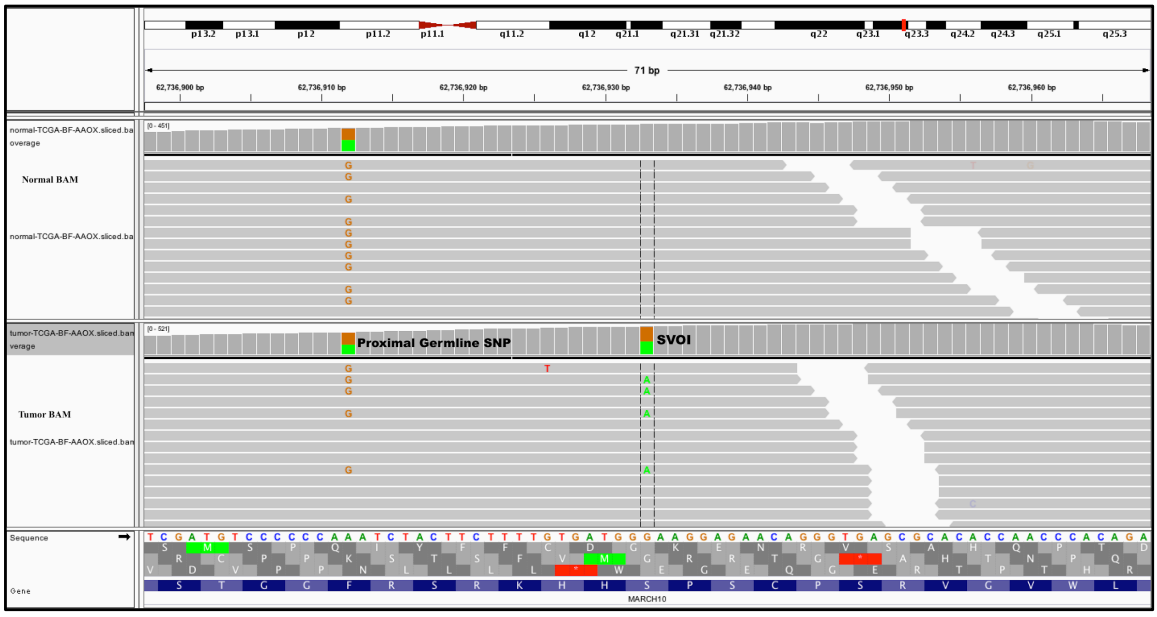


Figure 3.4 Example of a germline SNP within the proximity of a somatic SNV *An example from one of the TCGA melanoma samples with a missense SNV that overlaps a germline SNP (dbSNP ID: rs9891498), 21 nucleotides upstream. When translated, the germline SNP results*

in S357F (NP_001275708.1:p.Phe357Ser) alteration and is 7 amino acids downstream of the missense somatic variant F350S (NP_001275708.1:p.Ser350Phe) in MARCH10.

3.4 Discussion

There are some caveats/limitations of our approach. Firstly, the analysis was restricted only to single nucleotide changes (i.e. missense somatic SNVs that are near another germline or somatic SNV), and did not seek to evaluate whether other, potentially relevant types of variants were found nearby. These include insertions and deletions (both somatic and germline)[127] and different types of structural variants that often have a more significant impact on peptide sequences but also are rarer than SNVs. Phasing of indels and structural variants is also not currently handled by software such as GATK's ReadBackedPhasing. Secondly, our analysis 'window' (89 bp) was defined in genomic coordinates. It is substantially more complicated to consider this window size in the context of transcriptome coordinates, since intronic coordinates must be ignored when scanning upstream and downstream. This is further complicated in genes with alternative transcripts and hence multiple introns and exons to consider. Our ability to determine phase for variants separated by an intron would be limited in WGS or exome data (although could be evaluated in RNA-seq data with sufficient read lengths). Lastly, for this study, we only considered neoantigen binding predictions to MHC Class I molecules. MHC Class II peptides are much longer due to an open binding groove and hence, the subsequent impact of proximal variants on the peptide sequence would be even more pronounced. Due to these limitations, our results are likely an underestimation of the impact of PVC.

Moreover, even with seemingly small false discovery and false negative rates, the importance of accounting for the effect of proximal variants is clear when we consider clinical vaccine

design scenarios. For example, 10 or fewer peptides are usually selected for the final vaccine from a larger number of initial candidates. Given this scenario, we calculated the probability of choosing at least 1 weak binder or of omitting 1 strong binder in the final vaccine, without PVC. For the first probability, we calculated $1 - (1\text{-FDR})^{10} = 0.513$ and for the second, we calculated $1 - (1\text{-FNR})^{10} = 0.228$. The probability that at least one of these errors occurs for each patient evaluated, is $1 - (1\text{-FDR})^{10}*(1\text{-FNR})^{10} = 0.624$. Thus, for neoantigen identification in 100 patients, we can expect that approximately 51 patients would receive a suboptimal vaccine specifically due to receiving a neoantigen with an incorrect peptide sequence, 23 would receive a suboptimal vaccine specifically due to missing a strong-binding neoantigen, and 62 would receive a suboptimal vaccine due to at least one of these causes.

Design of personalized cancer vaccines is complex, time consuming, and expensive. Previous work has shown that only about 16-43% of the predicted neoantigenic peptides included in a vaccine formulation yield CD8+ T-cell response[69–72]. Our study demonstrates the importance of ensuring the selected neoantigens correctly represent the individual's genome and therefore maximize the likelihood of eliciting an immune response. PVC based on the patient's genome can eliminate errors during neoantigen candidate selection, potentially increasing the efficacy of personalized vaccines. Further studies may also demonstrate the importance of considering proximal variants when using neoantigen load to predict response to checkpoint blockade inhibition therapies.

3.5 Code availability

The proximal variant analysis code has now been added to the *proximal_variants* branch of the pVACtools GitHub repository (see URLs). We have also packaged this branch and

uploaded the package as an alpha release to TestPyPi. The alpha release can be installed by running `pip install -f https://test.pypi.org/project/pvactools/1.0.8/ pvactools==1.0.8` on the command line. The feature will be released with the main pVACtools package as part of the next software release cycle (version 1.1.0).

3.5.1 URLs

Picard: https://broadinstitute.github.io/picard/

pVACtools: http://pvactools.org/

Github repository for proximal variant analysis code: https://github.com/griffithlab/pVACtools/tree/proximal_variants

3.6 Data availability

Several of the in-house sequencing datasets used in the study have been previously published and deposited in various databases. All sequence data for the HER2+ breast cancer samples can be accessed via the Database of Genotypes and Phenotypes (dbGAP; study accession: phs001291)[128]. Data for oral squamous cell carcinoma project and hepatocellular carcinoma samples are part of other manuscripts currently in preparation, and can be accessed under dbGAP study accession phs001623 and phs001106, respectively. Results for the glioblastoma case[129] and small cell lung cancer[119] have been published and can be accessed under dbGAP study accessions phs001663 and phs001049, respectively. TCGA data can be accessed under dbGaP study accession phs000178.

3.7 Authors and Contributions

Jasreet Hundal[1], Susanna Kiwala[1], Yang-Yang Feng[1], Connor J. Liu [1], Ramaswamy Govindan[2,3], William C. Chapman[4], Ravindra Uppaluri[5], S Joshua Swamidass[6], Obi L. Griffith[1,2,3,7], Elaine R. Mardis[8]*, Malachi Griffith[1,2,3,7]*

Affiliations

(1) McDonnell Genome Institute, Washington University School of Medicine, St. Louis, Missouri, USA

(2) Siteman Cancer Center, Washington University School of Medicine, St. Louis, Missouri, USA

(3) Division of Oncology, Department of Internal Medicine, Washington University School of Medicine, St. Louis, Missouri, USA

(4) Department of Surgery, Washington University School of Medicine, St. Louis, Missouri, USA

(5) Department of Surgery/Otolaryngology, Brigham and Women's Hospital and Dana-Farber Cancer Institute, Boston, Massachusetts, USA

(6) Department of Pathology and Immunology, Washington University School of Medicine, St. Louis, Missouri, USA

(7) Department of Genetics, Washington University School of Medicine, St. Louis, Missouri, USA

(8) Institute for Genomic Medicine, Nationwide Children's Hospital, Columbus, Ohio, USA

* Corresponding authors. elaine.mardis@nationwidechildrens.org; mgriffit@wustl.edu

JH was involved in all aspects of this study including designing and developing the methodology, analyzing and interpreting data, and writing the manuscript, with input from CJL, SJS, OLG, ERM, and MG. SK was involved in development of neoantigen prediction software, and participated in the data analysis and writing the manuscript. YYF contributed to data analysis, interpretation, and writing the manuscript. RG, WCC and RU provided unpublished tumor datasets and provided critical feedback on the manuscript. ERM and MG supervised the study. All authors read and approved the final manuscript.

3.8 Acknowledgements

We are grateful to the patients and their families without whom this study would not be possible. We would like to thank G Dunn for early access to raw data for the published glioblastoma hypermutator case included in our analysis. We would also like to thank R Schreiber and B Carreno for the initial discussions that inspired the study, and for their expertise and guidance during the study. S Swamidass was supported by the NIH National Library of Medicine (R01LM012222 and R01LM012482). O Griffith was supported by the NIH NCI (K22CA188163 and U01CA209936). M Griffith was supported by the National Institutes of Health (NIH), National Human Genome Research Institute (NHGRI; award number R00HG007940) and the NIH National Cancer Institute (NCI) (U01CA209936).

Chapter 4: pVACtools- a computational toolkit to select and visualize cancer neoantigens

Hundal J et al. pVACtools: a computational toolkit to select and visualize cancer neoantigens (2018) Manuscript submitted. doi:10.1101/501817

4.1 Introduction

Increasing interest in identifying the numbers and types of predicted neoantigens encoded by a cancer genome has placed an emphasis on the facility and precision of related computational prediction tools[45]. Several such efforts have been published[112,113,130]. Typically, these tools start with a list of somatic variants (in VCF or other formats), with annotated protein changes, and predict the strongest MHC binding peptides (8-11-mer for class I MHC and 13-25-mer for class II) using one or more prediction algorithms[56,57,131]. The predicted neoantigens are then filtered or ranked based on quality metrics including sequencing read coverage, variant allele fraction, gene expression, and differential binding compared to the wild type peptide (agretopicity index score [85]). However, of the small number of such prediction tools **(Table 4.1)**, most lack some key functionality, including predicting neoantigens from gene fusions, aiding optimized vaccine design for DNA cassette vaccines, and including nearby germline or somatic alterations into the candidate neoantigens[132]. An intuitive graphical user interface to visualize and efficiently select the most promising candidates is also critical to facilitate involvement of clinicians and other researchers in the process of neoantigen evaluation.

To address these limitations and to add facility for all end-users, we created a comprehensive and extensible framework for computational identification, selection, prioritization and visualization of neoantigens - *'pVACtools',* that facilitates each of the major components of neoantigen identification. This computational framework can be used to identify neoantigens from a variety of somatic alterations, including gene fusions and insertion/deletion frameshift mutations, both of which potentially create very immunogenic neoantigens[133]. Further, pVACtools can facilitate both MHC class I and II predictions, and provides an interactive display of predicted neoantigens for review by the end user.

	pVACtools	**Vaxrank**[113]	**MuPeXI**[112]	**CloudNeo**[130]	**FRED2**[134]	**Epi-Seq**[85]	**ProTect** (https://github.com/BD2KGenomics/protect)
Variant calling (within the pipeline)	N	N	N	N	N	Y	Y
Variant types supported	SNVs, Inframe Indels, Frameshifts, Gene fusions (pVACfuse)	SNVs, Indels,	SNVs, Inframe Indels, Frameshifts	SNVs	SNVs, Inframe Indels, Frameshifts,	SNVs	SNVs
HLA typing (within the pipeline)	N	N	N	Y (PolySolver, HLAMiner)	Y (via separate installation of OptiType, Polysolver, Seq2HLA, & ATHLATES)	N	Y (Phlat)
RNA-Seq expression data filter	Y (pVACseq)	Y	Y	N	N	Y	Y
Sequence coverage & VAF filter	Y (pVACseq)	Y	Y (only compatible with MuTect2 variant calls)	N	N	Y	N
Algorithms used for epitope binding affinity prediction	IEDB (web and local) (MHC Class I: NetMHCpan, NetMHC, NetMHCcons, PickPocket,, SMM, SMMPMBEC) MHCflurry MHCnuggets (MHC Class II: NetMHCIIpan, SMMalign, NNalign MHCnuggets)	IEDB (web and local) MHCflurry, NetMHC, NetMHCpan, NetMHCIIpan, NetMHCcons,	netMHCpan (local only)	netMHC, netMHCpan (local only)	Local: NetMHC, NetMHCPan, NetMHCII, NetMHCIIpan, NetCTLpan, PickPocket Included: Select from, SMMPMBEC, syfpeithi, SMM, Tepitopepan, ARB,epidemix, comblibsidney, Unitope, HAMMER, SVMHC, BIMAS,	NetMHC (local only)	IEDB (local only)
Stability prediction	Y (NetMHCstabpan)	N	N	N	Y (NetMHCstabpan)	N	N
Cleavage site prediction	Y (NetChop)	Y (NetChop)	N	N	Y (ProteaSMM, PCM, Ginodi, NetChop)	N	N
Support for vector design/epitope assembly	Y (pVACvector)	N	N	N	Y (OptiVac)	N	N
Incorporation	Y	Y (RNA only)	N	N	N	Y	N

of proximal variants						(RefHap)	
Unique epitope ranking method	Y	Y ("Total Binding Score")	Y ("Priority Score")	N	Y (OptiTope immunogenicity score)	N	Y (Rankboost score)
Graphical User Interface	Y (pVACviz)	N	Y (http://www.cbs.dtu.dk/services/MuPeXI/)	Y (via NCI Cancer Genomics Cloud version)	Y (EpiToolKit 2.0)	N	N
Results visualization	Y (pVACviz)	N	N	N	N	N	N
HTTP REST API	Y (pVACapi)	N	N	N	Y	N	N
License	NPOSL-3.0	Apache 2.0	Unknown	Apache 2.0	3-clause BSD	Unknown	Apache 2.0

Table 4.1: Comparison of existing software and tools for cancer immunotherapy analysis

4.2 pVACtools workflow

The pVACtools workflow (***Figure 4.1***) is divided into flexible components that can be run independently. The main tools in the workflow are: (a) pVACseq: a significantly enhanced and reengineered version of our previous pipeline[111] for identifying and prioritizing neoantigens from a variety of tumor-specific alterations (b) pVACfuse: a tool for detecting neoantigens resulting from gene fusions (c) pVACviz: a graphical user interface web client for process management, visualization and selection of results from pVACseq (d) pVACvector: a tool for optimizing design of neoantigens and nucleotide spacers in a DNA vector that prevents high-affinity junctional epitopes, and (e) pVACapi: an OpenAPI HTTP REST interface to the pVACtools suite.

