UC Santa Cruz

UC Santa Cruz Electronic Theses and Dissertations

Title

Prediction Of T-cell Epitopes for Cancer Therapy

Permalink

https://escholarship.org/uc/item/0pk3b6wh

Author

Rao, Arjun Arkal

Publication Date

2018

Supplemental Material

https://escholarship.org/uc/item/0pk3b6wh#supplemental

Copyright Information

This work is made available under the terms of a Creative Commons Attribution-ShareAlike License, availalbe at https://creativecommons.org/licenses/by-sa/4.0/

Peer reviewed|Thesis/dissertation

eScholarship.org Powered by the California Digital Library
University of California

UNIVERSITY OF CALIFORNIA
SANTA CRUZ

PREDICTION OF T-CELL EPITOPES FOR CANCER THERAPY

A dissertation submitted in partial satisfaction of the
requirements for the degree of

DOCTOR OF PHILOSOPHY

in

BIOINFORMATICS

by

Arjun Arkal Rao

June 2018

The Dissertation of Arjun Arkal Rao
is approved:

Professor David Haussler, Chair

Professor Phillip Berman

Professor Nikolaos Sgourakis

Dean Tyrus Miller
Vice Provost and Dean of Graduate Studies

Copyright © by

Arjun Arkal Rao

2018

Table of Contents

List of Figures v

List of Tables viii

Abstract ix

Dedication xi

Acknowledgments xii

1 Introduction 1
- 1.1 Introduction . . . 1
- 1.2 Cancer and the immune system . . . 3
- 1.3 Innate and Adaptive Immunity . . . 4
- 1.4 Antigen presentation, MHC, and the T-cell response . . . 7
- 1.5 Cancer Immunoediting . . . 10
- 1.6 Cancer immunotherapy . . . 10
  - 1.6.1 Antibody-based therapies . . . 11
  - 1.6.2 Checkpoint Blockade . . . 12
  - 1.6.3 Adoptive cell therapy . . . 12
  - 1.6.4 Vaccine-based therapies . . . 16
- 1.7 Thesis statement . . . 16

2 Efficient workflow scheduling using Toil 18
- 2.1 Introduction . . . 18
- 2.2 The Toil paper . . . 19
- 2.3 The Toil paper supplementary . . . 23
  - 2.3.1 Supplementary Toil documentation, chapter 8 . . . 23

3 Translation of genomic mutations into therapeutically assayable peptides 28
- 3.1 Introduction . . . 28
- 3.2 The TransGene paper . . . 30

3.2.1 Abstract . . . 30
3.2.2 Introduction . . . 30
3.2.3 The TransGene workflow . . . 31
3.2.4 Application of TransGene to a test cohort . . . 34
3.2.5 Discussion . . . 35
3.3 The TransGene paper supplementary . . . 36
3.3.1 Supplementary Figures . . . 36
3.3.2 Supplementary Tables . . . 38

4 ProTECT: Prediction of T-cell Epitopes for Cancer Therapy 39
4.1 Introduction . . . 39
4.2 The ProTECT paper . . . 41
4.2.1 Abstract . . . 41
4.2.2 Introduction . . . 42
4.2.3 The ProTECT pipeline . . . 44
4.2.4 Materials . . . 50
4.2.5 Methods . . . 52
4.2.6 Results and Discussion . . . 53
4.2.7 Conclusion . . . 67
4.3 Supplementary information . . . 68
4.3.1 Supplementary Note 1 . . . 68
4.3.2 Supplementary Tables . . . 69
4.3.3 Supplementary Figures . . . 70

5 Identification of a potentially therapeutic hotspot neoepitope in Pediatric Neuroblastoma 75
5.1 Introduction . . . 75
5.2 The pediatric NBL study paper . . . 77
5.3 The pediatric NBL study supplementary . . . 95
5.3.1 Supplementary Table 1 . . . 95
5.3.2 Supplementary References . . . 97

6 Conclusion 99
6.1 Introduction . . . 99
6.2 Chapters . . . 99
6.3 Discussion . . . 101

7 Appendix 103
7.1 Figures . . . 103

Bibliography 105

List of Figures

1.1 The adaptive immune response begins with the activation of a dendritic cell. Activated CTLs and NK cells directly attack the infected cell. T-Helper cells interact with CTLs, B-cells, and other T-Helper cells. B cells produce antibodies which enable phagocytes to sense the infected cell. Figure obtained from Lambotin, et.al., Nature Reviews Microbiology [64] and modified to retain only relevant information. 6
1.2 Schematic representation of the methods of antigen presentation in cells. a) Intracellular proteins are processed via the MHCI pathway and are displayed to $CD8^+$ cells. b) Extracellular antigens are endocytosed and processed via the MHCII pathway in APCs before being displayed to $CD4^+$ cells. c) Dendritic cells can also "cross-present" extracellular antigens via the MHCI pathway. Figure obtained from Heath and Carbone, Nature Reviews Immunology [45]. . . . 8
1.3 The Three E's of Immunoediting; Elimination, Equilibrium, and Escape. Figure obtained from Dunn et.al., Nature Reviews Immunology [29]. 11
1.4 Schematic representation of adoptive T-cell therapy. Mutations in coding regions of the tumor DNA can be scanned for potential aberrant protein products and used to activate autologous T-cells. Figure obtained from Restifo et.al., Nature Reviews Immunology [93]. 14
1.5 Schematic representation of Engineered T-cells used in adoptive T-cell therapy. T-cells can be engineered with Chimeric Antigen Receptors(CARs) or T-cell Receptors grown in-vitro to enhance the immune reaction. Figure obtained from Restifo et.al., Nature Reviews Immunology [93]. 15
1.6 Schematic representation of a peptide vaccine workflow for cancer therapy. Figure obtained from Sahin and Türeci, Science [98]. 16

3.1 A) The TransGene workflow. B) A cartoon describing the output ImmunoActive Regions (IARs) for 4 mutation events. C) A cartoon describing how transgene handles mutations near exon boundaries. It also describes how transgene handles co-expressed mutations when RNA-Seq data is present. Transcript 1 produces no IARs since there are no reads supporting junctions E1:E2 and E2:E4, Transcript 2 produces 2 IARs from junctions E1:E3 and E3:E4, and Transcript 3 produces 2 IARs from junction E1:E4 (one with both mutations and one with only the mutant from E1 based on the read support/VAF) 32
3.2 An IGV screenshot of chromosome X for sample TCGA-CH-5792. The mutations at positions 48816540 (T>G) and 48816541 (G>C) affect different codons and are fully phased, hence they affect the same IAR at consecutive residues. 36
3.3 An IGV screenshot of chromosome 8 for sample TCGA-EJ-5525. The mutations at positions 135542701 (G>T) and 135542702 (C>T) affect different codons and are fully phased, hence they affect the same IAR at consecutive residues. 37

4.1 A schematic description of the ProTECT workflow. ProTECT can process FASTQs all the way through the prediction of ImmunoActive Regions, including alignment, HLA Haplotyping, variant calling, expression estimation, mutation translation, and pMHC binding affinity prediction. ProTECT also allows users to provide pre-computed inputs for various steps instead. 45
4.2 HLA Haplotypes called by ProTECT (using PHLAT) are fully concordant with POLYSOLVER haplotypes in only 67.5% of samples. 28.8% differ by 1 call and 3.7% by 2 calls. A majority of the miscalled HLA-A alleles are a documented PHLAT artifact. 61
4.3 Average runtimes on our cluster when ProTECT is run in a batch of 'n' samples. Each batch size is run with 5 unique sample sets and the range of runtimes is described by the whiskers at each datapoint. The grey bar describes the result of running ProTECT on a single sample on one machine. ProTECT takes considerably less time on average when run in a large group. 62
4.4 Fusion calls between ProTECT (STAR + Fusion-Inspector) are not concordant with INTEGRATE. A large number of INTEGRATE fusions have read support <5 (left) however some of these are called by ProTECT with >5 support (right). 65
4.5 HLA haplotypes called by HLAMiner in the INTEGRATE-Neo paper have very low overlap with ProTECT and POLYSOLVER. 66
4.6 ProTECT rejects 100/720 INTEGRATE calls for being transcriptional readthroughs (92) or for having a 5' non-coding RNA partner (8) (Left). ProTECT predicts 137 of the expected 155 epitopes called by INTEGRATE-Neo (Right). 66
4.7 MHC alleles called in samples with the chr21:41498119-chr21:38445621 TMPRSS2-ERG breakpoint 71

4.8 A UCSC genome browser showing the 5' breakpoint (highlighted) for the two missed epitopes from the ENSG00000231887-ENSG00000003056 fusion. The 5' partner was reported as PRH1 but the screenshot shows that the position is in the 5' UTR for PRH1. The overlapping PRR4 contains the epitope predicted by INTEGRATE-Neo. 72
4.9 A UCSC genome browser showing the 5' breakpoint (highlighted) for the two missed epitopes from the ENSG00000273294-ENSG00000164182 fusion. The 5' partner was reported as the readthrough transcript C1QTNF3-AMACR but the screenshot shows that the position is in the 5' UTR for C1QTNF3-AMACR. The overlapping AMACR contains the epitope predicted by INTEGRATE-Neo. 73
4.10 MHC alleles called in samples with the chr21:41507950-chr21:38445621 THMPRSS2-ERG breakpoint 74

7.1 A screengrab of the 'Contributors' tab on the BD2KGenomics/toil repository from the date of my first commit to the time of publication showing my substantial contribution. 104

List of Tables

4.1 Statistics for 323 samples with at least 1 accepted variant. We predict at least 1 MHCI IAR in 319 samples with a median of 11 per sample. As expected, SNVs are the dominant variant type. 55

4.2 Recurrent fusions called by ProTECT. PRAD is characterized by an abundance of TMPRSS2 fusions with genes in the ETV family (TMPRSS2-ERG being the most popular) 56

4.3 Recurrent TMPRSS2-ERG breakpoints in the cohort. IARs from 21:41498119-21:38445621 and 21:41507950-21:38445621 are recurrent suggesting their viability universal peptide vaccine candidates. We do not expect to see an IAR from fusions with 5'UTR breakpoints. [1] TransGene cannot handle de novo splice acceptors. [2] An Epitope will exist where the TMPRSS2 reads into the intron of ERG. [3] A frameshift is seen on the ERG side of the fusion. 57

4.4 Recurrent mutants in the SPOP gene target 3 codons. The F133V/C/I/L mutant may be of interest as a universal neoepitope due to the similar chemical properties of Leucine, Isoleucine and Valine. 59

4.5 Predicted binding affinities (better than *n* percent of a background set) of 9-mers arising from the SPOP mutants affecting p.F133 (FVQGKDWG **X** KKFIRRDF for X={V, I, L}) to HLA-A*02:01. The similar chemical properties of Valine, Leucine and Isoleucine lead to similar binding predictions of neoepitopes substituting them for Phenylalanine. Wildtype epitope affinity for reference. 60

4.6 ProTECT ranks for the validation data of PVAC-Seq. Some patients had biopsies taken at multiple time points. Green: Dominant epitope, Orange: Subdominant epitope, Red: Cryptic epitope. Non-highlighted ranks describe instances where PVAC-Seq failed to call the epitope. 64

Abstract

Prediction of T-cell Epitopes for Cancer Therapy

by

Arjun Arkal Rao

The human immune system can identify malfunctioning, damaged, or infected cells within the body. The integrity of nucleated cells is communicated to the immune system through short peptides which they display on their exterior using MHC proteins. Mutated peptides produced in cancer cells due to single nucleotide events, short insertion or deletion events, or due to large chromosomal rearrangements can help the immune system identify tumor cells as threats. Adoptive T-cell and Vaccine-based therapies that stimulate autologous T-cells to treat a patient's disease are more specific than standard chemo- and radiotherapies because they directly target tumor-specific mutations. They also potentially protect the body from recurrence of the tumor due to their unique memory potential. However, T-cells are often perceptive to only those cells originating from the patient's body, making this a personalized treatment for the disease. Rapid identification of tumor neoepitopes in a patient diagnosed with cancer will greatly reduce the time taken to develop a personalized therapy.

This thesis primarily covers ProTECT, a computational workflow for the Prediction of Epitopes for Cancer Therapy. ProTECT is an automated, scalable, and reproducible end-to-end pipeline that processes patient sequencing data to produce a ranked list of neoepitopes of therapeutic significance. ProTECT also attempts to predict the state of the MHC presentation pathway in the tumor, and the tumor response to immunotherapies based on previously

published immune signatures. I describe ProTECT and demonstrate its features using a cohort of 326 Prostate Adenocarcinoma patients. I then demonstrate its clinical utility through the identification of a neoepitope arising from a hotspot mutant in pediatric Neuroblastoma.

To enable ProTECT, I worked on two auxiliary projects, Toil and TransGene. Toil is a workflow manager developed at the Genomics Institute at UCSC, and is the backbone of the ProTECT workflow. I spent some of my time at UCSC working on enhancing the efficiency of Toil workflows. TransGene is a tool I wrote that translates genomic variants (single nucleotide, short insertions and deletions, and fusion genes) into peptides that are compatible with existing peptide:MHC binding affinity predictors. These methods are also included in this thesis.

To my Mother who did everything she could to get me here, who died a few days after I was accepted at UCSC; and to my Wife who refused to give up on my dream of attending UCSC, when even I had done so.

Acknowledgments

In addition to my committee, I would like to thank my family, and all the friends I've made over the past 5 years, the following people deserve special recognition.

Family Members

- Divyashree Kushnoor
- Vasudev Pai
- Ramana Rao
- Ajay Rao
- Shreeja Das
- Tanuja Kushnoor
- Vikram Kushnoor
- Renuka Shenoy
- Karthikeyan Sv
- Rohit Kamath
- Poorna Shenoy

Genomic Institute members

- Sofie Salama
- Benedict Paten
- Olena Morozova
- Hannes Schmidt
- Rob Currie
- Kelly Sauder
- Lynn Brazil
- Erich Weiler
- Ada Madejska
- Jacob Pfeil
- David Stewart

Collaborators

- John Maris
- Jugmohit Toor
- Paul Robbins
- Steven Rosenberg
- Bala Ganeshan
- Giridhar Basava
- Hemant Trivedi

Ph.D. Cohort

- Ian Fiddes
- John Vivian
- Edward Rice
- Nathan Schaefer

Other important people

- Afshin Mobramaein
- Ray and Alicia
- Tracie Tucker

Chapter 1

Introduction

1.1 Introduction

Cancer is one of the leading causes of human deaths worldwide. An estimated 1.6 million new cases of cancer will be diagnosed in 2015 with about 590,000 deaths resulting from various forms of the disease [107]. Mortality rates for 2013 recorded by the Center for Disease Control show cancer as the second leading cause for death in the US, trailing behind Heart Disease by a very small margin [59]. The prognosis of the cancer depends on the tissue of origin; for instance, brain cancers have poorer prognoses than lung cancers. The standard treatment for cancers include surgical resection of the tumor (if possible), followed by chemo- or radiation therapy, which are well known to have adverse acute, and chronic side effects.