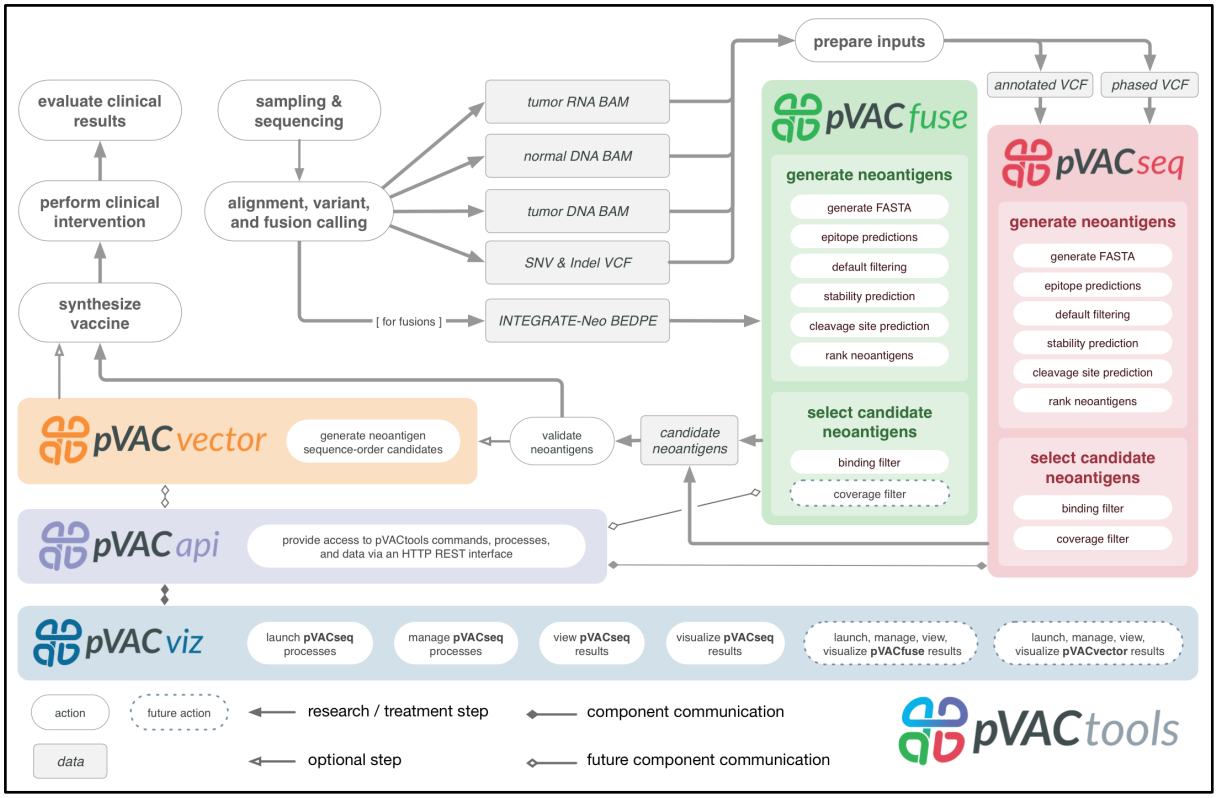


Figure 4.1: Overview of pVACtools workflow: *The pVACtools workflow is highly modularized and is divided into flexible components that can be run independently. The main tools under the workflow include pVACseq*[111] *for identifying and prioritizing neoantigens from a variety of somatic alterations (red inset box), pVACfuse (green) for detecting neoantigens resulting from gene fusions, pVACviz (blue) for process management, visualization and selection of results and pVACvector (orange) for optimizing design of neoantigens and nucleotide spacers in a DNA vector. All of these tools interact via the pVACapi (purple), an OpenAPI HTTP REST interface to the pVACtools suite.*

pVACseq[111] has been completely implemented in Python3 and extended to include many new features since our initial report of its use. pVACseq no longer requires a custom input format for variants, and now uses a standard VCF file annotated with VEP[4]. In our own neoantigen identification pipeline, this VCF is the result of merging results from multiple somatic variant callers and RNA expression tools. Information that is not natively available in the VCF output from somatic variant callers (such as coverage and variant allele fractions for RNA and DNA, as well as gene and transcript expression values) now can be added to the VCF using vcf-annotation-tools (vatools.org), a suite of accessory scripts that we created to accompany pVACtools. pVACtools queries these features directly from the VCF, enabling

prioritization and filtering of neoantigen candidates based on sequence coverage and expression information. In addition, pVACtools now makes use of phasing information provided in the VCF, taking into account all variants proximal to somatic variants of interest that alter neoantigen peptide sequences. Since proximal variants can change the neoantigenic peptide sequence and also affect neoantigen binding predictions, this is an important confounding factor to ensure that the selected neoantigens correctly represent the individual's genome[132]. We have also expanded the supported mutation types for neoantigen predictions to include in-frame indels and frameshift mutations. These capabilities expand the potential number of targetable neoantigens several-fold in many tumors.

To prioritize neoantigens, pVACseq now offers support for as many as eight different MHC Class I epitope prediction algorithms and four MHC Class II prediction algorithms. The tool does this by leveraging the Immune Epitope Database (IEDB)[135] and their suite of six different MHC class I prediction algorithms, as well as three MHC Class II algorithms. pVACseq supports local installation of these tools for power-users, or provides straightforward access by default via the IEDB RESTful web interface. In addition, pVACseq now contains an extensible framework for supporting new neoantigen prediction algorithms that has been used to add support for two new non-IEDB algorithms - MHCflurry[61] and MHCnuggets[62]. By creating a framework that integrates many tools we allow for (a) a broader ensemble approach than IEDB, and (b) a system that other users can leverage to develop improved ensemble ranking, or to integrate proprietary or not-yet-public prediction software. Importantly, this framework enables non-informatics-savvy users to predict neoantigens from sequence variant data sets.

Once neoantigens have been predicted, the pVACseq ranking score is used to prioritize them. This score takes into account gene expression, sequence read coverage, binding affinity predictions, and agretopicity. In addition to applying strict binding affinity cutoffs, the pipeline also offers support for MHC allele-specific cutoffs[136]. Taking a step further than most commonly used approaches, we also offer cleavage position predictions via optional processing through NetChop[137] as well as stability predictions made by NetMHCstab[138].

Previous studies have shown that the novel protein sequences produced by gene fusions frequently produce neoantigen candidates[139]. pVACfuse provides support for predicting neoantigens from such gene fusions. Fusion variants may be imported in annotated BEDPE format from any fusion caller (we used INTEGRATE-Neo[139]). These variants are then assessed for presence of fusion neo-epitopes using predictions against any of the pVACseq-supported binding prediction algorithms.

Implementing cancer vaccines in a clinical setting requires multidisciplinary teams, many of whom may not be informatics savvy. To support this growing community of users, we developed pVACviz, which is a browser-based user interface that assists in launching, managing, reviewing, and visualizing the results of pVACtools processes. Instead of interacting with the tools via terminal/shell commands, the pVACviz client provides a modern web-based user experience. Users complete a pVACseq process setup form that provides helpful documentation and suggests valid values for inputs. The client also provides views showing ongoing processes, their logs, and interim data files to aid in managing and troubleshooting. After a process has completed, users may examine the results as a filtered data table, or as a scatterplot visualization - allowing them to curate results and save them as a CSV file for further analysis.

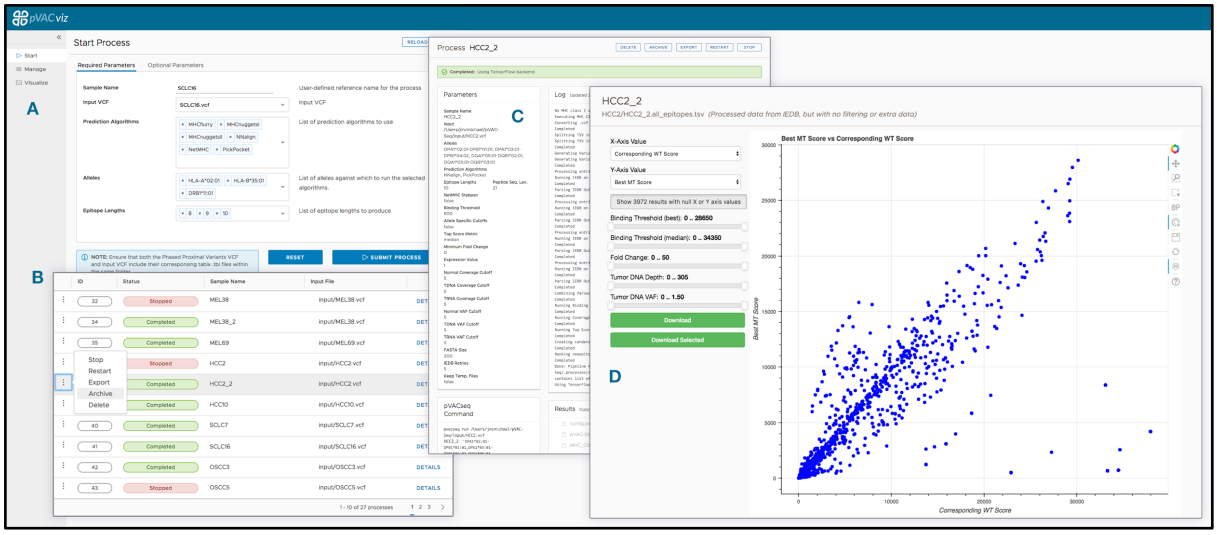


Figure 4.2: pVACviz GUI client: *pVACtools provides a browser-based graphic user interface, called pVACviz, that provides an intuitive means to launch pipeline processes, monitor their execution, and analyze, export, or archive their results. To launch a process, users navigate to the Start Page (A), and complete a form containing all of the relevant inputs and settings for a pVACseq process. Each form field includes help text, and provides typeahead completion where applicable. For instance, the Alleles field provides a typeahead dropdown menu that match available alleles. Once a process is launched, a user may monitor its progress on the Manage Page (B), which lists all running, stopped, and completed processes. The Details Page (C) shows a process' current log, attributes, and any results files as well as providing buttons for stopping, restarting, exporting and archiving the process. The results of pipeline processes may be analyzed on the Visualize Page (D), which displays a customizable scatterplot of a file's rows. The X and Y axis may be set to any column in the result set, and filters may be applied to values in any column. Additionally, points may be selected on the scatter plot or data grid (not visible in this figure) for further analysis or export as CSV files.*

Furthermore, to support informatics groups that want to incorporate or build upon the pVACtools features, we developed pVACapi, which provides a HTTP REST interface to the pVACtools suite. Currently, it provides the API that pVACviz uses to interact with the pVACtools suite. Advanced users could develop their own user interfaces, or use the API to control multiple pVACtools installations remotely over an HTTP network.

Once a list of neoantigen candidates has been prioritized and selected, the pVACvector utility can be used to aid in the construction of DNA-based cancer vaccines. The input is either the output file from pVACseq or a fasta file containing peptide sequences, and pVACvector

returns a neoantigen sequence ordering that minimizes the effects of junctional epitopes (which may create novel antigens) between the sequences. This is accomplished by using the core pVACseq services to predict the binding scores for each junctional peptide and by testing junctions with spacer[140] amino acid sequences that may help to reduce reactivity. The final vaccine ordering is achieved through a simulated annealing procedure that returns a near-optimal solution, when one exists.

pVACtools has been used to predict and prioritize neoepitopes for several neoantigen studies[141–143] and cancer vaccine clinical trials (e.g. NCT02348320 and NCT03122106). We also have a large external user community (the original 'pvacseq' package has been downloaded over 37,000 times from PyPi, and the 'pvactools' package has been downloaded over 9,000 times) that has been actively evaluating and using these packages for their neoantigen analysis, and has also helped in the subsequent refinement of pVACtools through feedback.

4.3 Methods

To demonstrate the utility and performance of the pVACtools package, we downloaded exome sequencing and RNA-Seq data from The Cancer Genome Atlas (TCGA)[14] from 100 cases each of melanoma, hepatocellular carcinoma and lung squamous cell carcinoma, and used patient-specific MHC Class I alleles (**Figure 4.3**) to determine neoantigen candidates for each cancer.

4.3.1 TCGA data pre-processing

Aligned tumor and normal BAMs from BWA[87] (version 0.7.12-r1039) as well as somatic variant calls from VarScan2[144,97] (in VCF format) were downloaded from the Genomic Data

Commons (GDC, https://gdc.cancer.gov/). Since the GDC does not provide germline variant calls for TCGA data, we used GATK's[145] HaplotypeCaller to perform germline variant calling using default parameters. These calls were refined using VariantRecalibrator in accordance with GATK Best Practices[121]. Somatic and germline missense variant calls from each sample were then combined using GATK's CombineVariants, and the variants were subsequently phased using GATK's ReadBackedPhasing algorithm.

Phased Somatic VCF files were annotated with RNA depth and expression information using VCF annotation tools (vatools.org). We restricted our analysis to only consider 'PASS' variants in these VCFs as these are higher confidence than the raw set, and the variants were annotated using the "--pick" option in VEP.

Existing *in silico* HLA typing information was obtained from The Cancer Immunome Atlas (TCIA) database[122].

4.3.2 Neoantigen prediction

The VEP-annotated VCF files were then run through pVACseq using all eight Class I prediction algorithms and for epitope lengths 8-11. The current MHC Class I algorithms supported by pVACseq are NetMHCpan[58], NetMHC[57,58], NetMHCcons(Karosiene et al. 2012), PickPocket[59], SMM[63], SMMPMBEC[64], MHCflurry[61] and MHCnuggets[62]. The four MHC Class II algorithms that are supported are NetMHCIIpan, SMMalign, NNalign, and MHCnuggets. For the demonstration analysis, we limited our prediction to only MHC Class I alleles due to availability of HLA typing information from TCIA, though predictions of Class II can be just as easily generated using pVACtools.

4.3.3 Ranking of Neoantigens

To help prioritize neoantigens, a ranking score is assigned where each of the following four criteria are assigned a rank-ordered value (where the worst = 1):

B = binding affinity

F = Fold Change between MT and WT alleles

M = mutant allele expression, calculated as (Gene expression * Mutant allele RNA Variant allele fraction)

D = DNA Variant allele fraction

A final ranking is based on a score obtained from combining these values:

Priority Score = B+F+(M*2)+(D/2). This score is not meant to be the final word on peptide suitability for vaccines, but was designed to be a useful metric.

4.3.4 Pipeline for creation of pVACtools input files

pVACtools is designed to support a standard VCF variant file format and thus, should be compatible with many existing variant calling pipelines. However, as a reference, we provide the following description of our current somatic and expression analysis pipeline (manuscript in preparation) which has been implemented using docker, CWL[146], and Cromwell[147]. The pipeline consists of workflows for alignment of exome/DNA- and RNA-Seq data, somatic and germline variant detection, RNA-Seq expression estimation as well as optional HLA typing.

This pipeline starts with raw patient tumor exome or cDNA capture[148] and RNA-seq data and produces annotated VCFs for neoantigen identification and prioritization with pVACtools.

Our pipeline consists of three main components: DNA alignment, variant detection and annotation, as well as RNA-seq data processing. More specifically, we use BWA-MEM[87] for aligning the patient's tumor and normal exome data. The output BAM then undergoes merging (Samtools Merge), query name sort (Picard SortSam), duplicate marking (Picard MarkDuplicates), position sorting followed by base quality recalibration (GATK BaseRecalibrator). GATK's HaplotypeCaller[145] is used for germline variant calling and the output variants are annotated using VEP[4] and filtered for coding sequence variants.

For somatic variant calling, our pipeline combines the output of four variant detection algorithms- Mutect2[149], Strelka[98], Varscan[144,97] and Pindel[150]. The combined variants are normalized using GATK's LeftAlignAndTrimVariants where the indels are left-aligned and common bases are trimmed. Vt[151] is used to split multi-allelic variants. Several filters such as gnomAD allele frequency, percentage of mapq0 reads, as well as pass-only variants are applied prior to annotation of the VCF using VEP. We use a combination of custom and standard plugins for VEP annotation (params: --format VCF --plugin Downstream --plugin Wildtype --symbol --term SO --transcript_version --tsl --coding_only --flag_pick --hgvs). Variant coverage is assessed using bam-readcount (https://github.com/genome/bam-readcount) for both the tumor and normal DNA exome data and is also annotated into the VCF output using VCF-annotation-tools (vatools.org).

Our pipeline also generates a phased-VCF file by combining both the somatic and germline variants and running the sorted combined variants through GATK ReadBackedPhasing.

For RNA-seq data, the pipeline first trims the adapter sequence using flexbar[152] and aligns the patient's tumor RNA-seq data using HISAT2[153]. Two different methods, Stringtie[154] and

Kallisto[155], are employed for evaluating both the transcript and gene expression values. Additionally, coverage support for variants in RNA-seq data can also be assessed through bam-readcount. This information is added to the VCF using VCF-annotation-tools and serves as an input for neoantigen prioritization using pVACtools.

Optionally, our pipeline can also run HLA-typing *in silico* using OptiType[48] when clinical HLA typing is not available.

4.3.5 Implementation of software

pVACtools is written in Python3. The individual tools are implemented as separate command line entry points that can be run using the `pvacseq`, `pvacfuse`, `pvacvector`, `pvacapi`, and `pvacviz` commands to run the respective tool. pVACapi is required to run pVACviz so both the `pvacapi` and `pvacviz` command need to be executed in separate terminals. For pVACseq, the PyVCF package is used for parsing the input VCF files. The mhcflurry and mhcnuggets packages are used to run the MHCflurry and MHCnuggets prediction algorithms, respectively. The pandas package is used for data management while filtering and ranking the neoantigen candidates in pVACseq and pVACfuse. The simanneal package is used for the simulated annealing procedure when running pVACvector. pVACapi is implemented using Flask and Bokeh. The pVACviz client is written in TypeScript using the Angular web application framework, the Clarity UI component library, and the ngrx library for managing application state. The test suite is implemented using the Python unittest framework and GitHub integration tests are run using travis-ci (travis-ci.org). Code changes are integrated using GitHub pull requests (https://github.com/griffithlab/pVACtools/pulls). Feature additions, user requests, and bug reports are managed using the GitHub issue tracking (https://github.com/griffithlab/pVACtools/issues). User documentation is written using the

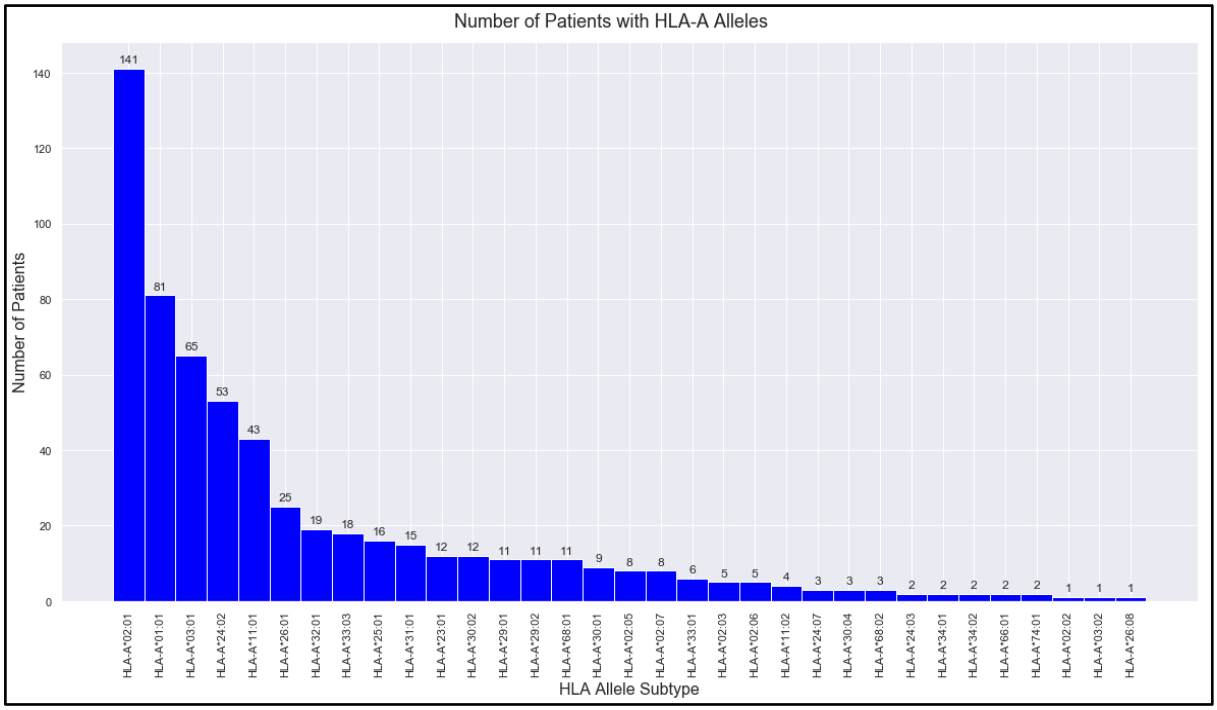
reStructuredText markup language and the Sphinx documentation framework (sphinx-doc.org). Documentation is hosted on Read The Docs (readthedocs.org).

4.4 Results and Discussion

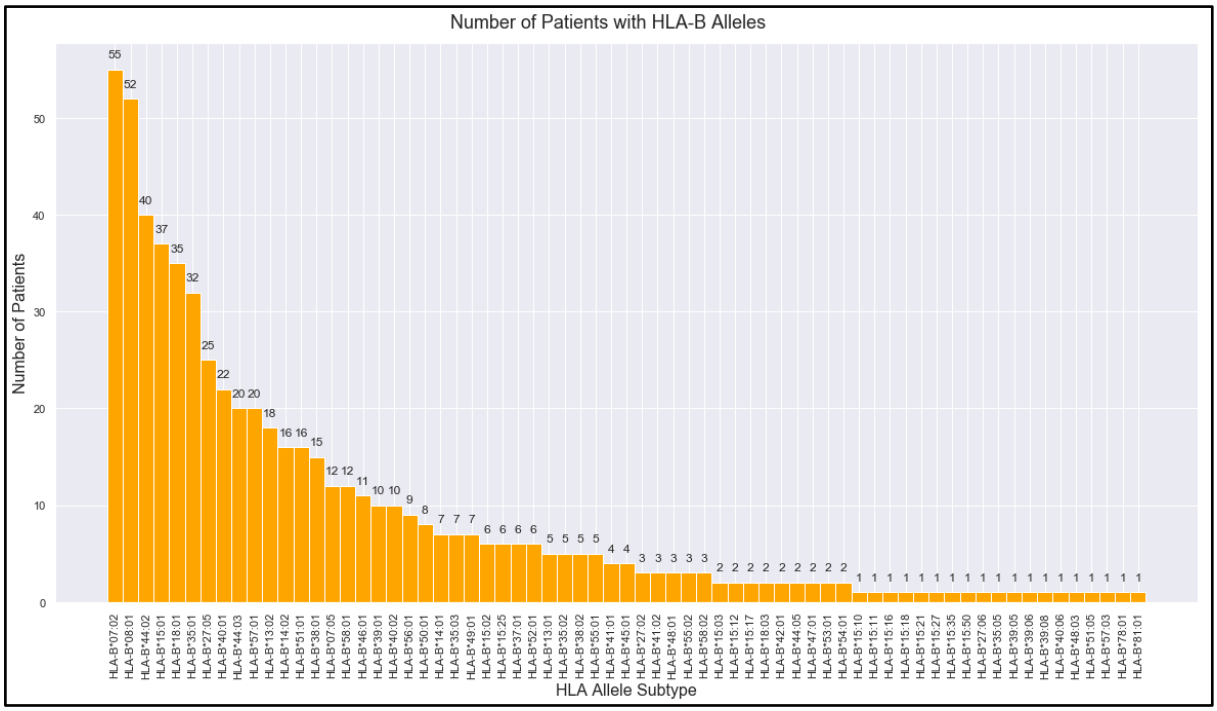
4.4.1 Analysis of TCGA data using pVACtools

To demonstrate the utility and performance of the pVACtools package, we downloaded exome sequencing and RNA-Seq data from The Cancer Genome Atlas (TCGA)[14] from 100 cases each of melanoma, hepatocellular carcinoma and lung squamous cell carcinoma, and used patient-specific MHC Class I alleles **(Figure 4.3**) to determine neoantigen candidates for each patient. There were a total of 64,422 VEP-annotated variants reported across 300 samples, with an average of 214 variants per sample. Of these, 61,486 were single nucleotides variants (SNVs), 479 were inframe insertions and deletions and 2,465 were frameshift mutations (**Figure 4.4**). We used this annotated list of variants as input to the pVACseq component of pVACtools to predict neoantigenic peptides. pVACseq reported 14,599,993 unfiltered peptide candidates. The original version of pVACseq[111] reported 10,284,467 peptides, and thus, by extending support for additional variant types as well as prediction algorithms (due to support for additional alleles), we produced 42% more raw candidate neoantigens.

(a)


Number of Patients with HLA-A Alleles
Number of Patients
HLA Allele Subtype
HLA-A*02:01
HLA-A*01:01
HLA-A*03:01
HLA-A*24:02
HLA-A*11:01
HLA-A*26:01
HLA-A*32:01
HLA-A*33:03
HLA-A*25:01
HLA-A*31:01
HLA-A*23:01
HLA-A*30:02
HLA-A*29:01
HLA-A*29:02
HLA-A*68:01
HLA-A*30:01
HLA-A*02:05
HLA-A*02:07
HLA-A*33:01
HLA-A*02:03
HLA-A*02:06
HLA-A*11:02
HLA-A*24:07
HLA-A*30:04
HLA-A*68:02
HLA-A*24:03
HLA-A*34:01
HLA-A*34:02
HLA-A*66:01
HLA-A*74:01
HLA-A*02:02
HLA-A*03:02
HLA-A*26:08

(b)


Number of Patients with HLA-B Alleles
Number of Patients
HLA Allele Subtype
HLA-B*07:02
HLA-B*08:01
HLA-B*44:02
HLA-B*15:01
HLA-B*18:01
HLA-B*35:01
HLA-B*27:05
HLA-B*40:01
HLA-B*44:03
HLA-B*57:01
HLA-B*13:02
HLA-B*14:02
HLA-B*51:01
HLA-B*38:01
HLA-B*07:05
HLA-B*58:01
HLA-B*46:01
HLA-B*39:01
HLA-B*40:02
HLA-B*56:01
HLA-B*50:01
HLA-B*14:01
HLA-B*35:03
HLA-B*49:01
HLA-B*15:02
HLA-B*15:25
HLA-B*37:01
HLA-B*52:01
HLA-B*13:01
HLA-B*35:02
HLA-B*38:02
HLA-B*55:01
HLA-B*41:01
HLA-B*45:01
HLA-B*27:02
HLA-B*41:02
HLA-B*48:01
HLA-B*55:02
HLA-B*58:02
HLA-B*15:03
HLA-B*15:12
HLA-B*15:17
HLA-B*18:03
HLA-B*42:01
HLA-B*44:05
HLA-B*47:01
HLA-B*53:01
HLA-B*54:01
HLA-B*15:10
HLA-B*15:11
HLA-B*15:16
HLA-B*15:18
HLA-B*15:21
HLA-B*15:27
HLA-B*15:35
HLA-B*15:50
HLA-B*27:06
HLA-B*35:05
HLA-B*39:05
HLA-B*39:06
HLA-B*39:08
HLA-B*40:06
HLA-B*48:03
HLA-B*51:05
HLA-B*57:03
HLA-B*78:01
HLA-B*81:01

(c)