Chemotherapy is a treatment regimen that uses chemical agents to destroy cancer cells. The chemicals destroy rapidly dividing cells, a characteristic of cancers, via programmed cell death or apoptosis [76, 42]. Certain tissues in the body also have high cell division rates,

such as the hair follicles or cells of the gastrointestinal tract, are adversely affected during chemotherapy [4]. Radiation therapy uses ionizing radiation to destroy cancerous cells. Ionizing radiation irreversibly damages DNA leading to cell death via apoptotic pathways. Radiation therapy is administered using an external source of radiation such as an X-ray generator [47], or using radioactive isotopes that operate from within the patient's body [101]. While both methods are carefully administered in order to target the radiation to the tumor, healthy cells in the vicinity of the tumor can get caught in the path of the X-rays and undergo cellular damage along with the target.

Both chemo- and radiation therapies are prone to damaging healthy, normal cells in the patient. Chemotherapy is commonly associated with alopecia due to cellular damage in hair follicles, nausea due to irritation caused in the gastrointestinal mucosa. Radiation therapy has more localized effects since it is targeted directly at the tumor as compared to systemic chemotherapy. Common side effects of radiation therapy include muscle soreness, swelling, long-term muscle fibrosis, and even secondary cancers due to the cellular damage. This is a serious drawback to these therapies, especially in the cases of brain cancers, and in all pediatric cancers, where collateral damage of healthy cells could lead to physical or cognitive disabilities. The body's in-built defense mechanism, the immune system, is designed to protect the body from threats at a cellular and sub-cellular level, such as viral infection, by selectively attacking the threat with minimal damage to self-tissue. It seems logical that this system should also detect cancer cells in the body as they pose a threat to the well-being of the body.

1.2 Cancer and the immune system

Almost all cells in the body communicate with the immune system using cell-surface class I Major Histocompatibility Complex (MHC) proteins. During the process of recycling old proteins – or nascent proteins that were formed in the cell with errors in their amino acid sequence – into amino acids, cells generate short fragments of the proteins that are processed and displayed to the immune system via these MHCI molecules. The immune system is constantly on the lookout for cells displaying "non-self" peptides on MHCI molecules. Since all cells contain the same genetic information, a cell displaying peptide from a protein not encoded within that genetic information is either malfunctioning or infected by a microorganism that is producing it. Both scenarios are detrimental to the body, and thus the immune system marks such cells for destruction [78].

Mutations in cellular DNA can arise when the replication machinery in the cell is compromised, as in the case of cancers. Mutations occurring in the coding regions of the DNA can have various effects on the final protein product. Mis-sense point mutations, short insertions or deletions, and fusion gene events in the DNA can all give rise to abnormal proteins that aren't seen in healthy, normal cells. These mutations often alter the function of the protein, render them inactive, or make them constitutively active. All three effects are seen in cancer and can drive the cancer towards a more aggressive form. Cancers also have mutations accumulated over time that don't contribute towards progression (passenger mutations). Mutated cancer-specific proteins are potentially non-self molecules and the immune system is programmed to destroy cells displaying such peptides on MHCI molecules. Hence MHCI displayed peptides

in cancer cells are ideal targets for the specific, precise killing of tumor cells. To understand exactly how the immune system would target these events, and how cancer cells can evade the immune system, we first need to look into how the immune system functions.

1.3 Innate and Adaptive Immunity

The immune system is broadly classified into two subsystems, namely the innate, and the adaptive immune systems. The innate immune system comprises of natural barriers to pathogenic infiltration such as skin and mucosal surfaces, and the complement system, which mainly protects the body from bacterial infection. Natural killer cells (NK cells) detect and destroy cells displaying low or no cell surface protein MHCI, which is a common consequence of viral infections. In fact, they were first detected back in 1975 by their ability to lyse tumor cells that were MHC deficient, without prior stimulation [46]. The innate immune system provides immediate non-specific response to infection and is an ancient system that is seen in one guise or another in almost all forms of life. The adaptive immune system is more specialized and takes longer to respond to an infection. However, it is more potent and can potentially remember a pathogen and react immediately to future infections.

Dendritic cells are specialized phagocytic cells that act as a link between the innate and adaptive immune responses. Dendritic cells sense pathogenic microorganisms in tissues through cell-surface and intracellular receptors. Upon detecting a pathogen, they are activated and process pathogen-derived proteins into smaller fragments for T cell activation. Activated dendritic cells migrate to the lymph nodes and recruit T cells by presenting them with pathogen

derived antigens displayed on MHCI and MHCII molecules, and the required co-stimulatory signals [7, 100, 109]. A subset of T Cells can subsequently activate B cells, increasing the potency of the adaptive immune response.

The adaptive immune system is a complex interaction between T cells and B cells, set in motion by the disease-causing agent. T-Helper cells (also called $CD4^+$ or just CD4 T cells due to cell surface expression of CD4 protein) are activated by pathogen-derived peptides presented to them via MHCII molecules on dendritic cells. Their primary role involves the modulation of the adaptive immune response - activation of B cells and Cytotoxic T cells at first, and inhibition of immune activity after the threat has been neutralized [99, 123]. Cytotoxic T cells (also known as killer T cells, or $CD8^+$ or CD8 cells due to expression of CD8 surface protein) are activated by pathogen-derived peptides presented to them on MHCI molecules. Activated CD8 cells detect specific antigenic peptides presented on infected cells via MHCI and induce cell death by depositing perforin and granzyme molecules into the infected cells [10, 77, 112]. It is important to mention that normal, healthy tissue graft from another individual could contain "foreign" MHC molecules, MHC molecules not produced in the self-genome. These foreign MHC molecules would be detected by T-cells as foreign, and be rejected by the system even if they were displaying a self-peptide.

B cells detect antigen via surface bound antibodies (also known as B-cell receptors). Receptor bound antigen is endocytosed and processed via the MHCII pathway. Helper-T cells detect MHCII presented antigens on B cells and can activate the presenter cell by providing the required co-stimulation. B cells can be activated by an alternate mechanism that does not require T cells, however that is not discussed here. T cell-activated B cells undergo rapid proliferation

after activation during which they also undergo selection to produce B cells that bind to antigen with high specificity. Activation of B cells also induces class switching which allows B cells to produce secreted antibodies that aid in the phagocytic activity of macrophages and dendritic cells. Figure 1.1 describes the adaptive immune response after activation of a dendritic cell.

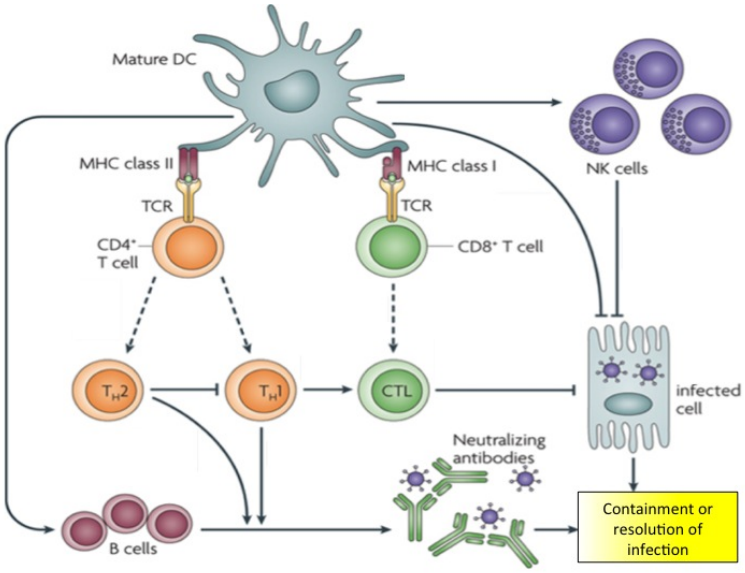


Figure 1.1: The adaptive immune response begins with the activation of a dendritic cell. Activated CTLs and NK cells directly attack the infected cell. T-Helper cells interact with CTLs, B-cells, and other T-Helper cells. B cells produce antibodies which enable phagocytes to sense the infected cell. Figure obtained from Lambotin, et.al., Nature Reviews Microbiology [64] and modified to retain only relevant information.

While B cells are highly effective at clearing out infections, they are limited by the requirement of cell-surface or secreted antigen for their action. In cancers, there can be a vast number of mutations which alter the sequence of proteins produced by the tumor. However, few (if any) of these mutations are present in the extracellular domains of cell-surface proteins and

thus the action of B cells is limited. T cells, on the other hand, can detect changes in intracellular proteins through a mechanism known as antigen presentation. The mechanism exists to detect the presence of non-self, pathogen-derived proteins in compromised cells, but works in favor of cancer therapy as T-cells would consider mutated proteins to be non-self.

1.4 Antigen presentation, MHC, and the T-cell response

The MHC- or HLA-type (Human Leukocyte Antigen) of a person is a representation of the set of MHC molecules encoded within the genome of the person. Naturally occurring variations in the MHC arise from polygeny (multiple MHC genes of the same class) and polymorphism (multiple types of the same MHC molecule). Through a combination of the two, humans have a total of 6 peptide-displaying MHC class I proteins encoded in their genome (1 each of HLA-A, HLA-B and HLA-C on each chromosome). The number of assembled peptide-displaying MHCII complexes can be larger as each HLA-DP, DQ and DR molecule depends on a combination of α and β sub-molecules in their respective loci. MHC molecules of different types differ in the type of peptide that they bind to [56, 78], hence heterozygosity (having different alleles) increases the potential number of peptides displayed on the cell surface, and the immune system's power in detecting a new non-self protein.

The short peptides produced during recycling of cellular proteins are transported into the endoplasmic reticulum (ER) of the cell via TAP transport proteins. The peptides are further processed within the ER before being loaded onto MHCI molecules and transported to the cell surface for display. The final displayed peptide is roughly 8-10 amino acids in length [56, 78].

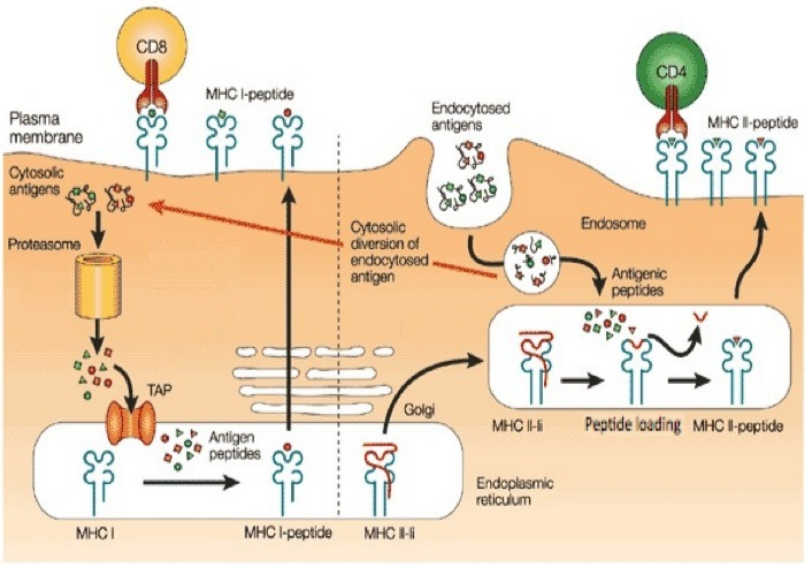


Figure 1.2: Schematic representation of the methods of antigen presentation in cells. a) Intracellular proteins are processed via the MHCI pathway and are displayed to $CD8^+$ cells. b) Extracellular antigens are endocytosed and processed via the MHCII pathway in APCs before being displayed to $CD4^+$ cells. c) Dendritic cells can also "cross-present" extracellular antigens via the MHCI pathway. Figure obtained from Heath and Carbone, Nature Reviews Immunology [45].

Figure 1.2a gives a schematic of the MHCI processing pathway. $CD8^+$ Cytotoxic T-Cells of the immune system can recognize MHCI displayed peptides and are programmed to destroy the cell if they sense that the peptide is derived from a non-self protein.

Class II MHC molecules are primarily produced by phagocytic cells. Cells that constitutively express both MHC I and II are known as Professional Antigen Presenting Cells(APCs). These include dendritic cells, macrophages, and some B cells. The role of MHCII in the immune response is the activation or "priming" of $CD4^+$ cells. MHCII displayed proteins are derived from extracellular protein endocytosed into the cell, or from dendritic cells "nibbling" on live cells [44]. The endocytosed material could be either secreted cell protein, small microorganisms, or cell proteins released into the cellular matrix after cell membrane rupture. The endocytosed protein is subjected to acid hydrolysis in endocytic vesicles and loaded onto MHCII molecules that are pre-assembled in the endoplasmic reticulum [56, 78]. Figure 1.2b shows the MHCII processing pathway of extracellular protein. $CD4^+$ cells detect MHCII displayed peptide and, given the right co-stimulation, can induce enhanced activity of the adaptive immune response.

Dendritic cells, in particular, also have a unique function known as cross presentation. In cross presentation, cells uptake extracellular pathogenic material via endocytosis and process it via the MHCI pathway instead of the MHCII pathway. In this manner, dendritic cells can prime $CD8^+$ cells for specificity to the antigen, without getting infected themselves (Figure 1.2c) [45].

1.5 Cancer Immunoediting

Cancers are frequently associated with a high degree of chromosomal instability and subsequently, a high mutational burden. The logical question to ask is, "If cancers have a high rate of mutation and produce peptides foreign to the body, why doesn't the immune system detect and destroy tumor cells?" The short answer is that the immune system can, and does mount a response to the tumor. However, tumors have ways to evade the immune system.

The tumoral response to the immune system is often referred to as the 'Three E's of cancer immunoediting', Elimination, Equilibrium, and Escape [30, 29, 62]. This idea suggests that in the early stage of tumor development, the innate and adaptive immune system react to the cancer and eliminate the cells in the growing tumor. Eventually, the cells within the tumor that do not harbor the mutations detected by the immune system gain a selective advantage and begin to grow in number. In this equilibrium state, the immune system continues to kill the cells it can target while other tumor cells slowly rise in number. At a certain point, the tumor develops enough cells to overcome the immune response and this leads to immune escape. Figure 1.3 shows the stages of Immunoediting in a cartoon form.