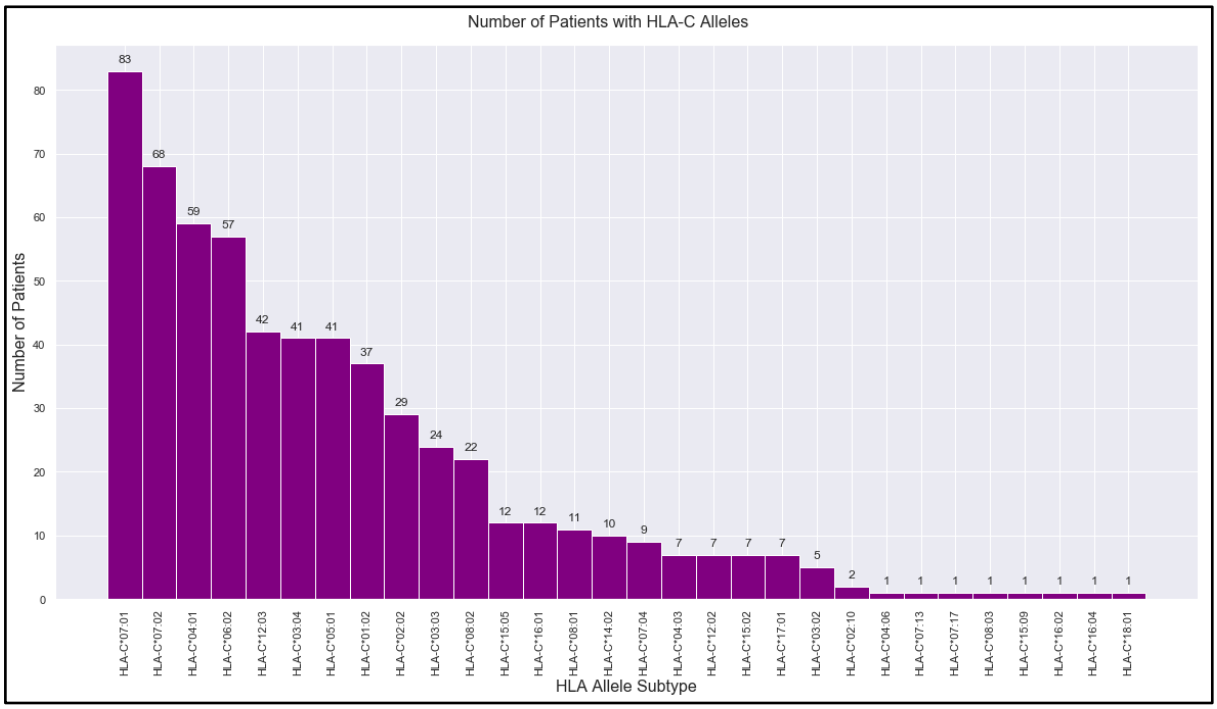


Figure 4.3: Patient counts per HLA allele subtype *Distribution of HLA-alleles for the entire cohort of 300 TCGA patients analyzed in this study is shown in each of the three plots for (a) HLA-A allele subtypes (33 unique alleles); (b) HLA-B allele subtypes (67 unique alleles); (c) HLA-C allele subtypes (30 unique alleles). The total number of unique HLA alleles found in the patient cohort was 130.*

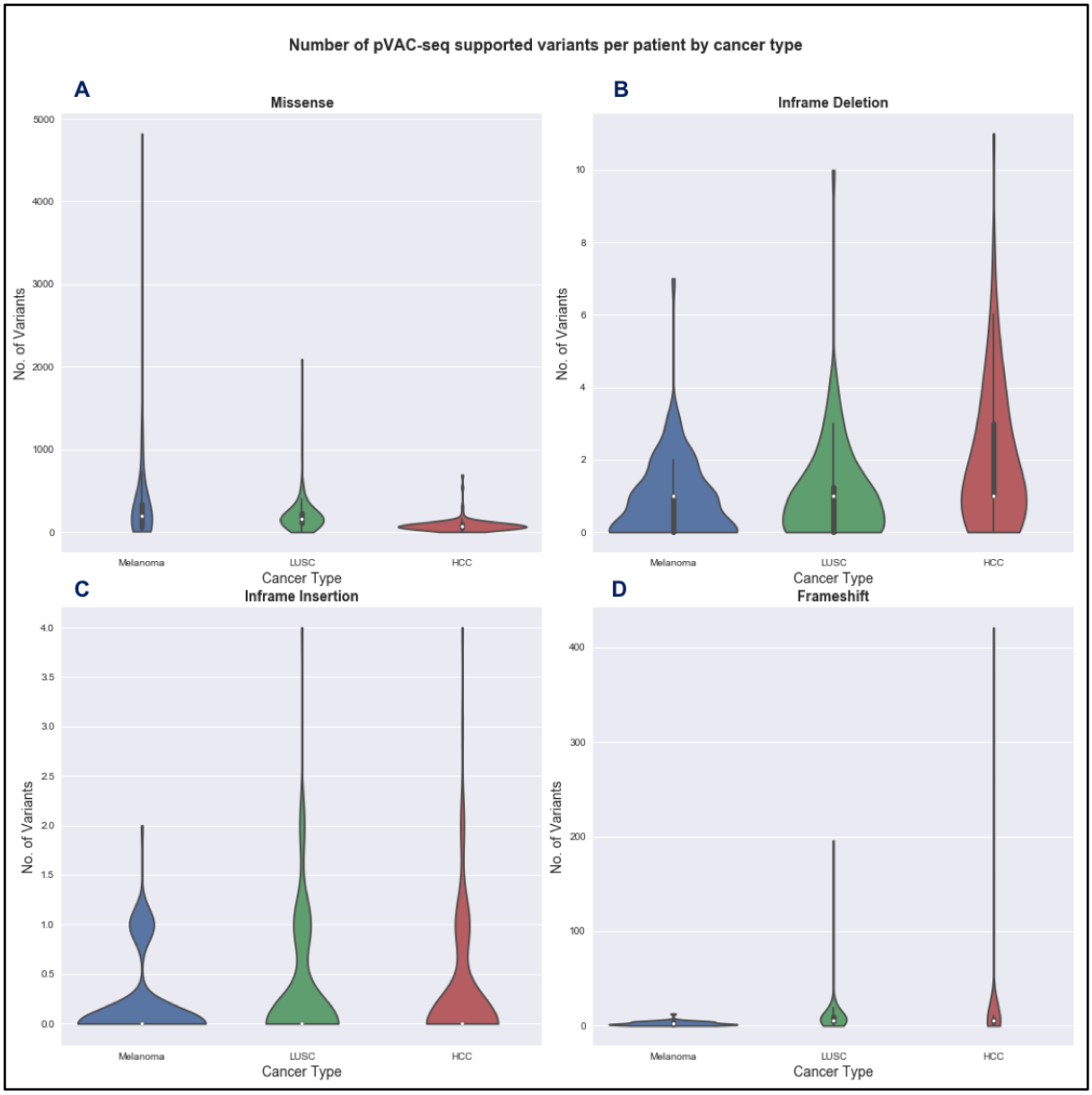


Figure 4.4: Violin plots showing the distribution of observed variants per cancer type summarized for each variant type supported by pVACseq*. (A) missense (total variant count: 61,486), (B) inframe deletion (total variant count: 389), (C) inframe insertion (total variant count: 81), and (D) frameshift (total variant count: 2,465).*

By applying our default median binding affinity cutoff of 500 nM across all eight MHC Class I prediction algorithms, there were 96,235 predicted strong binding neoantigens, derived from 34,552 somatic variants (32,788 missense SNVs, 1,603 frameshift variants and 131 in-frame indels). This set of strong binders was further reduced by filtering out mutant peptides with

median predicted binding affinities (across all prediction algorithms) greater than that of the corresponding wildtype peptide (i.e. mutant/wildtype binding affinity fold change > 1), resulting in 70,628 neoantigens from 28,588 variants (26,880 SNVs, 1,583 frameshift and 125 in-frame indels).

This set was subsequently filtered by evaluation of exome sequencing data coverage and our recommended defaults, as follows. By applying the default criteria of variant allele fraction (VAF) cutoff of > 25% in tumor and < 2% in normal sample, with coverage levels of at least 10X tumor coverage and at least 5X normal coverage, 10,730 neoantigens from 4,891 associated variants (4,826 SNVs, 56 frameshift and 9 in-frame indels) were obtained, with an average of 36 neoantigens predicted per case. Since RNA-seq data also were available, the filtering criteria included RNA-based coverage filters (tumor RNA VAF > 25% and tumor RNA coverage > 10X) as well as a gene expression filter (FPKM > 1). To condense the results even further, only the top ranked neoantigen was selected per variant across all alleles, lengths, and registers (position of amino acid mutation within peptide sequence), resulting in 4,891 total neoantigens with an average of 16 neoantigens per case. This list was then processed with pVACvector to determine the optimum arrangement of the predicted high quality neoantigens for a DNA-vector based vaccine design (**Figure 4.5**).

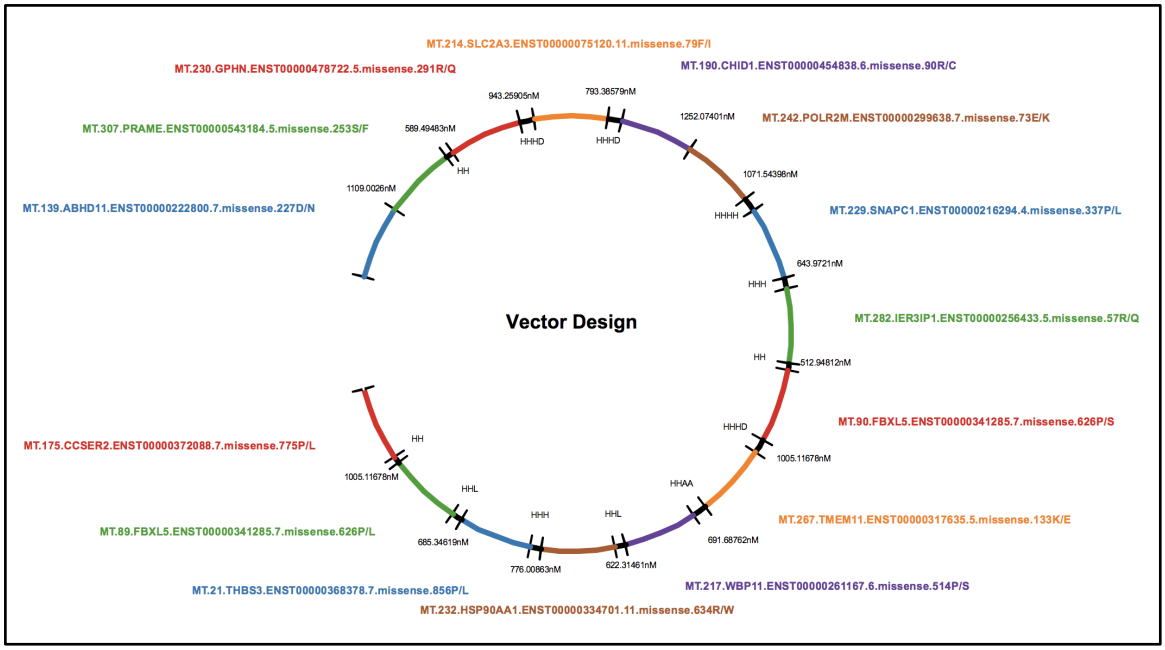


Figure 4.5: An example from pVACvector output showing the the optimum arrangement of candidate neoantigens for a DNA-vector based vaccine design. *The figure shows visualization of a circularized DNA insert carrying the encoded neoantigenic peptide sequences to be synthesized and encoded/cloned into a DNA plasmid. DNA sequences encoding each peptide are ordered (with use of spacer sequences where needed) to ensure there are no strong-binding junctional epitopes. Each neoantigenic peptide candidate is shown in Blue,Green, Red, Orange, Purple, and Brown. Spacer sequences, where added for minimizing junctional epitope affinity, are depicted in Black, along with the binding affinity value of the strongest binding junctional epitope. Labels depict MutantIdentifier:GeneName:TranscriptName:TranscriptNumber:TypeofMutation:AminoAcid Change.*

4.4.2 Comparison of epitope prediction software

Since we offer support for as many as eight different epitope prediction tools, we assessed agreement between these algorithms from a random subset of 100,000 peptides (**Figure 4.6**). The highest correlation was observed between the two stabilization matrix method (SMM)-based algorithms - SMM and SMMPMBEC. The next best correlation was observed between NetMHC and MHCflurry.

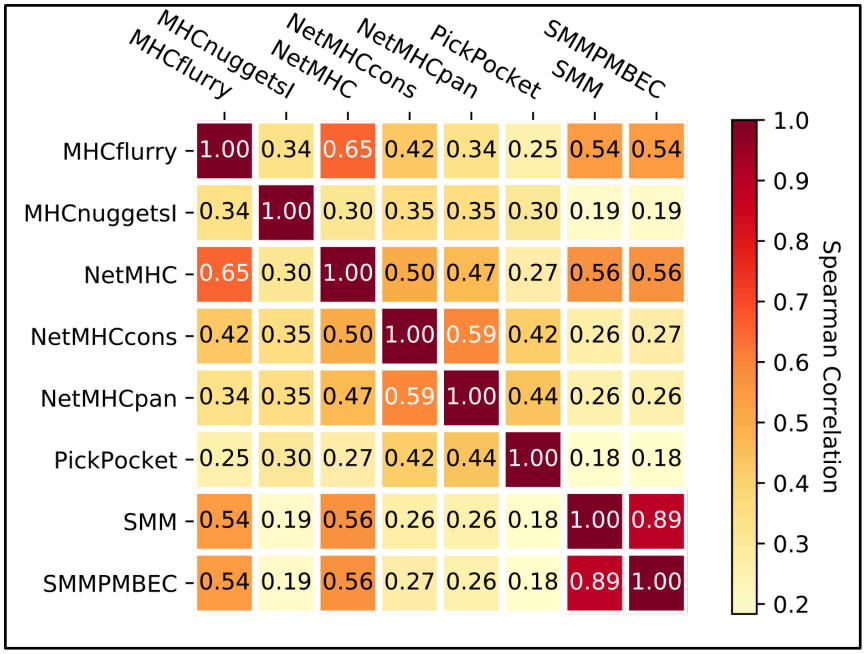
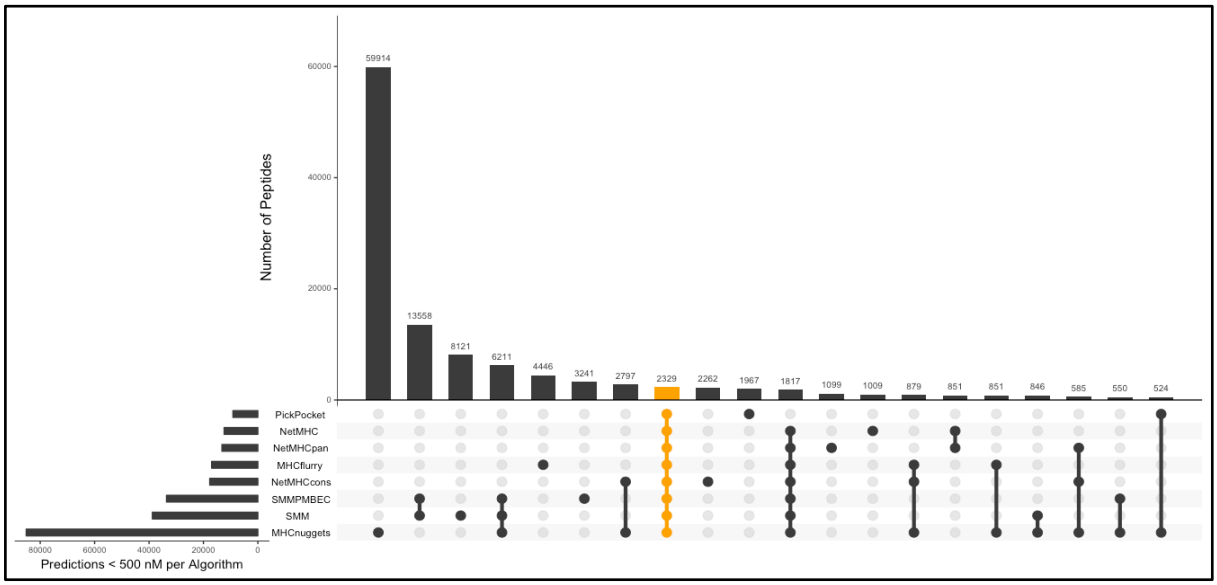