1.6 Cancer immunotherapy

Cancer immunotherapy involves using the immune system to treat cancers. There are 4 well defined cancer immunotherapy techniques: antibody based therapies, Checkpoint Blockade, Adoptive Cell Therapies (ACTs), and Vaccine-based Therapies. Each of these therapies directly

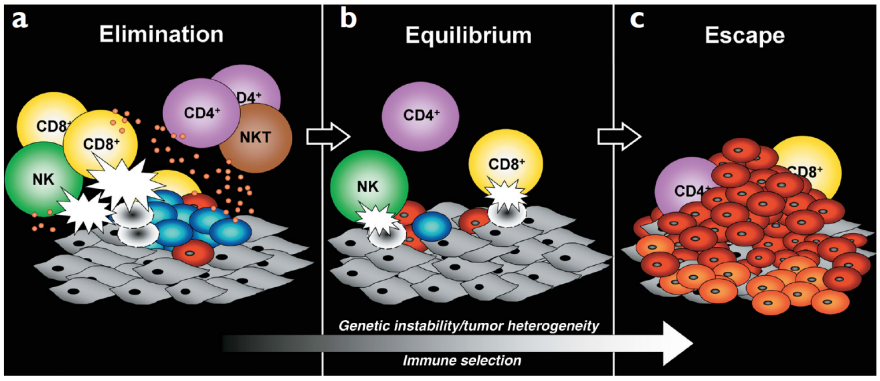


Figure 1.3: The Three E's of Immunoediting; Elimination, Equilibrium, and Escape. Figure obtained from Dunn et.al., Nature Reviews Immunology [29].

or indirectly uses elements of the patient's immune system to selectively attack the tumor.

1.6.1 Antibody-based therapies

Antibody based therapies take advantage of the high specificity of antibodies to their targets, and are used to selectively attack tumor cells that have unique cancer-specific cell-surface proteins. Antibody binding to the cell-surface proteins can lead to cell death via Antibody Dependent Cell-mediated Cytotoxicity (ADCC) or complement. In ADCC, immune cells (primarily of the NK lineage) bind to the Fc regions of antigen bound antibody and destroy the cell harboring the antigen. Complement involves the formation of a membrane attack complex via the alternate pathway of complement activation. Antibodies can also be used to deliver drugs selectively to the tumor cells [103, 116]. Antibody based methods (B-cell methods) are effective only if the tumor has a mutation in a cell-surface protein. The method does not consider the entire landscape of the tumoral DNA.

1.6.2 Checkpoint Blockade

Checkpoint Blockade is also an antibody mediated method, however instead of targeting the tumor cell for destruction, it prevents immune suppression via targeted blockade of specific immune signaling molecules [13, 104]. The immune system has safeguards in place to stop an immune reaction after a threat has been neutralized. These safeguards are often exploited by the tumor to escape the immune response [80, 38, 37, 55]. Targeting key molecules in the immune regulatory pathway such as Programmed Cell Death Receptor Protein 1 (PD-1), Programmed Death-Ligand 1 (PD-L1), and Cytotoxic T-Lymphocyte-Associated Protein 4 (CTLA-4) have been shown to improve antitumor immunity [108, 29, 25, 72, 53, 49, 50]. Targeting immune regulatory checkpoint molecules can however lead to autoimmune reactions [24, 31] and should not be administered over a long period of time. More recent practices include targeting multiple checkpoint molecules simultaneously, and combining checkpoint blockade with Adoptive Cell Therapies for increased effect.

1.6.3 Adoptive cell therapy

Adoptive cell therapies for cancer use autologously derived T-cells to treat a patient. The T-cells can either be cells isolated from the patient and assayed for tumor specificity, or T-cells from the patient that are genetically engineered to produce T-cell receptors (TCRs) specific for the patient's MHC restriction. Genetically modified TCRs can be obtained from transgenic mice or from human cells [92].

In traditional Adoptive T-cell therapy, the TCR repertoire naturally within the patient

is assayed for tumor specificity. Intravital imaging studies showed that T cell migration rapidly arrests when they encounter APCs in the tumor displaying their cognate antigen [26]. These T-cells are known as Tumor Infiltrating Lymphocytes (TILs). Early methods for priming patient T cells used autologous tumor and normal cells to test for reactivity in TILs [79, 92]. This method showed promise in a subgroup of patients diagnosed with melanoma [94]. In a cohort of 93 patients, 20 showed complete regression of the tumor with 3- and 5-year survival rates of 100% and 93% respectively; 32 patients showed partial response and had survival rates of 31% and 21%, while the 42 non-responders had 7% and 5% 3- and 5-year survival rates respectively. 19 of the complete responders showed ongoing complete regression 37 to 82 months after the first treatment. While the complete responders show great potential for the method, a 21.5% success rate is still not an ideal solution.

Not all TILs have the potential to attack tumor cells in vivo. Some TILs are rendered anergic due to a lack of co-stimulation after recognizing their cognate antigen, a fail-safe mechanism used by the immune system to prevent autoimmunity. TILs can also be repressed due to immunomodulatory action of the tumor [12]. Updated TIL based methods involve extracting TILs from the tumor followed by in-vitro stimulation using dendritic cells [113]. Mutations in the tumor are identified using computational methods and short contiguous stretches of DNA encoding the variant-containing regions are synthesized ("Tandem Minigenes") are transfected into dendritic cells. The dendritic cells translate the minigenes and process them to produce the epitopes predicted to be seen on the tumor surfaces. The dendritic cells can be used to stimulate tumor-responsive CTLs, which can be grown in-vitro before being re-infused into the patient. Figure 1.4 describes the adoptive T-cell therapy methodology.

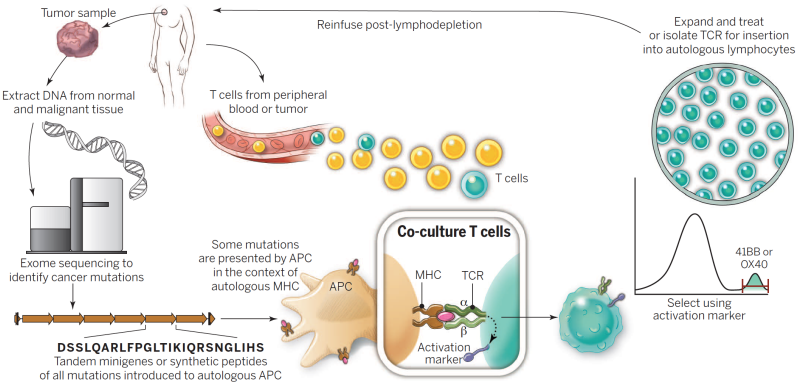


Figure 1.4: Schematic representation of adoptive T-cell therapy. Mutations in coding regions of the tumor DNA can be scanned for potential aberrant protein products and used to activate autologous T-cells. Figure obtained from Restifo et.al., Nature Reviews Immunology [93].

In cases where there is no evidence of the patient having already mounted an immune response to the tumor, engineered T-cells can be used to attack the tumor cells. Two common methods of genetically engineering T-cells are Chimeric Antigen Receptor (CAR) engineering, and TCR transformation. In CAR therapies, autologous T-Cells from the patient are extracted and engineered to express a Chimeric Antigen Receptor, a receptor that is a chimera between the antigen-recognition domains of an antibody and the ζchain of a TCR along with some co-stimulatory domains like CD28 and 4-1BB [32, 61, 52, 96, 88]. The recognition domains are targeted against cell-surface molecules on the tumor and antigen recognition stimulates IL2 production and an immune response. This method is highly work intensive, but has proved to be very effective in its action [70, 9, 84]. For example, CD-19 is a cell-surface marker present only in B-Cell lineages, making it a prime target for treatment of acute lymphoblastic leukemia (ALL), chronic lymphoblastic leukemia (CLL), and some forms of Hodgkin's lymphoma. In

2011, Kalos et.al. showed that anti-CD19 CAR T-cells against patients with advanced leukemia showed tumor reduction and even memory potential despite being engineered [57]. The drawback of CAR-T methods is that CARs can only recognize cell-surface proteins and very few tumors produce unique molecules that are absent on any other normal cell type. TCR transformation therapies involve identification of cells reactive to tumor antigen and transformation of the TCR genes from that cell into patient T-cells using retroviral or lentiviral vectors [54, 92]. The TCRs transformed into the patient need to be extracted from cells having the same MHC restriction as the patient for the therapy to work. TCRs can be obtained from cell lines, cells taken from other patients with the same MHC restriction, or humanized mice engineered to produce human MHCI molecules. Figure 1.5 describes how engineered T-cells can be used in therapy.

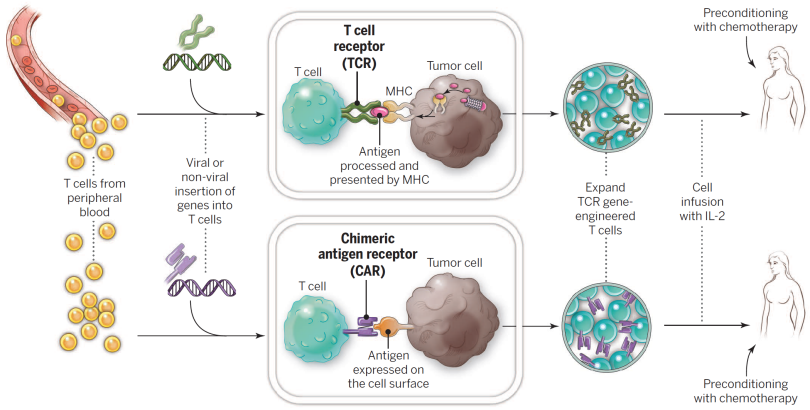


Figure 1.5: Schematic representation of Engineered T-cells used in adoptive T-cell therapy. T-cells can be engineered with Chimeric Antigen Receptors(CARs) or T-cell Receptors grown in-vitro to enhance the immune reaction. Figure obtained from Restifo et.al., Nature Reviews Immunology [93].

1.6.4 Vaccine-based therapies

Vaccine therapies are the most recent generation of cancer immunotherapy. These methods include vaccination of a patient with synthetic peptides [85] or neoepitope-encoding RNA [97] using a suitable vector. Dendritic cells in the skin uptake the vaccine molecules and process them into cell-surface MHCI/II displayed neoepitopes. $CD8^+$ stimulated by these dendritic cells are then capable of selectively targeting the tumor, while $CD4^+$ cells can recruit other cells of the immune system.

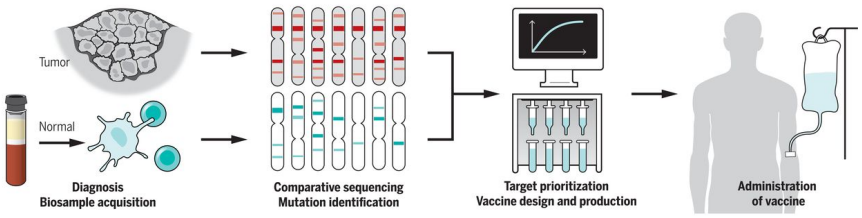


Figure 1.6: Schematic representation of a peptide vaccine workflow for cancer therapy. Figure obtained from Sahin and Türeci, Science [98].

1.7 Thesis statement

Identification of peptides displayed on a tumor's cell surface MHCs, and which ones will elicit a response, is currently a time- and labor intensive process due to lack of good prediction models. Patients diagnosed with cancers - especially cancers with poor prognoses such as glioblastomas - are in a race against time and need fast and accurate treatments. TIL- and Vaccine-based methods are limited by the number of minigenes synthesized and tested before arriving at a

therapeutic target. This thesis describes a workflow to automate the identification of ranked neoepitope targets in a patient that can be used to restrict the search space for a TIL-based or Vaccine-based immune therapy.

Chapter 2

Efficient workflow scheduling using Toil

2.1 Introduction

The genomics industry has seen huge growth in the past decade. The amount of sequencing data available on the internet has increased greatly with the advent of better, high throughput technologies. For example. the Sequence Read Archive (SRA) [67] at the National Center for Biotechnology Information (NCBI) has grown from 20.3 total Gigabases (Gb) in June 2007 to 16.7 Petabases (Pb) in May 2018.

Bioinformatics pipelines are often associated with high I/O tasks due to the large file-sizes of FASTQs, BAMs, etc. Toil caches the results from jobs such that child jobs running on the same node in a distributed cluster can directly use the same file objects, thereby eliminating the need for an intermediary transfer to the job store. Multiple jobs one the same node requesting the same immutable file from the jobstore share the same file object deposited in the cache, reducing the burden on the disk. This resulting drop in I/O allows pipelines to run faster,

and, by the sharing of files, allows users to run more jobs in parallel by reducing overall disk requirements.

My contribution to the Toil project included the implementation of caching to the file store along with numerous unit tests to ensure all the code works perfectly, and the detailed documentation for the manual (submitted as the supplementary for the paper). I made other small contributions to the software during my time working on the project. Toil was published as a proof-of-concept paper in Nature Biotechnology in April 2017 and is included in the next section. I helped write and edit all versions of the manuscript before publication. The relevant section of the supplementary with my contribution has been provided below and Figure 7.1 shows my contribution to the project was substantial from the time I started to the time of publication.

I want to acknowledge the contributions of John Vivian, the lead author of the paper who developed and ran the RNA-Seq analysis used to demonstrate Toil in the paper. I also want to Acknowledge Hannes Schmidt, the computational lead on the software whose brutal code reviews made me an infinitely better programmer.

2.2 The Toil paper

© 2017 Nature America, Inc., part of Springer Nature. All rights reserved.

open sharing of protocols. With a precise ontology to describe standardized protocols, it may be possible to share methods widely and create community standards.

We envisage that in future individual research laboratories, or clusters of co-located laboratories, will have in-house, low-cost automation work cells but will access DNA foundries via the cloud to carry out complex experimental workflows. Technologies enabling this from companies such as Emerald Cloud Lab (S. San Francisco, CA, USA), Synthace (London) and Transcriptic (Menlo Park, CA, USA) could, for example, send experimental designs to foundries and return output data to a researcher. This 'mixed economy' should accelerate the development and sharing of standardized protocols and metrology standards and shift a growing proportion of molecular, cellular and synthetic biology into a fully quantitative and reproducible era.

COMPETING FINANCIAL INTERESTS
The authors declare no competing financial interests.

David W McClymont[1] **& *Paul S Freemont***[1,2]

[1]*The London DNA Foundry, UK Synthetic Biology Innovation, Commercialisation and Industrial Translation Centre, London, UK.* [2]*Centre for Synthetic Biology and Innovation, Department of Medicine, South Kensington Campus, Imperial College London, London, UK.*
e-mail: d.mcclymont@imperial.ac.uk or p.freemont@imperial.ac.uk.

1. Baker, M. *Nature* **533**, 452–454 (2016).
2. Yachie, N. *et al. Nat. Biotechnol.* **35**, 310–312 (2017).
3. Hadimioglu, B., Stearns, R. & Ellson, R. *J. Lab. Autom.* **21**, 4–18 (2016).
4. ANSI SLAS 1–2004: Footprint dimensions; ANSI SLAS 2–2004: Height dimensions; ANSI SLAS 3–2004: Bottom outside flange dimensions; ANSI SLAS 4–2004: Well positions; (ANSI SLAS, 2004).
5. Mckernan, K. & Gustafson, E. in *DNA Sequencing II: Optimizing Preparation and Cleanup* (ed. Kieleczawa, J.) 9.128 (Jones and Bartlett Publishers, 2006).
6. Storch, M. *et al.* BASIC: a new biopart assembly standard for idempotent cloning provides accurate, single-tier DNA assembly for synthetic biology. *ACS Synth. Biol.* **4**, 781–787 (2015).

Toil enables reproducible, open source, big biomedical data analyses

To the Editor:

Contemporary genomic data sets contain tens of thousands of samples and petabytes of sequencing data[1–3]. Pipelines to process genomic data sets often comprise dozens of individual steps, each with their own set of parameters[4,5]. Processing data at this scale and complexity is expensive, can take an unacceptably long time, and requires significant engineering effort. Furthermore, biomedical data sets are often siloed, both for organizational and security considerations and because they are physically difficult to transfer between systems, owing to bandwidth limitations. The solution to better handling these big data problems is twofold: first, we need robust software capable of running analyses quickly and efficiently, and second, we need the software and pipelines to be portable, so that they can be reproduced in any suitable compute environment.

Here, we present Toil, a portable, open-source workflow software that can be used to run scientific workflows on a large scale in cloud or high-performance computing (HPC) environments. Toil was created to include a complete set of features necessary for rapid large-scale analyses across multiple environments. While several other scientific workflow software packages[6–8] offer some subset of fault tolerance, cloud support and HPC support, none offers these with the scale and efficiency to process petabyte and larger-scale data sets efficiently. This sets Toil apart in its capacity to produce results faster and for less cost across diverse environments.

We demonstrate Toil by processing >20,000 RNA-seq samples (**Fig. 1**). The resulting meta-analysis of five data sets is available to readers[9]. The large majority (99%) of these samples were analyzed in under 4 days using a commercial cloud cluster of 32,000 preemptable cores.

To support the sharing of scientific workflows, we designed Toil to execute common workflow language (CWL; **Supplementary Note 1**) and provide draft support for workflow description language (WDL). Both CWL and WDL are standards for scientific workflows[10,11]. A workflow comprises a set of tasks, or 'jobs', that are orchestrated by specification of a set of dependencies that map the inputs and outputs between jobs. In addition to CWL and draft WDL support, Toil provides a Python application program interface (API) that allows workflows to be declared statically, or generated dynamically, so that jobs can define further jobs during execution and therefore as needed (**Supplementary Note 2** and **Supplementary Toil Documentation**). The jobs defined in either CWL or Python can consist of Docker containers, which permit sharing of a program without requiring individual tool installation or configuration within a specific environment. Open-source workflows that use containers can be run regardless of environment. We provide a repository of genomic workflows as examples[12]. Toil supports services, such as databases or servers, that are defined and

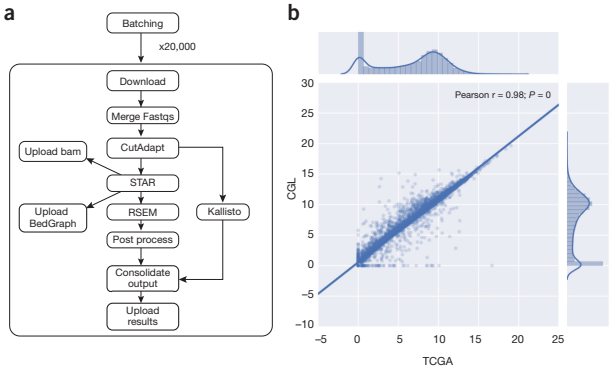


Figure 1 RNA-seq pipeline and expression concordance. (**a**) A dependency graph of the RNA-seq pipeline we developed (named CGL). CutAdapt was used to remove extraneous adapters, STAR was used for alignment and read coverage, and RSEM and Kallisto were used to produce quantification data. (**b**) Scatter plot showing the Pearson correlation between the results of the TCGA best-practices pipeline and the CGL pipeline. 10,000 randomly selected sample and/or gene pairs were subset from the entire TCGA cohort and the normalized counts were plot against each other; this process was repeated five times with no change in Pearson correlation. The unit for counts is: $\log_2$(norm_counts+1).

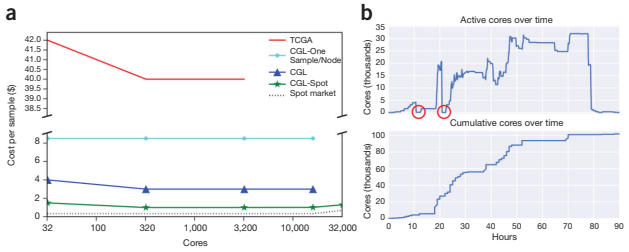


Figure 2 Costs and core usage. (**a**) Scaling tests were run to ascertain the price per sample at varying cluster sizes for the different analysis methods. TCGA (red) shows the cost of running the TCGA best-practices pipeline as re-implemented as a Toil workflow (for comparison). CGL-One-Sample/Node (cyan) shows the cost of running the revised Toil pipeline, one sample per node. CGL (blue) denotes the pipeline running samples across many nodes. CGL-Spot (green) is the same as CGL, but denotes the pipeline run on the Amazon spot market. The slight rise in cost per sample at 32,000 cores was due to a couple of factors: aggressive instance provisioning directly affected the spot price (dotted line), and saving *bam* and *bedGraph* files for each sample. (**b**) Tracking number of cores during the recompute. The two red circles indicate where all worker nodes were terminated and subsequently restarted shortly thereafter.

managed within a workflow. Through this mechanism it integrates with Apache Spark[13] (**Supplementary Fig. 4**), and can be used to rapidly create containerized Spark clusters[14] (**Supplementary Note 3**).

Toil runs in multiple cloud environments including those of Amazon Web Services (AWS; Seattle, WA, USA), Microsoft Azure (Seattle, WA, USA), Google Cloud (Mountain View, CA, USA), OpenStack, and in HPC environments running GridEngine or Slurm and distributed systems running Apache Mesos[15–17] (Forest Hill, MD, USA). Toil can run on a single machine, such as a laptop or workstation, to allow for interactive development, and can be installed with a single command. This portability stems from pluggable backend APIs for machine provisioning, job scheduling and file management (**Supplementary Note 4**). Implementation of these APIs facilitates straightforward extension of Toil to new compute environments. Toil manages intermediate files and checkpointing through a 'job store', which can be an object store like AWS's S3 or a network file-system. The flexibility of the backend APIs allow a single script to be run on any supported compute environment, paired with any job store, without requiring any modifications to the source code.

Toil includes numerous performance optimizations to maximize time and cost efficiencies (**Supplementary Note 5**). Toil implements a leader/worker pattern for job scheduling, in which the leader delegates jobs to workers. To reduce pressure on the leader, workers can decide whether they are capable of running jobs immediately downstream to their assigned task (in terms of resource requirements and workflow dependencies). Frequently, next-generation sequencing workflows are I/O bound, owing to the large volume of data analyzed. To mitigate this, Toil uses file caching and data streaming. Where possible, successive jobs that share files are scheduled on a single node, and caching prevents the need for repeated transfers from the job store. Toil is robust to job failure because workflows can be resumed after any combination of leader and worker failures. This robustness enables workflows to use low-cost machines that can be terminated by the provider at short notice and are currently available at a significant discount on AWS and Google Cloud. We estimate the use of such preemptable machines on AWS lowered the cost of our RNA-seq compute job 2.5-fold, despite encountering over 2,000 premature terminations (**Fig. 2**). Toil also supports fine-grained resource requirements, enabling each job to specify its core, memory and local storage needs for scheduling efficiency.

Controlled-access data requires appropriate precautions to ensure data privacy and protection. Cloud environments offer measures that ensure stringent standards for protected data. Input files can be securely stored on object stores, using encryption, either transparently or with customer managed keys. Compute nodes can be protected by SSH key pairs. When running Toil, all intermediate data transferred to and from the job store can be optionally encrypted during network transmission and on the compute nodes' drives using Toil's cloud-based job store encryption. These and other security measures help ensure protection of the input data, and as part of a broader security plan, can be used to ensure compliance with strict data security requirements.

To demonstrate Toil, we used a single script to compute gene- and isoform-level expression values for 19,952 samples from four studies: The Cancer Genome Atlas (TCGA)[1], Therapeutically Applicable Research To Generate Effective Treatments (TARGET; https://ocg.cancer.gov/programs/target), Pacific Pediatric Neuro-Oncology Consortium (PNOC; http://www.pnoc.us/), and the Genotype Tissue Expression Project (GTEx)[18]. The data set comprised 108 terabytes. The Toil pipeline uses STAR[19] to generate alignments and read coverage graphs, and performs quantification using RSEM[20] and Kallisto[21] (**Fig. 1** and **Supplementary Note 6**). Processing the samples in a single batch on ~32,000 cores on AWS took 90 h of wall time, 368,000 jobs and 1,325,936 core hours. The cost per sample was $1.30, which is an estimated 30-fold reduction in cost, and a similar reduction in time, compared with the TCGA best-practices workflow[5]. We achieved a 98% gene-level concordance with the previous pipeline's expression predictions (**Figs. 1**,**2** and **Supplementary Fig. 1**). Notably, we estimate that the pipeline, without STAR and RSEM, could be used to generate quantifications for $0.19/sample with Kallisto. To illustrate portability, the same pipeline was run on the I-SPY2 data set[22] (156 samples) using a private HPC cluster, achieving similar per sample performance (**Supplementary Table 1**). Expression-level signal graphs (read coverage) of the GTEx data (7,304 samples from 53 tissues, 570 donors) are available from a UCSC Genome Browser[23] public track hub (**Supplementary Fig. 2**). Gene and isoform quantifications for this consistent, union data set are publicly hosted on UCSC Xena[9] and are available for direct access through a public AWS bucket (**Supplementary Fig. 3** and **Supplementary Note 7**).

Although there is an extensive history of open-source workflow-execution software[6–8], the shift to cloud platforms and the advent of standard workflow languages is changing the scale of analyses. Toil is a portable workflow software that supports open community standards for workflow specification and enables researchers to move their computation according to cost, time and data location. For example, in our analysis the sample data were intentionally co-located in the same region as the compute servers in order to provide optimal bandwidth when scaling to thousands of simultaneous jobs (**Supplementary Note 8**). This type of flexibility enables larger, more

© 2017 Nature America, Inc., part of Springer Nature. All rights reserved.

comprehensive analyses. Further, it means that results can be reproduced using the original computation's set of tools and parameters. If we had run the original TCGA best-practices RNA-seq pipeline with one sample per node, it would have cost ~$800,000. Through the use of efficient algorithms (STAR and Kallisto) and Toil, we were able to reduce the final cost to $26,071 (**Supplementary Note 9**).

We have demonstrated the utility of Toil by creating one of the single largest, consistently analyzed, public human RNA-seq expression repositories, which we hope the community will find useful.

Editor's note: This article has been peer-reviewed.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

This work was supported by (BD2K) the National Human Genome Research Institute of the National Institutes of Health award no. 5U54HG007990 and (Cloud Pilot) the National Cancer Institute of the National Institutes of Health under the Broad Institute subaward no. 5417071-5500000716. The UCSC Genome Browser work was supported by the NHGRI award 5U41HG002371 (Corporate Sponsors). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health or our corporate sponsors.

AUTHOR CONTRIBUTIONS

J.V., A.A.R. and B.P wrote the manuscript. J.V., A.A.R., A.N., J.A., C.K., J.N., H.S., P.A., J.P., A.D.D., B.O. and B.P. contributed to Toil development. F.A.N. and A.M. contributed to Toil-Spark integration. J.V. wrote the RNA-seq pipeline and automation software. M.H. and C.B. contributed WDL and cloud support. P.A. and S.Z. contributed CWL support. J.Z., B.C. and M.G. hosted quantification results on UCSC Xena. K.R. hosted GTEx results in UCSC Genome Browser. W.J.K., J.Z., S.Z., G.G., D.A.P., A.D.J., M.C., D.H. and B.P. provided scientific leadership and project oversight.

Data availability. Data are available from this project at the Toil xena hub (https://genome-cancer.soe.ucsc.edu/proj/site/xena/datapages/?host=https://toil.xenahubs.net).

COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details are available in the online version of the paper.