Figure 4.6: Spearman Correlation between prediction values *This figure shows a heatmap of the Spearman correlation between binding affinity predictions from all eight algorithms generated from random subsample set of size 100,000 peptides*

keWe also evaluated if there were any biases within the algorithms to predict strong (i.e. binding affinity < = 500nM) or weak binding epitopes (**Figures 4.7 and 4.8**). We found that MHCnuggets predicts the highest number of strong-binding candidates alone. Of the total number of strong binding candidates predicted, 64.7% of these candidates were predicted by a single algorithm, 35.2% were predicted as strong-binders by two to seven algorithms, and only 1.8% of the strong-binding candidates were predicted as strong binders by the combination of all eight algorithms. Infact, even if one (or more) algorithms predict a peptide to be a strong binder, often another algorithm not only doesn't agree but disagrees by a large margin, in some cases predicting that same peptide as a very weak binder. This remarkable lack of agreement underscores the potential value of considering multiple algorithms.

(a)



(b)

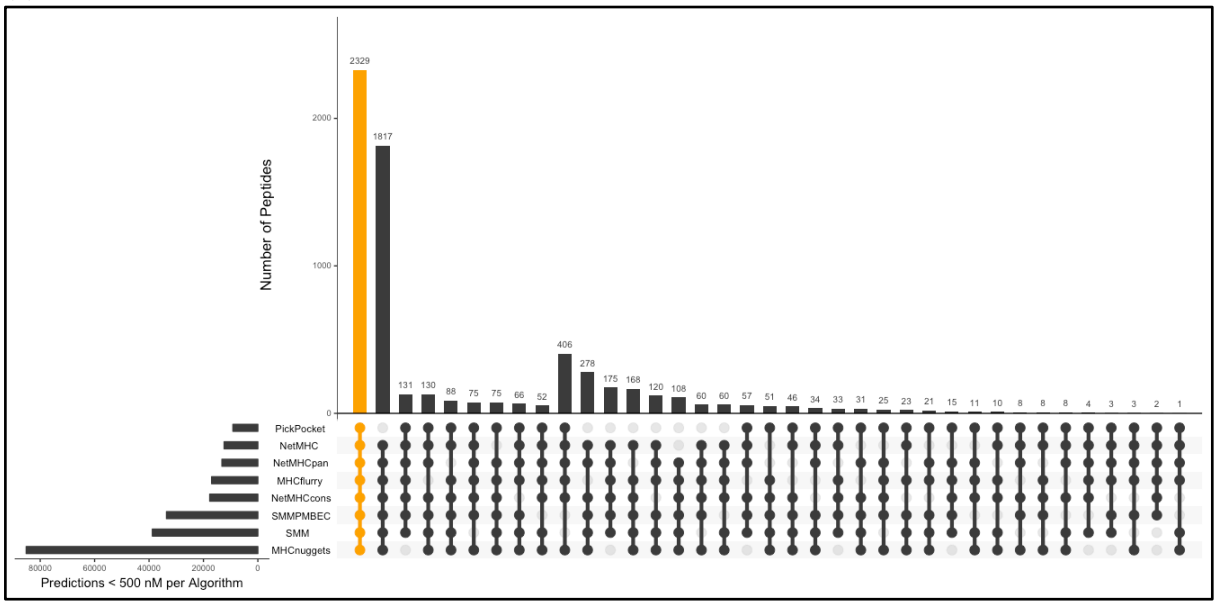


Figure 4.7: Upset plot between number of peptides and prediction algorithms *Intersection of peptide sequences predicted by different algorithms are shown using upset plots. The y-axis shows the number of overlapping unique neoantigenic peptides predicted for each combination of algorithm depicted on the x-axis. Each filled black circle shows the sets contained in an exclusive intersection (i.e. the identity of each algorithm), while the light gray circles represent the algorithm(s) that do not participate in this exclusive intersection. (a) Upset plot for the top 20 algorithm combinations ranked by the number of peptides predicted to be a good binder (mutant IC50 score < 500 nM). The combination of all eight algorithms (highlighted orange) ranks the 8th highest; (b) Upset plot for algorithm combinations where at least six algorithms agree on predicting a peptide to be a good binder (MT IC50 score < 500 nM). The combination of all eight algorithms (highlighted orange) ranks the highest.*

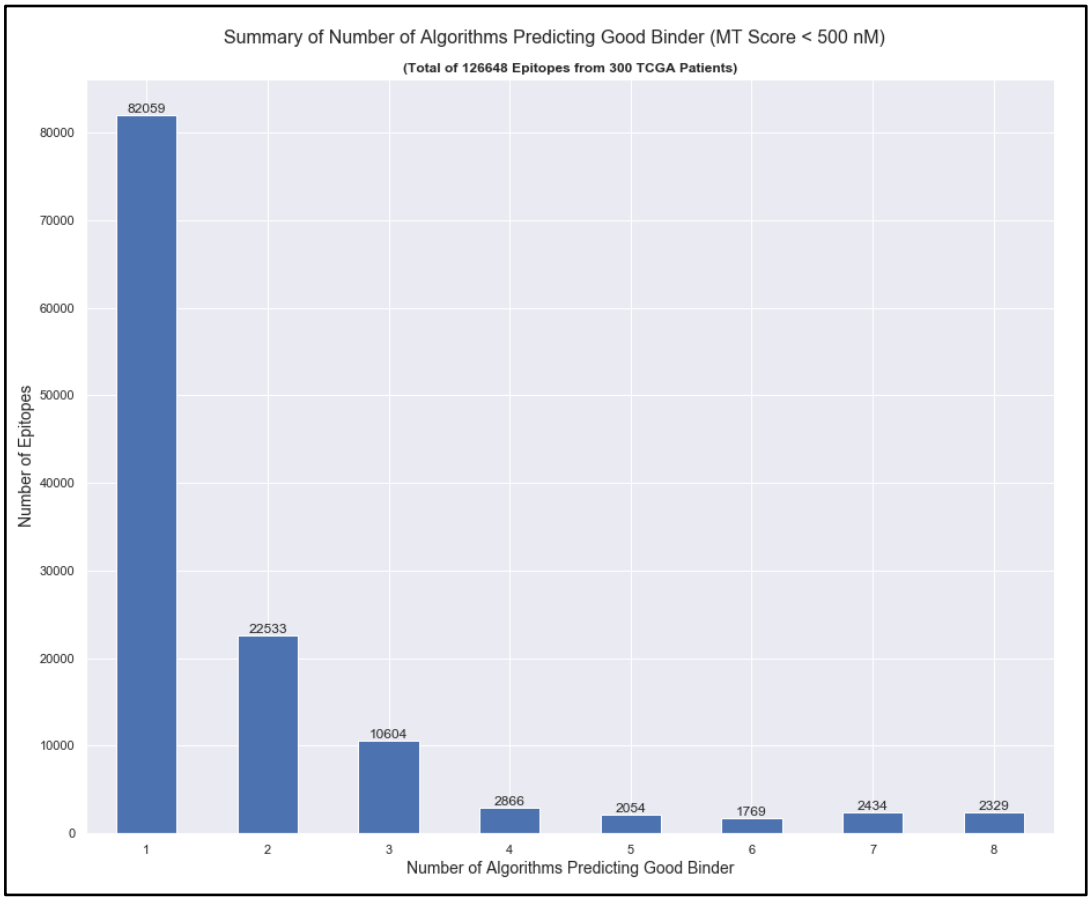
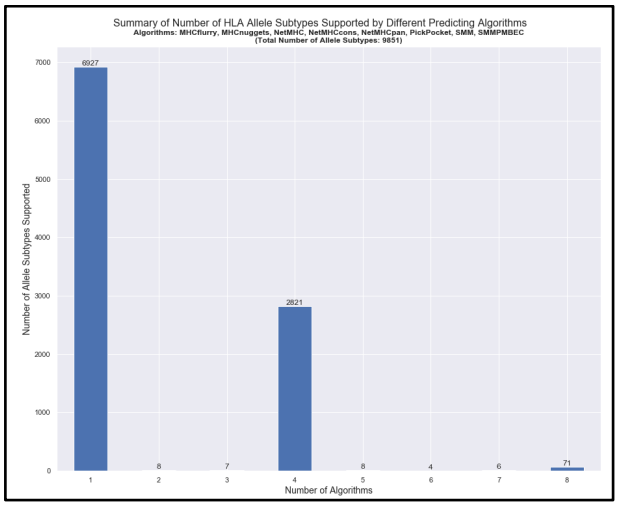


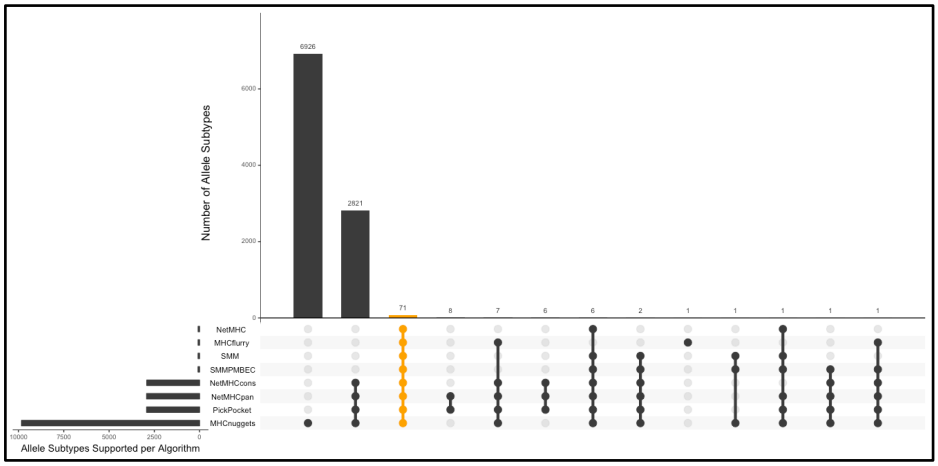
Figure 4.8: Number of peptides predicted to be good binders (MT IC50 Score < 500 nM) versus number of algorithms used.

Next we determined if the number of human HLA alleles supported by these eight algorithms differed. As shown (**Figure 4.9),** MHCnuggets supports the highest number of human HLA alleles.

(a)



(b)



(c)

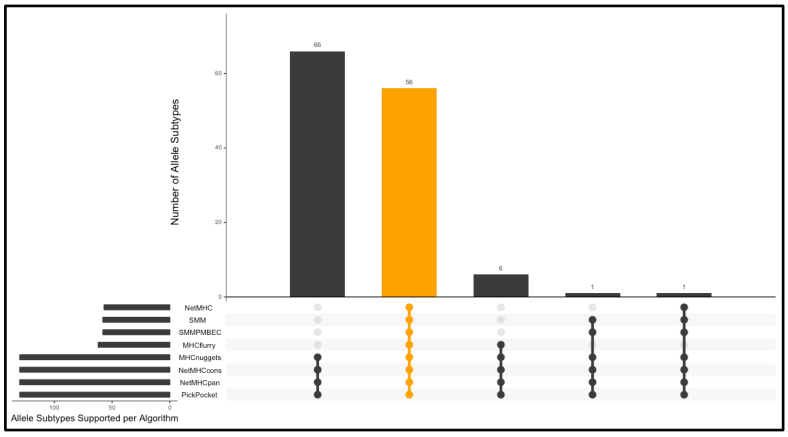


Figure 4.9: Human HLA Allele subtype support distribution by eight algorithms *(a) Number of allele subtypes supported versus number of algorithms; (b) Upset plot for algorithm combinations ranked by the number of allele subtypes supported by pVACseq. Total number of HLA allele subtypes supported by pVACseq: 9,851. The combination of all eight algorithms (highlighted orange) ranks the 3th highest; (C) Upset plot for algorithm combinations ranked by the number of allele subtypes supported for the 300 TCGA samples*

analyzed in this study. Total number of HLA alleles across patient cohort: 130. The combination of all eight algorithms (highlighted orange) ranks the 2nd highest.

4.4.3 Comparison of filtering criteria

Since pVACtools offers a multitude of ways to filter the list of predicted neoantigens, we evaluated and compared the effect of each filter in narrowing down high quality neoantigen candidates.