John Vivian[1], Arjun Arkal Rao[1], Frank Austin Nothaft[2,3], Christopher Ketchum[1], Joel Armstrong[1], Adam Novak[1], Jacob Pfeil[1], Jake Narkizian[1], Alden D Deran[1], Audrey Musselman-Brown[1], Hannes Schmidt[1], Peter Amstutz[4], Brian Craft[1], Mary Goldman[1], Kate Rosenbloom[1], Melissa Cline[1], Brian O'Connor[1], Megan Hanna[5], Chet Birger[5], W James Kent[1], David A Patterson[2,3], Anthony D Joseph[2,3], Jingchun Zhu[1], Sasha Zaranek[4], Gad Getz[5], David Haussler[1] & Benedict Paten[1]

[1]Computational Genomics Lab, UC Santa Cruz Genomics Institute, University of California Santa Cruz, Santa Cruz, California, USA. [2]AMP Lab, University of California Berkeley, Berkeley, California, USA. [3]UC Berkeley ASPIRE Lab, Berkeley, California, USA. [4]Curoverse, Somerville, Massachusetts, USA. [5]Broad Institute of Harvard and Massachusetts Institute of Technology (MIT), Cambridge, Massachusetts, USA.
e-mail: benedict@soe.ucsc.edu

1. Weinstein, J.N. *et al. Nat. Genet.* **45**, 1113–1120 (2013).
2. Zhang, J. et al. Database. http://dx.doi.org/10.1093/database/bar026 (2011)
3. Siva, N. *Lancet* **385**, 103–104 (2015).
4. McKenna, A. *et al. Genome Res.* **20**, 1297–1303 (2010).
5. UNC Bioinformatics. TCGA mRNA-seq pipeline for UNC data. https://webshare.bioinf.unc.edu/public/mRNAseq_TCGA/UNC_mRNAseq_summary.pdf (2013).
6. Albrecht, M., Michael, A., Patrick, D., Peter, B. & Douglas, T. in *Proceedings of the 1st ACM SIGMOD Workshop on Scalable Workflow Execution Engines and Technologies* (SWEET '12) 1. ACM (Association of Computing Machinery. http://dx.doi.org/10.1145/2443416.2443417 (2012).
7. Bernhardsson, E. & Frieder, E. Luigi. *Github* https://github.com/spotify/luigi (2016).
8. Goecks, J., Nekrutenko, A. & Taylor, J. *Genome Biol.* **11**, R86 (2010).
9. UCSC. Xena http://xena.ucsc.edu (2016).
10. Amstutz, P. Common workflow language. *Github* https://github.com/common-workflow-language/common-workflow-language (2016).
11. Frazer, S. Workflow description language. *Github* https://github.com/broadinstitute/wdl (2014).
12. Vivian, J. Toil scripts. *Github* https://github.com/BD2KGenomics/toil-scripts/tree/master/src/toil_scripts (2016).
13. Apache Software Foundation. Apache Spark http://spark.apache.org/ (2017).
14. Massie, M. *et al.* ADAM: genomics formats and processing patterns for cloud scale computing. University of California, Berkeley, Technical Report No. UCB/EECS-2013-207 (2013).
15. Gentzsch, W. in *Proceedings First IEEE/ACM International Symposium on Cluster Computing and the Grid* 35–36 http://dx.doi.org/10.1109/ccgrid.2001.923173 (IEEE, 2001).
16. Yoo, A.B., Jette, M.A. & Mark, G. in *Lecture Notes in Computer Science* 44–60 (2003) Springer, Berlin, Heidelberg.
17. Apache Software Foundation. Apache Mesos http://mesos.apache.org/
18. GTEx Consortium. *Science* **348**, 648–660 (2015).
19. Dobin, A. *et al. Bioinformatics* **29**, 15–21 (2013).
20. Li, B. & Dewey, C.N. *BMC Bioinformatics* **12**, 323 (2011).
21. Bray, N.L., Pimentel, H., Melsted, P. & Pachter, L. *Nat. Biotechnol.* **34**, 525–527 (2016).
22. Barker, A.D. *et al. Clin. Pharmacol. Ther.* **86**, 97–100 (2009).
23. Kent, W.J. *et al. Genome Res.* **12**, 996–1006 (2002).

Nextflow enables reproducible computational workflows

To the Editor:

The increasing complexity of readouts for omics analyses goes hand-in-hand with concerns about the reproducibility of experiments that analyze 'big data'[1–3]. When analyzing very large data sets, the main source of computational irreproducibility arises from a lack of good practice pertaining to software and database usage[4–6]. Small variations across computational platforms also contribute to computational irreproducibility by producing numerical instability[7], which is especially relevant to high-performance computational (HPC) environments that are routinely used for omics analyses[8]. We present a solution to this instability named Nextflow, a workflow management system that uses Docker technology for the multi-scale handling of containerized computation.

In silico workflow management systems are an integral part of large-scale biological analyses. These systems enable the rapid prototyping and deployment of pipelines that combine complementary software packages. In genomics the simplest pipelines, such as Kallisto and Sleuth[9], combine an RNA-seq quantification method with a differential expression module (**Supplementary Fig. 1**). Complexity rapidly increases when all aspects of a given analysis are included. For example, the Sanger Companion pipeline[10] bundles 39 independent software tools and libraries into a genome annotation suite. Handling such a large number of software packages, some of which may be incompatible, is a challenge. The conflicting requirements of frequent software updates and maintaining the reproducibility of original results provide another unwelcome wrinkle. Together with these problems, high-throughput usage of complex pipelines can also be burdened by the hundreds of intermediate files often produced by individual tools. Hardware fluctuations in these types of pipelines, combined with poor error handling, could result in considerable readout instability.

Nextflow (http://nextflow.io; **Supplementary Methods**, **Supplementary Note** and **Supplementary Code 1**) is designed to address numerical instability, efficient parallel execution, error tolerance, execution provenance and traceability. It is a domain-specific language that enables rapid pipeline development through the adaptation of existing pipelines written in any scripting language.

We present a qualitative comparison between Nextflow and other similar tools in **Table 1** (ref. 11). We found that multi-scale containerization, which makes it possible to

© 2017 Nature America, Inc., part of Springer Nature. All rights reserved.

2.3 The Toil paper supplementary

The entire supplementary pdf contains details regarding the RNA-Seq pipeline, Spark support, CWL support, etc, and contains the entire Toil manual for version 3.2.0a2 (including documentation for every API call in the package). Since most of this data is irrelevant to this thesis, only the relevant pages are provided below. The full supplementary pdf can be obtained from `https://media.nature.com/original/nature-assets/nbt/journal/v35/n4/extref/nbt.3772-S1.pdf`

2.3.1 Supplementary Toil documentation, chapter 8

CHAPTER

EIGHT

TOIL ARCHITECTURE

The following diagram layouts out the software architecture of Toil.

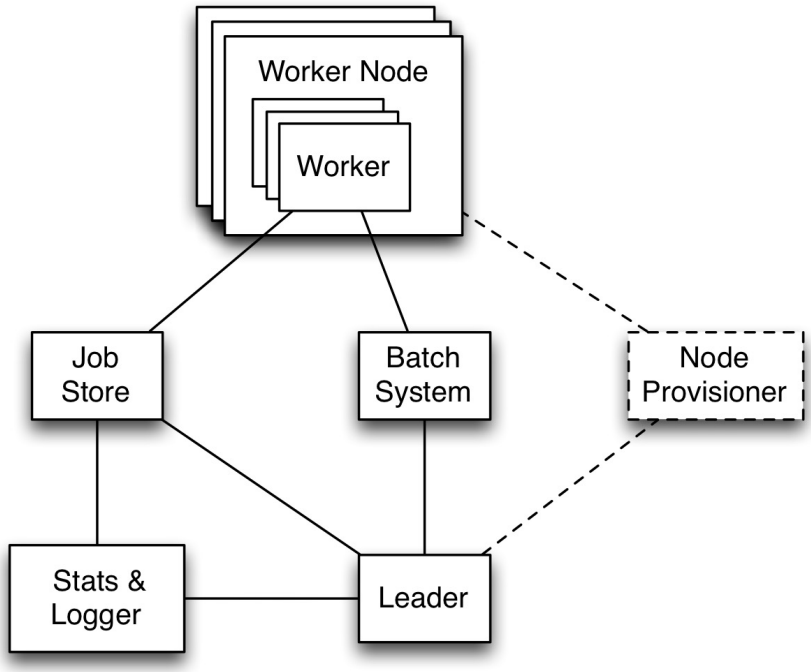


Fig. 8.1: Figure 1: The basic components of the toil architecture. Note the node provisioning is coming soon.

These components are described below:

- **the leader:** The leader is responsible for deciding which jobs should be run. To do this it traverses the job graph. Currently this is a single threaded process, but we make aggressive steps to prevent it becoming

Nature Biotechnology: doi:10.1038/nbt.3772

a bottleneck (see *Read-only Leader* described below).

- **the job-store:** Handles all files shared between the components. Files in the job-store are the means by which the state of the workflow is maintained. Each job is backed by a file in the job store, and atomic updates to this state are used to ensure the workflow can always be resumed upon failure. The job-store can also store all user files, allowing them to be shared between jobs. The job-store is defined by the abstract class `toil.jobStores.AbstractJobStore`. Multiple implementations of this class allow Toil to support different back-end file stores, e.g.: S3, network file systems, Azure file store, etc.
- **workers:** The workers are temporary processes responsible for running jobs, one at a time per worker. Each worker process is invoked with a job argument that it is responsible for running. The worker monitors this job and reports back success or failure to the leader by editing the job's state in the file-store. If the job defines successor jobs the worker may choose to immediately run them (see *Job Chaining* below).
- **the batch-system:** Responsible for scheduling the jobs given to it by the leader, creating a worker command for each job. The batch-system is defined by the abstract class class `toil.batchSystems.AbstractBatchSystem`. Toil uses multiple existing batch systems to schedule jobs, including Apache Mesos, GridEngine and a multi-process single node implementation that allows workflows to be run without any of these frameworks. Toil can therefore fairly easily be made to run a workflow using an existing cluster.
- **the node provisioner:** Creates worker nodes in which the batch system schedules workers. This is currently being developed. It is defined by the abstract class `toil.provisioners.AbstractProvisioner`.
- **the statistics and logging monitor:** Monitors logging and statistics produced by the workers and reports them. Uses the job-store to gather this information.

8.1 Optimizations

Toil implements lots of optimizations designed for scalability. Here we detail some of the key optimizations.

8.1.1 Read-only leader

The leader process is currently implemented as a single thread. Most of the leader's tasks revolve around processing the state of jobs, each stored as a file within the job-store. To minimise the load on this thread, each worker does as much work as possible to manage the state of the job it is running. As a result, with a couple of minor exceptions, the leader process never needs to write or update the state of a job within the job-store. For example, when a job is complete and has no further successors the responsible worker deletes the job from the job-store, marking it complete. The leader then only has to check for the existence of the file when it receives a signal from the batch-system to know that the job is complete. This off-loading of state management is orthogonal to future parallelization of the leader.

8.1.2 Job chaining

The scheduling of successor jobs is partially managed by the worker, reducing the number of individual jobs the leader needs to process. Currently this is very simple: if the there is a single next successor job to run and it's resources fit within the resources of the current job and closely match the resources of the current job then the job is run immediately on the worker without returning to the leader. Further extensions of this strategy are possible, but for many workflows which define a series of serial successors (e.g. map sequencing reads, post-process mapped reads, etc.) this pattern is very effective at reducing leader workload.

Nature Biotechnology: doi:10.1038/nbt.3772

8.1.3 Preemptable node support

Critical to running at large-scale is dealing with intermittent node failures. Toil is therefore designed to always be resumable providing the job-store does not become corrupt. This robustness allows Toil to run on preemptible nodes, which are only available when others are not willing to pay more to use them. Designing workflows that divide into many short individual jobs that can use preemptable nodes allows for workflows to be efficiently scheduled and executed.

8.1.4 Caching

Running bioinformatic pipelines often require the passing of large datasets between jobs. Toil caches the results from jobs such that child jobs running on the same node can directly use the same file objects, thereby eliminating the need for an intermediary transfer to the job store. Caching also reduces the burden on the local disks, because multiple jobs can share a single file. The resulting drop in I/O allows pipelines to run faster, and, by the sharing of files, allows users to run more jobs in parallel by reducing overall disk requirements.

To demonstrate the efficiency of caching, we ran an experimental internal pipeline on 3 samples from the TCGA Lung Squamous Carcinoma (LUSC) dataset. The pipeline takes the tumor and normal exome fastqs, and the tumor rna fastq and input, and predicts MHC presented neoepitopes in the patient that are potential targets for T-cell based immunotherapies. The pipeline was run individually on the samples on c3.8xlarge machines on AWS (60GB RAM,600GB SSD storage, 32 cores). The pipeline aligns the data to hg19-based references, predicts MHC haplotypes using PHLAT, calls mutations using 2 callers (MuTect and RADIA) and annotates them using SnpEff, then predicts MHC:peptide binding using the IEDB suite of tools before running an in-house rank boosting algorithm on the final calls.

To optimize time taken, The pipeline is written such that mutations are called on a per-chromosome basis from the whole-exome bams and are merged into a complete vcf. Running mutect in parallel on whole exome bams requires each mutect job to download the complete Tumor and Normal Bams to their working directories – An operation that quickly fills the disk and limits the parallelizability of jobs. The script was run in Toil, with and without caching, and Figure 2 shows that the workflow finishes faster in the cached case while using less disk on average than the uncached run. We believe that benefits of caching arising from file transfers will be much higher on magnetic disk-based storage systems as compared to the SSD systems we tested this on.

Nature Biotechnology: doi:10.1038/nbt.3772

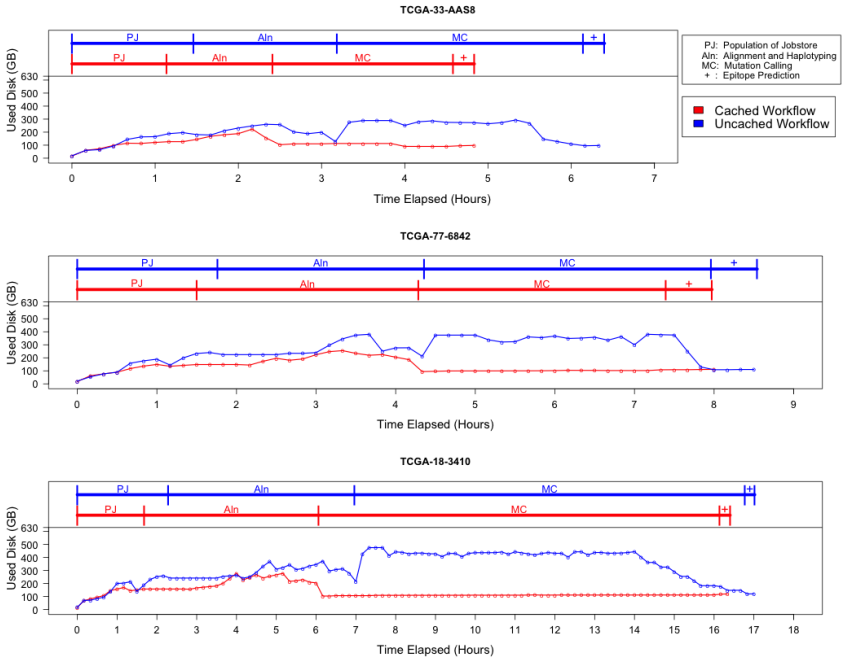


Fig. 8.2: Figure 2: Efficiency gain from caching. The lower half of each plot describes the disk used by the pipeline recorded every 10 minutes over the duration of the pipeline, and the upper half shows the corresponding stage of the pipeline that is being processed. Since jobs requesting the same file shared the same inode, the effective load on the disk is considerably lower than in the uncached case where every job downloads a personal copy of every file it needs. We see that in all cases, the uncached run uses almost 300-400GB more that the uncached run in the resource heavy mutation calling step. We also see a benefit in terms of wall time for each stage since we eliminate the time taken for file transfers.

Nature Biotechnology: doi:10.1038/nbt.3772

Chapter 3

Translation of genomic mutations into therapeutically assayable peptides

3.1 Introduction

In order to predict tumor neoepitopes of clinical significance, we need a reliable tool to translate the mutations from genomic space into peptides that can be assayed with existing peptide:MHC (pMHC) binding affinity predictors. Variant annotators like SNPEff [19], VEP [75] and TransVar [122] can annotate individual single nucleotide variants(SNVs), and short insertions and deletions (INDELs), providing details on the exact change in the amino acid sequences. Fusion genes (Fusions) are identified from the genomic breakpoints and are usually annotated by the Fusion caller. To predict neoepitopes of length n using an existing pMHC binding affinity predictor, we need at least $n-1$ amino acids flanking the variant. Since no freely available tool exists to predict mutated peptides, I developed TransGene.

TransGene is a standalone python tool capable of processing any combination of SNVs, INDELs, and Fusion Genes to variant-containing predict peptides from the tumor proteome. TransGene can use provided RNA-Seq data to confidently predict only peptides arising from transcribed variants, and DNA-Seq to filter reads arising from Oxidation of Guanosine (OxoG) [23] events. TransGene can also phase and merge events within a short distance of each other if they can potentially fit on the same neoepitope. The variant containing peptides produced by TransGene are called ImmunoActive Regions (IARs) as they can potentially stimulate the immune system.

With the exception of a few auxiliary modules used in fusion gene handling, I developed every module in the TransGene package. This includes the modules to process calls from every version of SNPEff, process VEP calls, use RNA-Seq to filter variants with low allele frequency, use DNA-Seq to filter OxoG events, and the modules to predict IARs from SNVs, INDELs and Fusions. The first pass at handling fusion genes was written by Jacob Pfeil under my guidance. Since then, I have added code to filter artifacts arising from the fusion gene prediction process and completely rewrote the way IARs are predicted from breakpoints.

I ran TransGene on the input dataset, analyzed and documented the results, and wrote the entire paper with edits from Dr. Sofie Salama and Dr. David Haussler. The text and supplementary for the proposed paper are provided below.

3.2 The TransGene paper

3.2.1 Abstract

Somatic mutations in cancer cells can produce peptides recognized by autologous T-cells as non-self, resulting in an immune reaction to the tumor. We developed TransGene to translate somatic mutants identified by DNA sequence analysis into peptides that can be used to identify therapeutic targets in cancer patients. We demonstrate TransGene on 240 samples from The Cancer Genome Atlas (TCGA) Prostate Adenocarcinoma (PRAD) cohort. More information on the mutational rate of this dataset can be seen in Chapter 4. TransGene is developed entirely in python 2.7 and is available at `www.github.com/arkal/transgene`. TransGene can be installed using pip, and dockerized versions are available from aarjunrao/transgene.

3.2.2 Introduction

Genomic mutations acquired by cancer cells can result in the production of proteins with amino acid sequence alterations. Studies have shown that these alterations can be recognized by autologous T-cells as non-self, resulting in an immune response [21, 114, 74]. These analyses currently use somatic mutations called using next generation sequencing of the tumor and matched normal DNA, and often use sequencing of the RNA to identify only expressed variants. Published workflows for neoepitope prediction are specific in their input requirements. For example, pVAC-Seq [51] require VEP annotated VCFs (with custom downstream plugins), and INTEGRATE-Neo [121] use fusion gene (Fusion) calls from INTEGRATE. We present

TransGene, a standalone tool to translate Fusions, SNPEff- [20] or VEP- [75] annotated Single Nucleotide Variants (SNVs), and short insertions and deletions (INDELs) into peptides that can be used with any pMHC binding predictor. TransGene enables researchers with more control over their neoepitope prediction workflows.

3.2.3 The TransGene workflow

TransGene accepts any combination of SNVs, INDELs, and Fusions, and generates the set of all mutant peptide sequences of a given length derived from those variants. TransGene operates in 3 steps (3.1A), 1) parse input mutations, 2) process SNVs and INDELs 3) process Fusions. Identification of expressed variants and potential oxidation of guanine to 8-oxoguanine (OxoG) events [23] occurs in step 1 if the user provides an input RNA BAM or Whole Genome/Exome (WGS/WXS) BAM (sequencing data) respectively.

In step 1, TransGene accepts SNVs and INDELs in a single VCF annotated using SNPEff or VEP. For each VCF record the following sequence ontology terms are accepted: *missense_variant*, *synonymous_variant* and *stop_gained* for SNVs, *frameshift_variant*, *inframe_insertion*, *inframe_deletion*, *protein_altering_variant*, *disruptive_inframe_insertion*, and *disruptive_inframe_deletion* for INDELs. If an RNA BAM is provided, the variant allele frequency (VAF) of the mutant in the expressed transcript is calculated and values below a user-defined threshold are rejected. If a WGS/WXS BAM is provided, TransGene will identify OxoG events. We found OxoG filtering using TransGene to be critical for analyzing pediatric neuroblastoma tumor sequencing data from the TARGET [87] cohort [111], for example. If sequencing data is provided, TransGene will output a processed VCF file with read coverage,

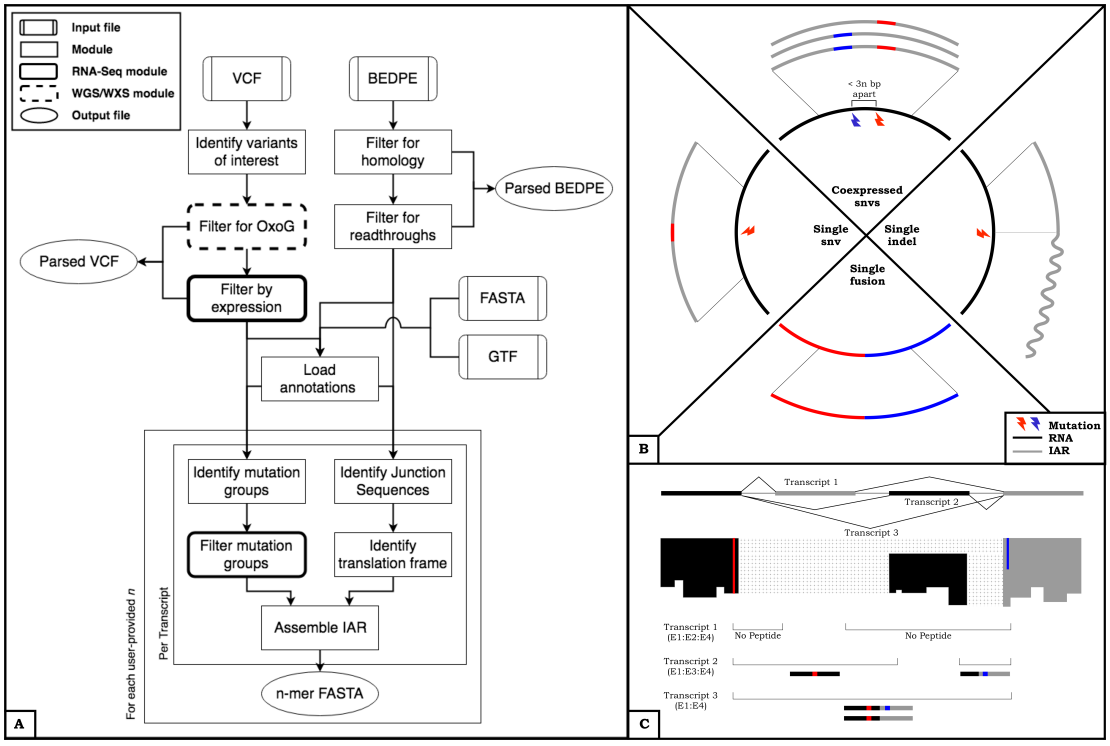


Figure 3.1: A) The TransGene workflow. B) A cartoon describing the output ImmunoActive Regions (IARs) for 4 mutation events. C) A cartoon describing how transgene handles mutations near exon boundaries. It also describes how transgene handles co-expressed mutations when RNA-Seq data is present. Transcript 1 produces no IARs since there are no reads supporting junctions E1:E2 and E2:E4, Transcript 2 produces 2 IARs from junctions E1:E3 and E3:E4, and Transcript 3 produces 2 IARs from junction E1:E4 (one with both mutations and one with only the mutant from E1 based on the read support/VAF)

VAF and a reason for all rejected calls.

Fusion calls are accepted in the BEDPE format with the Fusion name in EnsEMBL terms in the "name" column, and four additional columns "JunctionSeq1", "JunctionSeq2", "HUGO1" and "HUGO2" containing the inferred junction sequences on either side of the Fusion if available (can be blank) and the HUGO names for each gene. Fusions are optionally filtered to remove potential false positive events including Fusions between mitochondrial, immunoglobulin, or lncRNA genes. Transcriptional read-throughs below a user-defined distance threshold can be filtered. Once all mutations are read to memory, transgene parses a provided gencode annotation and genomic FASTA to selectively load transcript sequences and annotations from Fusion- and INDEL-containing genes into memory.

In step 2, TransGene handles the SNVs and INDELs on a per-transcript basis and prints ImmunoActive Regions (IARs) to an output file. An IAR is the region comprising a mutant amino acid along with the $n-1$ flanking amino acids on either side such that every n-mer from the region contains the mutant (3.1B). Multiple mutations can be chained together if the distance between two consecutive events is less than n residues. The IAR for a co-expressed group is the $n-1$ bases on either side of the extreme mutations along with all wild-type and mutant amino acids that lie between them. If an RNA-Seq BAM is provided, TransGene will phase co-expressed mutations and filter the set of all combinations into those with expression evidence (3.1C). Frameshift mutants are assumed to disrupt splicing, and are processed for a user-provided number of residues downstream to the mutant (30 amino acids by default) in search of a stop codon. Every generated IAR has a corresponding wildtype peptide that can be used to calculate the Agretopicity Index [28].

In step 3, TransGene handles Fusion calls from the BEDPE file, analyzing every record in the file to identify combinations of transcripts for both Fusion partners. TransGene infers junction sequences for each combination or use user-provided sequences if provided in the input BEDPE. If the Fusion results in an in-frame change, the junction sequences are concatenated and translated to obtain the IAR. In a frameshift event, the acceptor sequence is the sequence downstream to the breakpoint on the 3' partner.

The output from transgene includes a peptide FASTA containing IARs for each peptide length *n* specified by the user, a corresponding FASTA of wildtype peptides, a processed VCF and a processed BEDPE. The sequence name for each IAR contains the information about the gene, transcript, mutation and VAF.

3.2.4 Application of TransGene to a test cohort

We processed 240 samples from the TCGA Prostate Adenocarcinoma (PRAD) cohort [81]. Transgene was run in an automated manner with all Fusion filters, and default values for all other parameters to produce 9-, 10- and 15-mer-containing IARs. MHCI epitopes are most commonly 9 or 10 amino acids long, and MHCII peptides can be larger than 20 amino acids but netMHC uses 15 as the size for predicting the binding core. The entire workflow was orchestrated with the Toil workflow manager [115]. The RNA-Seq BAM and VCF (MuTect2 + VEP) for each selected sample from the TCGA PRAD cohort were downloaded on the fly from the National Center for Biotechnology Information (NCBI) Genomics Data Commons(GDC). The Mutect2 VCFs were parsed to retain only calls passing all filters. Fusion data for each sample was obtained from the supplementary information in the INTEGRATE-Neo paper [121] and parsed

into a BEDPE file for each sample using a python script. A full explanation (including all scripts used in this analysis) is provided at `https://www.github.com/arkal/TransGene_supplementary`.

The dataset contained 15,430 SNVs and INDELs (0-835/sample), and 2,040 Fusions (1-47/sample). The median number of SNVs, INDELs and fusions per sample were 50, 5, and 8 respectively. On average, TransGene accepted 30% of the SNVs and INDELs, based on expression and 88% of fusions based on all filters. A total of 4,958 9-mer-, 4,998 10-mer- and 4,742 15-mer IARs were generated. The median number of 9-mer IARs from SNVs, INDELs and Fusions per samples was 9, 0.5, and 5.5 respectively, similar to the values for 10-mers and 15-mers (Supplementary Table 1). Seven samples had co-expressed mutants affecting the same IAR, of which five affected the same codon (Supplementary Table 2, Supplementary Figures 3.2 and 3.3).

The analysis was run on a Microsoft Azure D14_v2 VM with 112 GB of RAM and 16 2.4GHz Intel Xeon®E5-2673 v3 (Haswell) processors. Each sample took on average 435 seconds to run. Since TransGene is a novel tool, we do not have another tool to benchmark this against. However, we can say that 7 minutes of runtime is relatively small compared to alignment and variant calling, making this a short, yet effective node in a neoepitope prediction workflow.

3.2.5 Discussion

TransGene is a standalone tool capable of handling patient mutation data with a combination of sequencing data for detailed putative neoepitope identification. The IARs produced

by TransGene can be supplied as input to any pMHC binding prediction tool to identify therapeutic neoepitopes for a cancer patient. We believe that TransGene will allow researchers more freedom in choosing and prioritizing neoepitopes for Adoptive Cell-, and Vaccine-based immunotherapies. TransGene works with variant data off the internet (as demonstrated by this paper), making it a good resource for identification of neoepitope trends across cancer types. TransGene is an integral part of our in-house workflow for predicting neoepitopes, ProTECT (Chapter 4).

3.3 The TransGene paper supplementary

3.3.1 Supplementary Figures

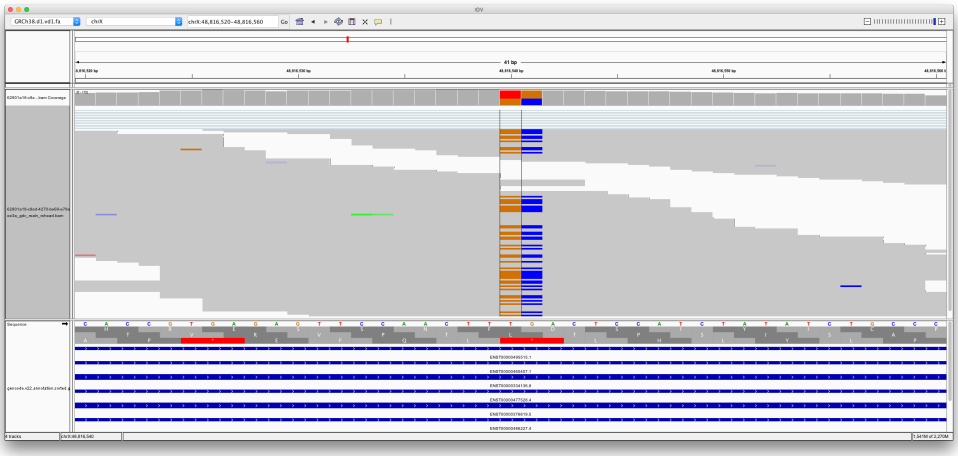


Figure 3.2: An IGV screenshot of chromosome X for sample TCGA-CH-5792. The mutations at positions 48816540 (T>G) and 48816541 (G>C) affect different codons and are fully phased, hence they affect the same IAR at consecutive residues.