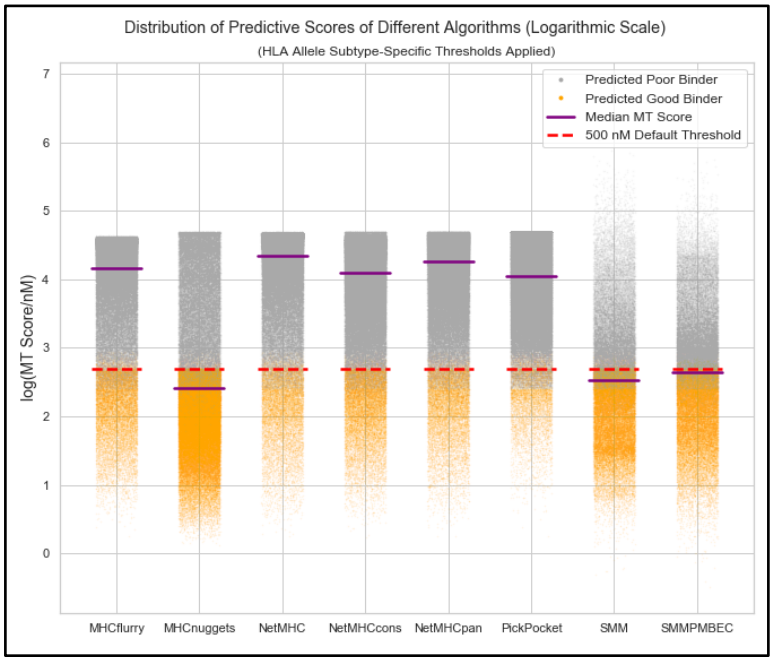
We first compared the effect of running pVACtools using the commonly used standard binding score cutoff of <= 500nM (parameters: -b 500) versus the newly added allele-specific score filter (parameters: -a). These cutoffs were, by default, applied to the "median" binding score of all prediction algorithms. 96,235 neoantigens (average 320.78 per patient) were predicted using the 500 nm binding score cutoff compared to 94,068 neoantigens (average 313.56 per patient) using allele-specific filters. About 79% neoantigens were shared between the two sets.

We also narrowed further to include only those predictions where the default "median" predicted binding affinities (< = 500 nm) are lower than each corresponding median wildtype peptide affinity (parameters: -b 500 -c 1) (i.e. a binding affinity mutant/wild type ratio or aggretopicity value indicating that the mutant version of the peptide is a stronger binder). Using the aggretopicity value filter, 70,628 neoantigens (average 235.43 per patient) were predicted versus the previously reported 96,235 neoantigens without this filter.

We also evaluated the effect of the aggretopicity value filter when applied to the set of epitopes filtered on "lowest" binding score of < = 500 nm (parameters: -b 500 -c 1 -m lowest). This filter reports peptides where at least one of the algorithms predicts a strong

binder, instead of calculating a median score and requiring that to meet the 500 nm threshold (**Figure 4.10).** Using the lowest binding score filter resulted in an 11-fold increase in the number of candidates predicted (827,423 candidates, average 2,758.08 per patient).

(a)



(b)

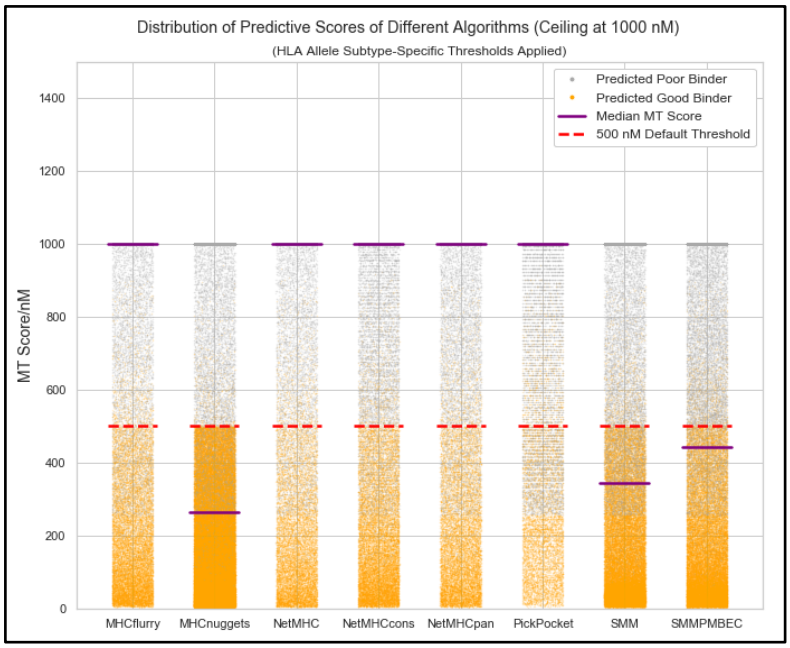


Figure 4.10: Overall of distribution of binding affinity scores (nM) for peptides where at least one of the algorithms predicts a strong binder. *HLA allele subtype-specific thresholds*

are applied when available, otherwise the default cutoff binding affinity of 500 nM is used. Peptides with predicted MT IC50 scores lower than their respective cutoff scores are highlighted in orange. The median MT IC50 scores of each algorithm's prediction are marked for reference. (a) Original data on logarithmic scale; (b) A ceiling of 1000 nM is applied to the MT IC50 scores.

Using the median (default) binding affinity filtering criteria, we next applied coverage and expression based filters. First, we filtered using the recommended defaults i.e. greater than 5X normal DNA coverage, less than 2% normal VAF, greater than 10X tumor RNA and DNA coverage and greater than 25% tumor RNA and DNA VAF, along with FPKM > 1 for transcript level expression (parameters: --normal-cov 5 --tdna-cov 10 --trna-cov 10 --normal-vaf 0.02 --tdna-vaf 0.25 --trna-vaf 0.25 --expn-val 1). A total of 10,730 neoantigens were shortlisted across all samples with an average of 35 neoantigens per case. We then compared this set with a slightly more stringent criteria using tumor DNA and RNA VAF of 40% (parameters: --tdna-vaf 0.40 --trna-vaf 0.40). This shortened our list of predicted neoantigens to 4,073 candidates averaging to 13 candidates per patient.

Lastly, we applied our top binding score filter to select the best candidate neoantigen per variant across all alleles and all lengths (one result per variant) using the previously described default filters. This resulted in a final list of 4,891 neoantigens across 300 patients.

4.4.4 Demonstration of neoantigen analysis using pVACfuse

To demonstrate the potential of neoantigens resulting from gene fusions, we analyzed TCGA prostate cancer RNA-seq data from 302 patients. This dataset was used as a demonstration set for the neoantigen prediction supported by Integrate-Neo. We wanted to assess the difference (if any) in neoantigens candidates reported by INTEGRATE-Neo[139] using the one MHC

Class I prediction algorithm it supports (NetMHC) versus an ensemble of eight Class I prediction algorithms supported in pVACfuse.

Using 1,619 gene fusions across 302 samples as input, pVACfuse reported 2,104 strong binding neoantigens (binding affinity <= 500nM) resulting from 739 gene fusions. On average, there were about 7 neoantigens per sample resulting from an average of 2 fusions per case. This is an eight fold increase in the number of strong binding neoantigens predicted by pVACfuse versus the ones reported by INTEGRATE-Neo, which reported 261 neoantigens across 210 fusions.

4.5 Conclusion

As reported from our demonstration analysis, a typical tumor has too many possible neoantigen candidates to be practical for a vaccine. There is therefore a critical need for a tool that takes in the input from a standard sequencing analysis pipeline and reports a filtered and prioritized list of neoantigens. pVACtools enables a streamlined, accurate and user-friendly analysis of neoantigenic peptides from NGS cancer datasets. This suite offers a complete and easily configurable end-to-end analysis, starting from somatic variants and gene fusions (pVACseq and pVACfuse respectively), through filtering, prioritization, and visualization of candidates (pVACviz), and determining the best arrangement of candidates for a DNA vector vaccine (pVACvector). Furthermore, by supporting additional classes of variants as well as gene fusions, we offer an increase in the number of predicted epitopes which is even more important in the case of low mutational burden tumors. Finally, by extending support for multiple binding prediction algorithms, we allow for a consensus approach. The need for this

integrated approach is made abundantly clear by the high disagreement between these algorithms observed in our demonstration analyses.

The results from pVACtools analyses are already being used in cancer immunology studies, including studying the relationship between tumor mutation burden and neoantigen load to predict response in checkpoint blockade therapy trials and the design of cancer vaccines in ongoing clinical trials. We anticipate that pVACtools will make such analyses more robust, reproducible, and facile as these efforts continue.

4.6 Data availability

Data from 100 cases each of melanoma, hepatocellular carcinoma and lung squamous cell carcinoma were obtained from TCGA and downloaded via the Genomics Data Commons (GDC). This data can be accessed under dbGaP study accession phs000178. Data for demonstration and analysis of fusion neoantigens was downloaded from the Github repo for Integrate (https://github.com/ChrisMaherLab/INTEGRATE-Vis/tree/master/example).

4.7 Software availability

The pVACtools codebase is hosted publicly on GitHub at https://github.com/griffithlab/pVACtools and https://github.com/griffithlab/BGA-interface-projects (pVACviz). User documentation is available at pvactools.org. This project is licensed under the Non-Profit Open Software License version 3.0 (NPOSL-3.0, https://opensource.org/licenses/NPOSL-3.0). pVACtools has been packaged and uploaded to PyPi under the “pvactools” package name and can be installed on Linux systems by running the `pip install pvactools[API]` command. Installation requires a Python 3.5 environment

which can be emulated by using Conda. Versioned Docker images are available on DockerHub (https://hub.docker.com/r/griffithlab/pvactools/).

4.8 Authors and Contributions

Jasreet Hundal[1+], Susanna Kiwala[1+], Joshua McMichael[1], Christopher A. Miller[1,2,4], Alexander T. Wollam[1], Huiming Xia[1], Connor J. Liu[1], Sidi Zhao[1] , Yang-Yang Feng[1], Aaron P. Graubert[1], Amber Z. Wollam[1], Jonas Neichin[1], Megan Neveau[1], Jason Walker[1], William E Gillanders[4,5], Elaine R. Mardis[3], Obi L. Griffith[1,2,4,6,*], Malachi Griffith[1,2,4,6,*]

Affiliations

(1) McDonnell Genome Institute, Washington University School of Medicine, St. Louis, MO, USA

(2) Division of Oncology, Department of Medicine, Washington University School of Medicine, St. Louis, MO, USA

(3) Institute for Genomic Medicine, Nationwide Children's Hospital, 575 Children's Crossroad, Columbus OH, USA

(4) Siteman Cancer Center, Washington University School of Medicine, St. Louis, MO, USA

(5) Department of Surgery, Washington University School of Medicine, St. Louis, MO, USA

(6) Department of Genetics, Washington University School of Medicine, St. Louis, MO, USA

+ These authors contributed equally to this work

* Corresponding authors. obigriffith@wustl.edu, mgriffit@wustl.edu

JH and SK were involved in all aspects of this study including designing and developing the methodology, and writing the manuscript, with help from CAM, ERM, OLG and MG.

JH analyzed and interpreted data with input from HX, CJL, SZ and Y-YF. SK wrote the software code with help from ATW, APG, AZW, MN and JW. JM developed pVACviz and pVACapi with assistance from APG. JN, MN, and CAM developed pVACvector. WEG provided input about clinical needs for vaccine design to help improve the tool features. OG and MG supervised the project and revised the paper. All authors read and approved the final manuscript.

4.9 Acknowledgements

We thank the patients and their families for donation of their samples and participation in clinical trials. We also thank our growing user community for testing the software and providing useful input, critical bug reports as well as suggestions for improvement and new features. We are grateful to Drs. Robert Schreiber, Gavin Dunn and Beatriz Carreno for their expertise and guidance on foundational work on cancer immunology using neoantigens and suggestions on improving the pipeline.

Chapter 5: Conclusion and future directions

Neoantigen prediction accuracy is a critical factor in the widespread adoption of cancer vaccine-based immunotherapy, as vaccine efficacy depends on selection of the best neoantigens. The approaches developed in this work should help to identify, evaluate and characterize tumor-specific neoantigens in a much-reduced time, thereby increasing the applicability of cancer vaccines for clinical use.

Preliminary studies in human patients have shown that only a subset of candidate antigens predicted by existing methods elicit an immune response when administered as a vaccine[72]. In fact, only about 16–43% of the predicted neoantigenic peptides included in a vaccine formulation for any reported clinical trial to-date yield a CD8+ T cell response. This lack of accuracy in prediction may emerge from several sources of uncertainty in neoantigen prediction, including the inability to identify which of the predicted neoantigenic peptides with high binding affinity will be processed and presented to the immune system by the MHC. This aspect is not readily predicted computationally, although there are several algorithms available to perform these predictions [137]. Hence, there is a critical need for an approach that learns from the response data obtained from clinical cancer vaccine trials and uses the information to refine neoantigen predictions for future trials. We have hypothesized that a machine learning approach might help in refining these predictions such that a higher percentage of the neoantigens are able to induce T-cell immunity when administered in the cancer vaccine. Hence, we have initiated an effort to curate these data from multiple clinical trials and to implement a machine learning-based approach.

Pursuit of machine learning will require several components, outlined below:

Construct target variable(s)

To assess 'good' vaccine candidates, we will first need to define appropriate criteria to construct a target variable(s) as an indicator of immunogenicity. Different immunology labs may use different metrics for such an analysis, and since not all labs actively use neoantigens for vaccination per se, some data mining will be required to determine which assays are used routinely. This will help in narrowing down a common endpoint across different projects to assess immunogenicity. Some examples of these assays include ELISPOT/ELISA that measures IFNγ production, Peptide–MHC tetramers/dextramer assays, or a combination of both.

Obtain data from published and ongoing research

There are several in-house vaccine clinical trials currently underway as collaborations between the McDonnell Genome Institute and the Siteman Cancer Center. For example, the Komen Promise Vaccine trial which aims to design and treat triple negative breast cancer patients during the window of opportunity following neoadjuvant chemotherapy, surgery and radiation therapy, is a phase I clinical trial that compares synthetic polypeptide- (https://clinicaltrials.gov/ct2/show/NCT02427581) to DNA-based vaccines (https://clinicaltrials.gov/ct2/show/NCT02348320). Initial results were already published for a first-in-human vaccine trial in melanoma patients (https://clinicaltrials.gov/ct2/show/NCT00683670). There were five patients enrolled in this phase 1 vaccine clinical trial employing autologous, functionally mature, interleukin (IL)-12p70-producing dendritic cells (DC) carrying neoantigens identified by our pipeline.

We also are actively curating previously published research to inform our machine learning approach[37,41,42,66,156,157]. One caveat for these published datasets is the absence of transcriptome data in some cases, which will need to be imputed while building the predictive model.

Engineer features based on dataset properties

Since genomic and/or transcriptomic data are now used routinely for assessment of neoantigens, it is essential to use this information from such massively parallel sequencing studies to select appropriate features for building the machine learning classifier. We may use basic aggregators on the given data to engineer meaningful features for the model. Some of these features include:

a) Tumor coverage from DNA and/or RNA

b) Tumor variant allele fraction from DNA and/or RNA

c) Gene expression: this may be in the form of FPKM values obtained from RNA-(Cap) Seq data or quantitative expression from qPCR.

d) Binding affinity from epitope prediction algorithms

e) Fold change in the predicted binding of the mutant epitope versus the wildtype

Additional features such as HLA-type, type of amino acids in the peptide, properties of amino acids (hydrophobic, hydrophilic, polar), position of the mutation in the peptide sequence (anchor residues, TCR-facing residues) etc. could also be investigated.

Select feature subset

Since there are several different features associated with sequencing datasets, it may be possible that not all features are relevant when building a predictor. Popular feature selection

methods such as chi squared and analysis of variance (ANOVA) could be used to retrieve a subset of features that have a strong correlation with the target variable. This subset would be used by the predictive model to carry out the classification.

Perform predictive modeling

Once a set of features is finalized, a feature matrix is constructed based on the features selected for the study as well the target variable. A predictive model would then be developed which would use the feature matrix as input to predict the best set of neoantigens to be used for analysis and/or vaccination. There are several different machine learning approaches that can be evaluated including neural networks, random forest, support vector machine and deep learning approaches. Evaluation is generally done by cross validation techniques such as N--fold cross validation, leave one out cross validation, etc. Various metrics like area under the ROC curve (AUC), accuracy of the model, precision, recall, F1 score, etc. can be used to determine the appropriate classifier.

As we learn from ongoing early human trials, the methods developed in this thesis will help in identification, selection and characterization of high quality neoantigens from different cancer sequencing datasets. Furthermore, these methods could be directly applied to assist in development of immunotherapy based clinical trials and basic immunology studies such as studying the relationship between tumor mutation burden and neoantigen load to predict response in checkpoint blockade therapy trials and the design of personalized cancer vaccines. The computational framework developed in this work (pVACtools) will help optimize the composition of personalized cancer vaccines with high precision that will hasten vaccine design to enable growing clinical demand. Additionally, as mentioned previously,

these methods could even be extended by incorporating the feedback from the growing number of trials to improve the prediction of neoantigens and their immunogenicity.

References

1. International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature* **431,** 931–945 (2004).
2. Mardis, E. R. A decade's perspective on DNA sequencing technology. *Nature* **470,** 198–203 (2011).
3. Mardis, E. R. Next-generation sequencing platforms. *Annu. Rev. Anal. Chem.* **6,** 287–303 (2013).
4. McLaren, W. *et al.* The Ensembl Variant Effect Predictor. (2016). doi:10.1101/042374
5. McLaren, W. *et al.* Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics* **26,** 2069–2070 (2010).
6. Kumar, P., Henikoff, S. & Ng, P. C. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat. Protoc.* **4,** 1073–1081 (2009).
7. Ng, P. C. & Henikoff, S. Predicting the Effects of Amino Acid Substitutions on Protein Function. *Annu. Rev. Genomics Hum. Genet.* **7,** 61–80 (2006).
8. Carter, H. *et al.* Cancer-specific high-throughput annotation of somatic mutations: computational prediction of driver missense mutations. *Cancer Res.* **69,** 6660–6667 (2009).
9. Adzhubei, I. A. *et al.* A method and server for predicting damaging missense mutations. *Nat. Methods* **7,** 248–249 (2010).
10. Reva, B., Antipin, Y. & Sander, C. Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic Acids Res.* **39,** e118 (2011).
11. Kumar, R. D., Swamidass, S. J. & Bose, R. Unsupervised detection of cancer driver

mutations with parsimony-guided learning. *Nat. Genet.* **48,** 1288–1294 (2016).

12. Ley, T. J. *et al.* DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. *Nature* **456,** 66–72 (2008).
13. Ding, L., Wendl, M. C., Koboldt, D. C. & Mardis, E. R. Analysis of next-generation genomic data in cancer: accomplishments and challenges. *Hum. Mol. Genet.* **19,** R188–96 (2010).
14. Cancer Genome Atlas Research Network *et al.* The Cancer Genome Atlas Pan-Cancer analysis project. *Nat. Genet.* **45,** 1113–1120 (2013).
15. Zhang, J. *et al.* International Cancer Genome Consortium Data Portal--a one-stop shop for cancer genomics data. *Database* **2011,** bar026–bar026 (2011).
16. Garraway, L. A. & Lander, E. S. Lessons from the Cancer Genome. *Cell* **153,** 17–37 (2013).
17. Lawrence, M. S. *et al.* Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499,** 214–218 (2013).
18. Mardis, E. R. The translation of cancer genomics: time for a revolution in clinical cancer care. *Genome Med.* **6,** 22 (2014).
19. Berger, M. F. & Mardis, E. R. The emerging clinical relevance of genomics in cancer medicine. *Nat. Rev. Clin. Oncol.* **15,** 353–365 (2018).
20. Townsend, A. R. M., Gotch, F. M. & Davey, J. Cytotoxic T cells recognize fragments of the influenza nucleoprotein. *Cell* **42,** 457–467 (1985).
21. Babbitt, B. P., Allen, P. M., Matsueda, G., Haber, E. & Unanue, E. R. Binding of immunogenic peptides to Ia histocompatibility molecules. *Nature* **317,** 359–361 (1985).
22. Simpson, A. J. G., Caballero, O. L., Jungbluth, A., Chen, Y.-T. & Old, L. J. Cancer/testis antigens, gametogenesis and cancer. *Nat. Rev. Cancer* **5,** 615–625 (2005).
23. van der Bruggen, P. *et al.* A gene encoding an antigen recognized by cytolytic T

lymphocytes on a human melanoma. *Science* **254,** 1643–1647 (1991).

24. De Plaen, E. *et al.* Immunogenic (tum-) variants of mouse tumor P815: cloning of the gene of tum- antigen P91A and identification of the tum- mutation. *Proc. Natl. Acad. Sci. U. S. A.* **85,** 2274–2278 (1988).
25. Monach, P. A., Meredith, S. C., Siegel, C. T. & Schreiber, H. A unique tumor antigen produced by a single amino acid substitution. *Immunity* **2,** 45–59 (1995).
26. Dubey, P. *et al.* The Immunodominant Antigen of an Ultraviolet-induced Regressor Tumor Is Generated by a Somatic Point Mutation in the DEAD Box Helicase p68. *J. Exp. Med.* **185,** 695–706 (1997).
27. Sahin, U. *et al.* Human neoplasms elicit multiple specific immune responses in the autologous host. *Proceedings of the National Academy of Sciences* **92,** 11810–11813 (1995).
28. Khan, G., E., S., Denniss, F., Sigurdardottir, D. & Gui, B.-A. Identification and Validation of Targets for Cancer Immunotherapy: From the Bench-to-Bedside. in *Novel Gene Therapy Approaches* (2013).
29. Caron, M., Choquet-Kastylevsky, G. & Joubert-Caron, R. Cancer immunomics using autoantibody signatures for biomarker discovery. *Mol. Cell. Proteomics* **6,** 1115–1122 (2007).
30. Martelange, V., De Smet, C., De Plaen, E., Lurquin, C. & Boon, T. Identification on a human sarcoma of two new genes with tumor-specific expression. *Cancer Res.* **60,** 3848–3855 (2000).
31. Klade, C. S. *et al.* Identification of tumor antigens in renal cell carcinoma by serological proteome analysis. *Proteomics* **1,** 890–898 (2001).
32. Schumacher, T. N. & Schreiber, R. D. Neoantigens in cancer immunotherapy. *Science* **348,** 69–74 (2015).

33. Zhou, J., Dudley, M. E., Rosenberg, S. A. & Robbins, P. F. Persistence of multiple tumor-specific T-cell clones is associated with complete tumor regression in a melanoma patient receiving adoptive cell transfer therapy. *J. Immunother.* **28,** 53–62 (2005).

34. Lennerz, V. *et al.* The response of autologous T cells to a human melanoma is dominated by mutated neoantigens. *Proceedings of the National Academy of Sciences* **102,** 16013–16018 (2005).

35. Gubin, M. M., Artyomov, M. N., Mardis, E. R. & Schreiber, R. D. Tumor neoantigens: building a framework for personalized cancer immunotherapy. *J. Clin. Invest.* **125,** 3413–3421 (2015).

36. Segal, N. H. *et al.* Epitope landscape in breast and colorectal cancer. *Cancer Res.* **68,** 889–892 (2008).

37. Castle, J. C. *et al.* Exploiting the Mutanome for Tumor Vaccination. *Cancer Res.* **72,** 1081–1091 (2012).

38. Matsushita, H. *et al.* Cancer exome analysis reveals a T-cell-dependent mechanism of cancer immunoediting. *Nature* **482,** 400–404 (2012).

39. Gubin, M. M. *et al.* Checkpoint blockade cancer immunotherapy targets tumour-specific mutant antigens. *Nature* **515,** 577–581 (2014).

40. van Rooij, N. *et al.* Tumor Exome Analysis Reveals Neoantigen-Specific T-Cell Reactivity in an Ipilimumab-Responsive Melanoma. *J. Clin. Oncol.* **31,** e439–e442 (2013).

41. Robbins, P. F. *et al.* Mining exomic sequencing data to identify mutated antigens recognized by adoptively transferred tumor-reactive T cells. *Nat. Med.* **19,** 747–752 (2013).

42. Rajasagi, M. *et al.* Systematic identification of personal tumor-specific neoantigens in

chronic lymphocytic leukemia. *Blood* **124,** 453–462 (2014).

43. Linnemann, C. *et al.* High-throughput epitope discovery reveals frequent recognition of neo-antigens by CD4+ T cells in human melanoma. *Nat. Med.* **21,** 81–85 (2015).
44. Houghton, A. N. & Guevara-Patiño, J. A. Immune recognition of self in immunity against cancer. *J. Clin. Invest.* **114,** 468–471 (2004).
45. Liu, X. S., Shirley Liu, X. & Mardis, E. R. Applications of Immunogenomics to Cancer. *Cell* **168,** 600–612 (2017).
46. Bauer, D. C., Zadoorian, A., Wilson, L. O. W. & Thorne, N. P. Evaluation of computational programs to predict HLA genotypes from genomic sequencing data. *Brief. Bioinform.* **19,** 179–187 (2018).
47. Shukla, S. A. *et al.* Comprehensive analysis of cancer-associated somatic mutations in class I HLA genes. *Nat. Biotechnol.* **33,** 1152–1158 (2015).
48. Szolek, A. *et al.* OptiType: precision HLA typing from next-generation sequencing data. *Bioinformatics* **30,** 3310–3316 (2014).
49. Liu, C. *et al.* ATHLATES: accurate typing of human leukocyte antigen through exome sequencing. *Nucleic Acids Res.* **41,** e142 (2013).
50. Warren, R. L. *et al.* Derivation of HLA types from shotgun sequence datasets. *Genome Med.* **4,** 95 (2012).
51. Hundal, J. *et al.* Cancer Immunogenomics: Computational Neoantigen Identification and Vaccine Design. *Cold Spring Harb. Symp. Quant. Biol.* **81,** 105–111 (2016).
52. Backert, L. & Kohlbacher, O. Immunoinformatics and epitope prediction in the age of genomic medicine. *Genome Med.* **7,** 119 (2015).
53. DiBrino, M. *et al.* Identification of the Peptide Binding Motif for HLA-B44, One of the Most Common HLA-B Alleles in the Caucasian Population. *Biochemistry* **34,** 10130–10138 (1995).

54. Rammensee, H.-G. *et al.* SYFPEITHI: database for MHC ligands and peptide motifs. *Immunogenetics* **50,** 213–219 (1999).

55. Reche, P. A., Glutting, J.-P. & Reinherz, E. L. Prediction of MHC class I binding peptides using profile motifs. *Hum. Immunol.* **63,** 701–709 (2002).

56. Andreatta, M. & Nielsen, M. Gapped sequence alignment using artificial neural networks: application to the MHC class I system. *Bioinformatics* **32,** 511–517 (2016).

57. Nielsen, M. *et al.* Reliable prediction of T-cell epitopes using neural networks with novel sequence representations. *Protein Sci.* **12,** 1007–1017 (2003).

58. Hoof, I. *et al.* NetMHCpan, a method for MHC class I binding prediction beyond humans. *Immunogenetics* **61,** 1–13 (2008).

59. Zhang, H., Lund, O. & Nielsen, M. The PickPocket method for predicting binding specificities for receptors based on receptor pocket similarities: application to MHC-peptide binding. *Bioinformatics* **25,** 1293–1299 (2009).

60. Karosiene, E., Lundegaard, C., Lund, O. & Nielsen, M. NetMHCcons: a consensus method for the major histocompatibility complex class I predictions. *Immunogenetics* **64,** 177–186 (2012).

61. O'Donnell, T. J. *et al.* MHCflurry: Open-Source Class I MHC Binding Affinity Prediction. *Cell Syst* **7,** 129–132.e4 (2018).

62. Bhattacharya, R. *et al.* Evaluation of machine learning methods to predict peptide binding to MHC Class I proteins. (2017). doi:10.1101/154757

63. Peters, B. & Sette, A. 10.1186/1471-2105-6-132. *BMC Bioinformatics* **6,** 132 (2005).

64. Kim, Y., Sidney, J., Pinilla, C., Sette, A. & Peters, B. Derivation of an amino acid similarity matrix for peptide: MHC binding and its application as a Bayesian prior. *BMC Bioinformatics* **10,** 394 (2009).

65. Nielsen, M. *et al.* Quantitative Predictions of Peptide Binding to Any HLA-DR

Molecule of Known Sequence: NetMHCIIpan. *PLoS Comput. Biol.* **4,** e1000107 (2008).

66. Snyder, A. *et al.* Genetic basis for clinical response to CTLA-4 blockade in melanoma. *N. Engl. J. Med.* **371,** 2189–2199 (2014).
67. Rizvi, N. A. *et al.* Cancer immunology. Mutational landscape determines sensitivity to PD-1 blockade in non-small cell lung cancer. *Science* **348,** 124–128 (2015).
68. Zhang, X., Sharma, P. K., Peter Goedegebuure, S. & Gillanders, W. E. Personalized cancer vaccines: Targeting the cancer mutanome. *Vaccine* **35,** 1094–1100 (2017).
69. Ott, P. A. *et al.* An immunogenic personal neoantigen vaccine for patients with melanoma. *Nature* **547,** 217–221 (2017).
70. Sahin, U. *et al.* Personalized RNA mutanome vaccines mobilize poly-specific therapeutic immunity against cancer. *Nature* **547,** 222–226 (2017).
71. Carreno, B. M. *et al.* A dendritic cell vaccine increases the breadth and diversity of melanoma neoantigen-specific T cells. *Science* **348,** 803–808 (2015).
72. Linette, G. P. & Carreno, B. M. Neoantigen Vaccines Pass the Immunogenicity Test. *Trends Mol. Med.* **23,** 869–871 (2017).
73. Tran, E. *et al.* Cancer Immunotherapy Based on Mutation-Specific CD4 T Cells in a Patient with Epithelial Cancer. *Science* **344,** 641–645 (2014).
74. Rosenberg, S. A. *et al.* Durable complete responses in heavily pretreated patients with metastatic melanoma using T-cell transfer immunotherapy. *Clin. Cancer Res.* **17,** 4550–4557 (2011).
75. Boon, T. Tumor Antigens Recognized by T Lymphocytes. *Annu. Rev. Immunol.* **12,** 337–365 (1994).
76. Trajanoski, Z. *et al.* Somatically mutated tumor antigens in the quest for a more efficacious patient-oriented immunotherapy of cancer. *Cancer Immunol. Immunother.* **64,** 99–104 (2015).