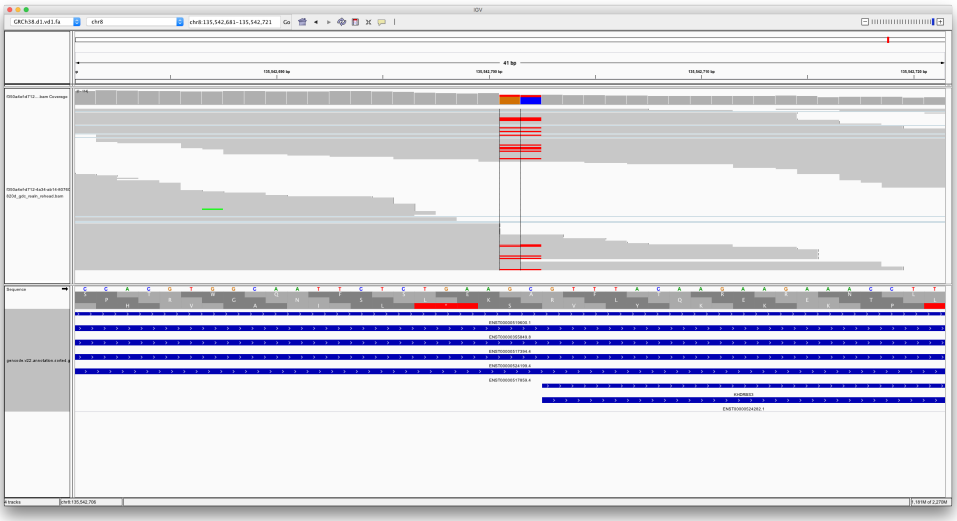


Figure 3.3: An IGV screenshot of chromosome 8 for sample TCGA-EJ-5525. The mutations at positions 135542701 (G>T) and 135542702 (C>T) affect different codons and are fully phased, hence they affect the same IAR at consecutive residues.

3.3.2 Supplementary Tables

All supplementary tables are provided as separate sheets in a Microsoft excel document and are named chapter3-supplementary_tableX where X is the number of the table.

- **Supplementary Table 1:** Information on the number of mutants called and accepted for each input sample, along with the number of 9-mer, 10-mer and 15-mer IARs predicted.
- **Supplementary Table 2:** Co-Expressed variants detected in the cohort. A) Co-expressed variants within 9 amino acids of each other, affecting the same 10-mer IAR. B) Co-expressed variants within 3 amino acids of each other, affecting the same codon.

Chapter 4

ProTECT: Prediction of T-cell Epitopes for Cancer Therapy

4.1 Introduction

This chapter describes the ProTECT pipeline. ProTECT, an acronym for **Pr**ediction **o**f **T**-cell **E**pitopes for **C**ancer **T**herapy, is a computational workflow used to identify short tumor-produced peptides of immunotherapeutic significance in cancer patients in a scalable, automated, reproducible, and efficient manner. The development of ProTECT was driven by the lack of tools in the field to predict cancer neoepitopes. Since I started my work on ProTECT, two tools, pVAC-Seq [51] and INTEGRATE-Neo [121] were published. pVAC-Seq generates a filtered list of neoepitope candidates from SNVs and INDELs, and INTEGRATE-Neo identified a list of neoepitopes from Fusion genes. Neither of these tools are automated or process data from FASTQs.

I identified every required tool (with ideal input parameters) for every module used in the pipeline, and developed the remaining required modules that had no off-the-shelf tool – TransGene (chapter 3) to translate genomic mutations into peptides, and Rankboost to rank the plethora of peptide:MHC (pMHC) predictions made for a sample. I containerized each module into individual Docker images tagged with the version such that the pipeline would use reproducible modules. I wrote the entire pipeline in the Toil framework, in the form of individual packages for each major step in the pipeline. This allows users to easily use my code to make new custom pipelines. I wrote all the auxiliary code required to process outputs from modules into the input for others allowing ProTECT to run with no manual intervention. I made the pipeline available on the Python Package Index (PyPI) for greater availability.

In addition to all of the above, I mentored/oversaw the contribution of the following individuals. Jacob Pfeil (Lab roton/early addition to the lab at the time) developed the first pass at Fusion support for ProTECT. Christine Cho, Danlin Lillemark, and Shreya Guha (summer high school interns through the UCSC SIP program in Summer 2016) curated a list of immune pathways implicated with tumor progression or regression. Ada Madejska (undergraduate research assistant) completed the work started by the high schoolers and converted it into a reporting module in ProTECT. Ada also identified clinical trials involving CAR therapies and made that an additional report output by ProTECT. Aisling O'Farrell (undergraduate research assistant) created a working dockerized version of ProTECT with additional help from Christopher Ketchum from the Genomics Institute. David Stewart (undergraduate research assistant) implemented support for mixing and matching mutation callers and specifying thresholds for consensus calling providing users better control over the variants called by ProTECT.

4.2 The ProTECT paper

4.2.1 Abstract

Somatic mutations in cancers affecting protein coding genes can give rise to potentially therapeutic neoepitopes. These neoepitopes can guide Adoptive Cell therapies (ACTs) and Peptide Vaccines (PVs) to selectively target tumor cells in the body using autologous patient cytotoxic T-cells. Currently, researchers have to independently align their data, call somatic mutations and haplotype the patient's HLA to use existing neoepitope prediction tools. We present ProTECT, a fully automated, reproducible, scalable, and efficient end-to-end analysis pipeline to identify and rank therapeutically relevant tumor neoepitopes in terms of immunogenicity starting directly from patient sequencing data. The ProTECT pipeline encompasses alignment, HLA haplotyping, mutation calling (single nucleotide variants, short insertions and deletions, and gene fusions), pMHC binding prediction, and ranking of final candidates. ProTECT also allows the user to start from pre-computed data. We demonstrate ProTECT on 326 samples from the TCGA Prostate Adenocarcinoma cohort, and compare it with published tools. ProTECT processes an average TCGA sample starting from input FASTQs in 26 minutes when run in a batch of 50 samples on our cluster. ProTECT is available at `https://www.github.com/BD2KGenomics/protect`.

4.2.2 Introduction

Tumor recognition by the adaptive immune system has been known as early as 1988, when Muul et.al. described a 55% objective response rate in a cohort of melanoma samples [79]. However, at the time, T-cell responses were observed to be short lived, often lasting very few days. Later studies showed that tumors are adept at suppressing the immune system in many ways [34, 105, 36, 39].

Checkpoint blockade therapy has seen a great increase in interest in the past few years with numerous drugs being approved by the FDA for clinical treatment [91, 15, 50]. Prevention of PD-1:PD-L1 [5] and CTLA-4:B7.1/2 [17] binding via monoclonal antibodies re-enables the immune attack against the tumor, however it can leave the patient open to development of autoimmunity or other toxicities associated with unchecked immune action [95, 31].

Adoptive Cell Therapies (ACTs) use T-cells specifically targeted against the tumor to reduce the collateral damage associated with conventional therapies. Tumor Infiltrating Lymphocytes (TILs) from patient tumors can be activated and expanded in-vitro using minced autologous tumor or experimentally primed autologous dendritic cells [94, 119] to selectively target cell-surface MHC-presented antigen produced by the tumor. Cancer vaccines attempt to produce the same result as ACT therapies by stimulating dendritic cells in-vivo via synthetically produced peptides (or coding RNA) delivered subcutaneously to the patient [85, 97]. This removes the requirement for extracting and assaying T-Cell interactions in-vitro. Experimentally primed Dendritic cells and Peptide vaccine therapies require prior knowledge of the mutations in the tumor to identify the potentially targetable sequence. This can be achieved using a com-

bination of Next Generation Sequencing (NGS) and bioinformatics.

Currently, there are two published methods to identify neoepitopes from tumor variants. pVAC-Seq [51] is an automated bioinformatics pipeline that identifies neoepitopes generated from a pre-computed VCF file of variants annotated using VEP [75] (and some specialized plug-ins). INTEGRATE-Neo [121] identifies neoepitopes from fusion genes provided in a pre-computed BEDPE file. Both these tools require a user to previously align the sequencing data to a reference of choice and call variants before following the same logical paradigm of identifying mutant peptides and predicting pMHC binding affinity (often via netMHC [71]). Other unpublished (or proprietary) pipelines exist [2], that differ in their degree of automation, input mutation type and annotation, and presence or absence of a ranking schema, however to our knowledge, none of them fully automate the process from end-to end, beginning at the raw FASTQs emitted by the sequencer.

We describe ProTECT, a fully automated, open-source tool for the Prediction of T-cell Epitopes for Cancer Therapy. ProTECT accepts an input trio of sequencing data from a patient consisting of the tumor DNA, normal DNA, and the tumor RNA sequencing reads in the FASTQ format and processes the data from end-to-end including Alignment, in-silico HLA haplotyping, expression profiling, mutation calling and neoepitope prediction. We showed the utility of ProTECT in a previous publication where we identified a potentially therapeutic neoepitope from the hotspot ALK:R1275Q mutant in pediatric Neuroblastoma [111] (Chapter 5). The prediction was biochemically validated and was also shown to be recognized by PBMCs from HLA-matched donors using a tetramer assay.

To demonstrate ProTECT in this manuscript, we use the Prostate Adenocarcinoma

(PRAD) cohort from The Cancer Genome Atlas (TCGA) project [81]. The PRAD cohort consists of 500 patients in total, 331 of which have trios of genomic data (Tumor DNA, Normal DNA, and Tumor RNA). The TCGA PRAD cohort has an average of 24.8 exonic mutations per sample [66] and 157 of the 500 are predicted to contain fusion transcript [118]. In addition to the PRAD analysis, we compare our results to the published pVAC-Seq and INTEGRATE-Neo pipelines.

4.2.3 The ProTECT pipeline

4.2.3.1 Description

ProTECT identifies therapeutically significant neoepitopes in a given samples using a series of methods that are orchestrated in a single automated pipeline using the Toil workflow manager [115]. Input sequencing data is processed to identify the patient's HLA haplotype, and to generate alignments against a reference genome (sequence alignment). The aligned reads are used to identify somatic variants (mutation calling), and to estimate the gene- and isoform-level expression within the tumor (expression profiling). Somatic variants are translated into short peptides that contain a new sequence (or neoepitope) that isn't seen in the self-immunopeptidome. The binding affinity for these peptides to MHC alleles in the patient's haplotype is assayed in-silico and the final output is a ranked list of sequences that can be used to guide a neoepitope-based immune therapy. The schematic for the ProTECT workflow is shown in figure4.1. For convenience, ProTECT can also be run using pre-computed alignments, or using pre-computed somatic variants.

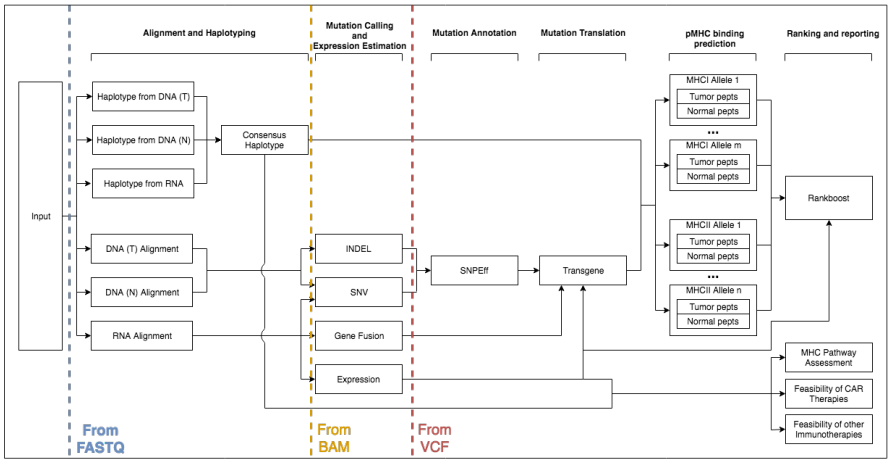


Figure 4.1: A schematic description of the ProTECT workflow. ProTECT can process FASTQs all the way through the prediction of ImmunoActive Regions, including alignment, HLA Haplotyping, variant calling, expression estimation, mutation translation, and pMHC binding affinity prediction. ProTECT also allows users to provide pre-computed inputs for various steps instead.

4.2.3.1.1 Sequence Alignment

Variant calling on sequencing data requires the input FASTQs to be aligned to a genomic reference. DNA sequence alignment on the Tumor and Normal sequencing data is carried out using the burrows wheeler aligner (BWA) [69] with default parameters. The aligned SAM file is processed to properly format the header, and is then converted to a coordinate-sorted BAM file with a corresponding index. RNA sequence alignment is carried out using the ultra-fast aligner, STAR [27]. The parameters for the run are optimized for fusion detection (Supplementary note 1). ProTECT creates both a genome-mapped BAM (Used in variant calling) and a transcriptome-mapped one (Used in expression estimation).

Alternatively, ProTECT accepts pre-aligned BAMs as an input if the MHC haplotype

is provided as well. ProTECT assumes that the user has aligned the DNA and RNA using the same reference genome.

4.2.3.1.2 Haplotyping

The HLA haplotype of the patient is crucial information when predicting neoepitopes as TCRs will only react favorably so self-MHC -displayed neoepitopes. The HLA haplotype is predicted using PHLAT [6], taking each input source of sequencing data. A consensus haplotype is generated based on agreement between the three individual haplotype predictions. Due to limitations in the tool, we only predict HLA-A, HLA-B and HLA-C for MHCI, and HLA-DPA/B and HLA-DRB for MHCII.

4.2.3.1.3 Expression profiling

The gene- and isoform-level expression profiles for the patient are used to identify which variant-containing genes and transcripts are expressed in good quantities (relative to other expressed molecules). This is estimated using RSEM [68] with default parameters.