77. Houghton, A. N. & Guevara-Patiño, J. A. Immune recognition of self in immunity against cancer. *J. Clin. Invest.* **114,** 468–471 (2004).

78. Reche, P. A., Glutting, J.-P. & Reinherz, E. L. Prediction of MHC class I binding peptides using profile motifs. *Hum. Immunol.* **63,** 701–709 (2002).

79. Bhasin, M. & Raghava, G. P. S. A hybrid approach for predicting promiscuous MHC class I restricted T cell epitopes. *J. Biosci.* **32,** 31–42 (2007).

80. Lundegaard, C. *et al.* NetMHC-3.0: accurate web accessible predictions of human, mouse and monkey MHC class I affinities for peptides of length 8–11. *Nucleic Acids Res.* **36,** W509–W512 (2008).

81. Soria-Guerra, R. E., Nieto-Gomez, R., Govea-Alonso, D. O. & Rosales-Mendoza, S. An overview of bioinformatics tools for epitope prediction: implications on vaccine development. *J. Biomed. Inform.* **53,** 405–414 (2015).

82. Vita, R. *et al.* The immune epitope database (IEDB) 3.0. *Nucleic Acids Res.* **43,** D405–D412 (2014).

83. Duarte, A. *et al.* Prediction of CD8 Epitopes in Leishmania braziliensis Proteins Using EPIBOT: In Silico Search and In Vivo Validation. *PLoS One* **10,** e0124786 (2015).

84. Schubert, B., Brachvogel, H.-P., Jürges, C. & Kohlbacher, O. EpiToolKit—a web-based workbench for vaccine design. *Bioinformatics* **31,** 2211–2213 (2015).

85. Duan, F. *et al.* Genomic and bioinformatic profiling of mutational neoepitopes reveals new rules to predict anticancer immunogenicity. *J. Exp. Med.* **211,** 2231–2248 (2014).

86. Griffith, M. *et al.* Genome Modeling System: A Knowledge Management Platform for Genomics. *PLoS Comput. Biol.* **11,** e1004274 (2015).

87. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25,** 1754–1760 (2009).

88. Picard.

89. Warren, R. L. & Holt, R. A. Targeted assembly of short sequence reads. *PLoS One* **6,** e19816 (2011).

90. Robinson, J. *et al.* The IMGT/HLA database. *Nucleic Acids Res.* **41,** D1222–D1227 (2012).

91. Website. Available at: Hercus, C. 2009. 'Novocraft Short Read Alignment Package.' www.novocraft.com. (Accessed: 12th December 2018)

92. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26,** 841–842 (2010).

93. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25,** 2078–2079 (2009).

94. Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27,** 2987–2993 (2011).

95. Larson, D. E. *et al.* SomaticSniper: identification of somatic point mutations in whole genome sequencing data. *Bioinformatics* **28,** 311–317 (2012).

96. Koboldt, D. C. *et al.* VarScan: variant detection in massively parallel sequencing of individual and pooled samples. *Bioinformatics* **25,** 2283–2285 (2009).

97. Koboldt, D. C. *et al.* VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res.* **22,** 568–576 (2012).

98. Saunders, C. T. *et al.* Strelka: accurate somatic small-variant calling from sequenced tumor–normal sample pairs. *Bioinformatics* **28,** 1811–1817 (2012).

99. Flicek, P. *et al.* Ensembl 2013. *Nucleic Acids Res.* **41,** D48–55 (2013).

100. Zhang, H., Lundegaard, C. & Nielsen, M. Pan-specific MHC class I predictors: a benchmark of HLA class I pan-specific prediction methods. *Bioinformatics* **25,** 83–89 (2008).

101. Lundegaard, C., Lund, O. & Nielsen, M. Prediction of epitopes using neural network based methods. *J. Immunol. Methods* **374,** 26–34 (2011).

102. Nielsen, M. *et al.* NetMHCpan, a Method for Quantitative Predictions of Peptide Binding to Any HLA-A and -B Locus Protein of Known Sequence. *PLoS One* **2,** e796 (2007).

103. Trapnell, C., Pachter, L. & Salzberg, S. L. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25,** 1105–1111 (2009).

104. Kim, D. *et al.* TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* **14,** R36 (2013).

105. Trapnell, C. *et al.* Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc.* **7,** 562–578 (2012).

106. Dodt, M., Roehr, J. T., Ahmed, R. & Dieterich, C. FLEXBAR-Flexible Barcode and Adapter Processing for Next-Generation Sequencing Platforms. *Biology* **1,** 895–905 (2012).

107. bam-readcount.

108. Robinson, J. T. *et al.* Integrative genomics viewer. *Nat. Biotechnol.* **29,** 24–26 (2011).

109. Thorvaldsdottir, H., Robinson, J. T. & Mesirov, J. P. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief. Bioinform.* **14,** 178–192 (2012).

110. Hackl, H., Charoentong, P., Finotello, F. & Trajanoski, Z. Computational genomics tools for dissecting tumour–immune cell interactions. *Nat. Rev. Genet.* **17,** 441–458 (2016).

111. Hundal, J. *et al.* pVAC-Seq: A genome-guided in silico approach to identifying tumor neoantigens. *Genome Med.* **8,** 11 (2016).

112. Bjerregaard, A.-M., Nielsen, M., Hadrup, S. R., Szallasi, Z. & Eklund, A. C. MuPeXI:

prediction of neo-epitopes from tumor sequencing data. *Cancer Immunol. Immunother.* (2017). doi:10.1007/s00262-017-2001-3

113. Rubinsteyn, A., Hodes, I., Kodysh, J. & Hammerbacher, J. Vaxrank: A Computational Tool For Designing Personalized Cancer Vaccines. (2017). doi:10.1101/142919
114. Meydan, C., Otu, H. H. & Sezerman, O. U. Prediction of peptides binding to MHC class I and II alleles by temporal motif mining. *BMC Bioinformatics* **14 Suppl 2,** S13 (2013).
115. Rammensee, H. G., Friede, T. & Stevanoviíc, S. MHC ligands and peptide motifs: first listing. *Immunogenetics* **41,** 178–228 (1995).
116. Griffith, M. *et al.* Optimizing cancer genome sequencing and analysis. *Cell Syst* **1,** 210–223 (2015).
117. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics* **26,** 589–595 (2010).
118. Griffith, M. *et al.* Comprehensive genomic analysis reveals FLT3 activation and a therapeutic strategy for a patient with relapsed adult B-lymphoblastic leukemia. *Exp. Hematol.* **44,** 603–613 (2016).
119. Wagner, A. H. *et al.* Recurrent WNT pathway alterations are frequent in relapsed small cell lung cancer. *Nat. Commun.* **9,** 3787 (2018).
120. Barnell, E. K. *et al.* Standard operating procedure for somatic variant refinement of tumor sequencing data. (2018). doi:10.1101/266262
121. Van der Auwera, G. A. *et al.* From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr. Protoc. Bioinformatics* **43,** 11.10.1–33 (2013).
122. Charoentong, P. *et al.* Pan-cancer Immunogenomic Analyses Reveal Genotype-Immunophenotype Relationships and Predictors of Response to Checkpoint Blockade. *Cell Rep.* **18,** 248–262 (2017).

123. Chicz, R. M. *et al.* Predominant naturally processed peptides bound to HLA-DR1 are derived from MHC-related molecules and are heterogeneous in size. *Nature* **358,** 764–768 (1992).

124. Poplin, R. *et al.* Scaling accurate genetic variant discovery to tens of thousands of samples. (2017). doi:10.1101/201178

125. Łuksza, M. *et al.* A neoantigen fitness model predicts tumour response to checkpoint blockade immunotherapy. *Nature* **551,** 517–520 (2017).

126. Sette, A. *et al.* The relationship between class I binding affinity and immunogenicity of potential cytotoxic T cell epitopes. *J. Immunol.* **153,** 5586–5592 (1994).

127. Turajlic, S. *et al.* Insertion-and-deletion-derived tumour-specific neoantigens and the immunogenic phenotype: a pan-cancer analysis. *Lancet Oncol.* (2017). doi:10.1016/S1470-2045(17)30516-8

128. Lesurf, R. *et al.* Genomic characterization of HER2-positive breast cancer and response to neoadjuvant trastuzumab and chemotherapy-results from the ACOSOG Z1041 (Alliance) trial. *Ann. Oncol.* **28,** 1070–1077 (2017).

129. Johanns, T. M. *et al.* Immunogenomics of Hypermutated Glioblastoma: A Patient with Germline POLE Deficiency Treated with Checkpoint Blockade Immunotherapy. *Cancer Discov.* **6,** 1230–1236 (2016).

130. Bais, P., Namburi, S., Gatti, D. M., Zhang, X. & Chuang, J. H. CloudNeo: a cloud pipeline for identifying patient-specific tumor neoantigens. *Bioinformatics* **33,** 3110–3112 (2017).

131. Jurtz, V. *et al.* NetMHCpan-4.0: Improved Peptide–MHC Class I Interaction Predictions Integrating Eluted Ligand and Peptide Binding Affinity Data. *The Journal of Immunology* **199,** 3360–3368 (2017).

132. Hundal, J. *et al.* Accounting for proximal variants improves neoantigen prediction. *Nat.*

Genet. (2018). doi:10.1038/s41588-018-0283-9

133. Turajlic, S. *et al.* Insertion-and-deletion-derived tumour-specific neoantigens and the immunogenic phenotype: a pan-cancer analysis. *Lancet Oncol.* **18,** 1009–1021 (2017).

134. Schubert B, E. al. FRED 2: an immunoinformatics framework for Python. - PubMed - NCBI. Available at: https://www.ncbi.nlm.nih.gov/pubmed/27153717. (Accessed: 11th December 2018)

135. Vita, R. *et al.* The Immune Epitope Database 2.0. *Nucleic Acids Res.* **38,** D854–D862 (2009).

136. Paul, S. *et al.* HLA Class I Alleles Are Associated with Peptide-Binding Repertoires of Different Size, Affinity, and Immunogenicity. *The Journal of Immunology* **191,** 5831–5839 (2013).

137. Keşmir, C., Nussbaum, A. K., Schild, H., Detours, V. & Brunak, S. Prediction of proteasome cleavage motifs by neural networks. *Protein Eng.* **15,** 287–296 (2002).

138. Jørgensen, K. W., Rasmussen, M., Buus, S. & Nielsen, M. NetMHCstab - predicting stability of peptide-MHC-I complexes; impacts for cytotoxic T lymphocyte epitope discovery. *Immunology* **141,** 18–26 (2014).

139. Zhang, J., Mardis, E. R. & Maher, C. A. INTEGRATE-neo: a pipeline for personalized gene fusion neoantigen discovery. *Bioinformatics* **33,** 555–557 (2017).

140. Schubert, B. & Kohlbacher, O. Designing string-of-beads vaccines with optimal spacers. *Genome Med.* **8,** 9 (2016).

141. Miller, A. *et al.* High somatic mutation and neoantigen burden are correlated with decreased progression-free survival in multiple myeloma. *Blood Cancer J.* **7,** e612 (2017).

142. Balachandran, V. P. *et al.* Identification of unique neoantigen qualities in long-term survivors of pancreatic cancer. *Nature* **551,** 512 (2017).

143. Formenti, S. C. *et al.* Radiotherapy induces responses of lung cancer to CTLA-4 blockade. *Nat. Med.* **24,** 1845 (2018).

144. Koboldt, D. C., Larson, D. E. & Wilson, R. K. Using VarScan 2 for Germline Variant Calling and Somatic Mutation Detection. in *Current Protocols in Bioinformatics* 15.4.1–15.4.17 (2013).

145. DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43,** 491–498 (2011).

146. Peter Amstutz, Michael R. Crusoe, Nebojša Tijanić (editors), Brad Chapman, John Chilton, Michael Heuer, Andrey Kartashov, Dan Leehr, Hervé Ménager, Maya Nedeljkovich, Matt Scales, Stian Soiland-Reyes, Luka Stojanovic. Common Workflow Language, v1.0. (2016). doi:10.6084/m9.figshare.3115156.v2

147. Voss K, G. J. A. V. der A. G. Full-stack genomics pipelining with GATK4 + WDL + Cromwell [version 1; not peer reviewed]. *F1000Res.* (2017). doi:10.7490/f1000research.1114631.1

148. Cabanski, C. R. *et al.* cDNA hybrid capture improves transcriptome analysis on low-input and archived samples. *J. Mol. Diagn.* **16,** 440–451 (2014).

149. Cibulskis, K. *et al.* Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat. Biotechnol.* **31,** 213–219 (2013).

150. Ye, K., Schulz, M. H., Long, Q., Apweiler, R. & Ning, Z. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics* **25,** 2865–2871 (2009).

151. Tan, A., Abecasis, G. R. & Kang, H. M. Unified representation of genetic variants. *Bioinformatics* **31,** 2202–2204 (2015).

152. Roehr, J. T., Dieterich, C. & Reinert, K. Flexbar 3.0 - SIMD and multicore parallelization. *Bioinformatics* **33,** 2941–2942 (2017).

153. Kim, D., Langmead, B. & Salzberg, S. L. HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods* **12,** 357–360 (2015).

154. Pertea, M. *et al.* StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* **33,** 290–295 (2015).

155. Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.* **34,** 525–527 (2016).

156. Brown, S. D. *et al.* Neo-antigens predicted by tumor genome meta-analysis correlate with increased patient survival. *Genome Res.* **24,** 743–750 (2014).

157. Pritchard, A. L. *et al.* Exome Sequencing to Predict Neoantigens in Melanoma. *Cancer Immunol Res* **3,** 992–998 (2015).