4.2.3.1.4 Variant Calling

SNV calling occurs in a per-chromosome manner across any combination of 5 mutation prediction algorithms, MuTECT [18], MuSE [35], RADIA [90], Somatic Sniper [65], and Strelka [102]. The choice of mutation callers was guided by the ICGC DREAM mutation calling challenge [33]. The SNV predictions from the callers are merged into a common file and only events supported by n or more predictors (user-provided, or $(x+1)/2$ where x is the number of input

predictors) advance to the translation step.

In addition to SNV calling, Strelka also predicts somatic INDELs. These INDELs are added to the same VCF that contains the consensus-called SNVs.

Fusion genes are predicted using STAR-Fusion [40] with default parameters. The fusions are then annotated using Fusion-Inspector [3] along with an optional assembly step using Trinity [41].

4.2.3.1.5 Mutation Translation

Variants called in the previous step need to be translated into peptides that can be assayed in-silico for MHC binding affinity. MHCI molecules generally bind to 9- and 10-mer peptides and the binding groove of an MHCII molecule is considered to be about 15 amino acids (This value is used by our pMHC binding prediction predictor).

SNVs are translated from genomic coordinates to proteomic coordinates using SNPEff [19] and coding variants are processed using TransGene (Chapter 3). TransGene filters the input SNPEff-ed VCF to exclude non-expressed calls based on the gene expression data. Mutations lying within 27, 30, and 45 bp of each other (for 9-mer-, 10-mer- and 15-mer-containing peptides respectively) are chained together into an "ImmunoActive region" (IAR), or a region that will potentially produce an immunogenic peptide. These events are phased using the RNA-Seq to ensure that they truly are co-expressed on the same chromosome.

Fusion peptides are generated using the breakpoints present in the input BEDPE file. TransGene uses provided junction sequences or infers them from the input annotation file. The predicted IAR contains $(n-1)\times 3 \;\; \forall n \in \{9,10,15\}$ bp on either side of the fusion junction.

TransGene outputs 10 files in total, A filtered VCF, potential 9-, 10- and 15-mer neoepitopes as peptide FASTQs, the corresponding 9-, 10- and 15-mer wild type epitopes, and a map file for each of the tumor FASTQs that maps each peptide FASTQ name to the gene, mutation, and list of transcript isoforms of origin.

4.2.3.1.6 MHC:Peptide binding prediction

The predicted neoepitopes are assayed against each MHCI (9- and 10-mers) and MHCII (15-mers) predicted to be in the patient's HLA haplotype using the IEDB MHCI and MHCII binding predictions tools.

The IEDB tools run a panel of methods [83, 82, 58, 14, 86, 120, 110] on each input query (peptide FASTQ + MHC allele) and provide a consensus "percentile rank" that describes on average, how well did each peptide predict to bind against a background set of 100,000 UniPROT derived peptides. pMHC calls predicted to bind within the top 5% of all binders for that MHC are selected for further study.

The normal counterpart for each selected neoepitope is then assayed against the MHC(s) identified to bind with the neoepitope. The data is then tabulated into one table each for MHCI and MHCII.

4.2.3.1.7 Neo-epitope ranking

Neoepitope:MHC calls are consolidated by the candidate IAR of origin. IARs are ranked using a boosting strategy that rewards candidates satisfying certain biologically relevant criteria including the number of calls originating from the IAR (diversity), the number of MHCs stimulated by

peptides from the region (promiscuity), the combined expression of the isoforms displaying the neoepitope-generating mutation, the number of neoepitopes in the region predicted to bind to an MHC better than their wild type counterpart, and the number of events where a 10-mer, and a 9-mer subset of it both bind well to the MHC (only in peptide:MHCI ranking). The ranked results are tabulated into a single file that can be used to guide a neoepitope based therapy. An output table is generated for MHCI using the 9- and 10-mer pMHC calls, and for MHCII using the 15-mer calls. The table contains information for each potential IAR including the amino acid sequence, numeric values for each of the ranking criteria, the final rank, and the list of binding MHCs.

4.2.3.2 Reproducibility

Every tool used the pipeline, from established aligners to the in-house script used to translate mutations, is wrapped in a Docker image [89]. Docker allows developers to wrap software along with all requirements into an image, that can be instantiated into a container on any other machine. The software within the container is guaranteed to run in the same manner on any machine, under the same environmental constraints. Since every tool in ProTECT is containerized using Docker, ProTECT results from different machines are directly comparable to each other.

4.2.3.3 Automation, Scalability and Efficiency

ProTECT is built to be run end-to-end without any user intervention. ProTECT is written in the Toil framework and will attempt to run the pipeline on the given input samples in a resource-

efficient manner. The pluggable backend Toil APIs allow ProTECT to run on a single machine, a grid engine cluster, or a mesos cluster setup on a local network, or on AWS. Toil allows users to deploy scripts on Azure and the Google cloud as well, however ProTECT does not support these platforms. Users provide ProTECT a configuration file that details the input files, and the various indexes and versions of tools to use during the run. ProTECT downloads the files to a "file store" and then spawns a graph of jobs for each input sample culminating in a ranked list of epitopes. The nodes in the graph are tuned to request an appropriate number of CPUs (for multithreaded jobs), memory and disk space – And Toil ensures that these jobs are parallelized to the maximum extent.

4.2.4 Materials

4.2.4.1 Input samples

Genomic Trio (Tumor DNA, Normal DNA, and Tumor RNA) BAMs from 326 samples in the Cancer Genome Atlas (TCGA) Prostate Adenocarcinoma (PRAD) cohort were downloaded from the Genomics Data Commons (GDC) at the National Cancer Institute using the GDC data transfer tool. The downloaded BAM files were converted back to FASTQ using the SamToFastq module from Picard version 1.125 [1]. MHCI haplotype calls using POLYSOLVER [106] for all samples for benchmarking were obtained and used with the permission of Dr. Catherine Wu.

Genomic Trios from 3 additional samples (Mel-21, Mel-38, Mel-218) were downloaded from the SRA [67] via Bioproject PRJNA278450/dbGaP accession phs001005. These patients were diagnosed with stage III resected cutaneous melanoma and had all previously received ipilimumab. Data from seven HLA-A*02:01 restricted vaccines tested for each patient

were obtained from the supplementary information of the original manuscript [16].

The input data for the INTEGRATE-Neo comparison included haplotype and fusion calls from parsed from 240 samples from the supplementary data of the publication. Fusion gene predictions for each samples were extracted from Supplementary Excel Sheet 1 parsed into individual BEDPE files. HLA haplotypes were extracted from Supplementary Excel Sheet 3 into individual haplotype list files with one MHC allele per line.

4.2.4.2 Tool indexes

Indexes for the various tools were generated using the hg38 reference sequence obtained from the UCSC genome browser [60]. GENCODE [43] v25 was chosen as the reference annotation and was used to filter the background SNV databases and SNPEff. Every hg38 index used in the analysis is available in our AWS S3 bucket, 'protect-data'. A detailed list of commands used to create the various indexes is available in the same bucket in the README.

4.2.4.3 Compute Infrastructure

The 326-sample PRAD analysis and the INTEGRATE-Neo comparison were both conducted on a Mesos [48] cluster with one leader (12 CPUs, 62GB RAM, 500GB Local disk) and eight identical agents (56 CPUs, 250GB RAM, 1.8TB local disk). The pVAC-Seq comparison on the 3 Melanoma samples was conducted on an Amazon Web Services EC2 c3.8xlarge instance (32 CPUs, 60GB Ram, 640GB Local disk) and the data was stored securely using SSE-C encryption on S3.

4.2.5 Methods

We ran 3 experiments to demonstrate our pipeline. The first experiment was run on 326 samples from the TCGA PRAD cohort and highlights the scalability, efficiency and utility of ProTECT. The second experiment compares ProTECT to the published SNV- and INDEL-based neoepitope prediction pipeline, pVAC-Seq. The third experiment compares ProTECT to the published fusion-based neoepitope predictor, INTEGRATE-Neo. In all experiments, ProTECT was run using a consensus of 2 out of 5 mutation callers and all TransGene fusion filters to remove inter-mitochondrial, inter-immunoglobulin, 5' lncRNA, and transcriptional readthrough events. Results were tabulated using a mix of python scripts and manual curation on a local machine.

4.2.5.1 326-sample PRAD compute

The 326 samples were run in batches of 1, 2, 5, 10, 20, or 50 samples in order to gauge the efficiency and scalability of toil. Each batch size was run 5 times with unique samples to normalize the runtime information. The configuration file for each run was generated from a template containing all the required tool options and paths to the input reference files on the NFS storage server. Each batch was run once on the mesos cluster using all nodes and an NFS-based Toil file job store to save the state of the pipeline. The five single-sample batches were also run separately without mesos on individual nodes of the cluster using an NFS-based Toil file job store to document the time taken per sample on a single machine.

4.2.5.2 Comparison with PVAC-Seq

To compare our results with PVAC-Seq, we ran ProTECT on the input samples on AWS EC2 using an S3-based cloud job store. The input configuration for the run included paths to HG38-mapped reference files from our AWS S3 bucket 'protect-data' and paths to the input FASTQ files in another secure bucket. The results were stored on AWS S3 in the same bucket as the input. This analysis was conducted consistent with the mandatory cloud data use limitations on the input dataset.

4.2.5.3 Comparison with INTEGRATE-Neo

To compare our results with INTEGRATE-Neo, we had to convert the data from the supplement into files acceptable by ProTECT (Methods). The initial input configuration file consisted of links to the fusion BEDPE for each of 240 samples, along with haplotype and expression data called from the 326 sample run. The final analysis included fusion and inferred haplotype calls from 83 INTEGRATE-Neo samples along with ProTECT expression estimates. All ProTECT runs were conducted on the mesos cluster.

4.2.6 Results and Discussion

4.2.6.1 326 Sample run

To describe the scalability, utility and efficiency of ProTECT, we ran ProTECT on a total of 326 genomic trios from the TCGA PRAD cohort. We called a median of 77 SNVs, 3 INDELs, and 7 Fusion Genes per sample, and accepted 19, 1, and 4 respectively for the production of IARs

(Table 4.1). We identified a median of 11 IARs per sample. Of the 326 samples, 3 samples were predicted to have no actionable mutants (Expressed missense SNV, INDEL, or fusion). TCGA-CH-5743 was part of the TCGA data freeze list and was known to have no somatic SNVs or INDELs. INTEGRATE has 4 fusions called for the sample (including one TMPRSS2-ERG event) with 1 read support each but ProTECT did not detect any of them. TCGA-G9-6370 was not in the INTEGRATE list but was expected to have one variant, chr4:88494316G>T, as reported in the TCGA MAF. This call was rejected by both MuTECT and RADIA for insufficient ALT allele support. TCGA-FC-A66V was not in the data freeze and did not have any INTEGRATE calls. Of the remaining 323 with at least 1 actionable mutant, only 4 were predicted to produce no IARs. The entire results table is submitted as Supplementary Table 1. We detected a total of 12 recurrent gene-fusion pairs (Table 4.2) and 71 recurrent SNVs and INDELs (Supplementary Table 2).

We detected the well-documented TMPRSS2-ERG fusion gene in 131 samples. We predicted at least one IAR each arising from 5 of the 10 unique breakpoints predicted (Table 4.3). 4 of the breakpoints are located in the 5' UTR of TMPRSS2 and will not result in a neoepitope. The last breakpoint we did not call an epitope from has a 5' intronic breakpoint and a 3' exonic one, and the resulting neoepitope should contain the translated product from the last few bases of TMPRRS2 Exon 1 and the first bases after the de novo splice acceptor is reached in ERG. This case is not handled by TransGene at this time. We predict the IAR "DNSKMALNSEALSVVSED" from the junction chr21:41498119-chr21:38445621 in 37 of the 48 unique samples harboring that junction (11% of the entire cohort). Among the MHC molecules predicted in the group to bind to a peptide from this IAR are HLA-

Metric	Called Variants			Accepted Variants			MHCI IARs
	SNVs	INDELs	Fusions	SNVs	INDELs	Fusions	
Mean	94.74	5.69	8.82	22.67	2.74	4.68	12.85
SD	111.58	28.88	5.68	33.89	18.54	4.45	19.71
Median	77	3	8	19	1	4	11
Min	5	0	0	0	0	0	0
Max	1536	504	49	448	329	46	273

Table 4.1: Statistics for 323 samples with at least 1 accepted variant. We predict at least 1 MHCI IAR in 319 samples with a median of 11 per sample. As expected, SNVs are the dominant variant type.

A*02:01 (Allele frequency: 0.26) and HLA-C*07:01 (Allele frequency: 0.17) (Figure 4.7), two alleles that are frequently seen in Caucasian populations. Similarly, we predict "SGCEERGAAGSLISCE" from 22/35 samples with chr21:41507950-chr21:38445621, binding to HLA-C*07:01, HLA-C*04:01, and HLA-B*44:02 (0.14, 0.12, 0.08 allele frequency respectively) (Figure 4.10). These results suggest the possibility of a universal epitope for vaccinating patients with TMPRSS2-ERG. The epitope that will benefit any patient harboring the hotspot breakpoint and an HLA haplotype capable of detecting a neoepitope produced from that fusion. Further experimental validation would be required to assess the viability of the two IARs as vaccine candidates.

Gene name ENSEMBL	Gene name HUGO	Count
ENSG00000184012-ENSG00000157554	TMPRSS2-ERG	131
ENSG00000158715-ENSG00000157554	SLC45A3-ERG	7
ENSG00000184012-ENSG00000175832	TMPRSS2-ETV4	6
ENSG00000158715-ENSG00000006468	SLC45A3-ETV1	4
ENSG00000184012-ENSG00000006468	TMPRSS2-ETV1	3
ENSG00000139865-ENSG00000258601	TTC6-RP11-81F13.2	3
ENSG00000167861-ENSG00000243069	HID1-ARHGEF26-AS1	2
ENSG00000146963-ENSG00000006468	LUC7L2-ETV1	2
ENSG00000152894-ENSG00000093144	PTPRK-ECHDC1	2
ENSG00000128891-ENSG00000122565	C15orf57-CBX3	2
ENSG00000184012-ENSG00000110108	TMPRSS2-TMEM109	2
ENSG00000203727-ENSG00000111961	SAMD5-SASH1	2

Table 4.2: Recurrent fusions called by ProTECT. PRAD is characterized by an abundance of TMPRSS2 fusions with genes in the ETV family (TMPRSS2-ERG being the most popular)

Breakpoint	Count	5' breakpoint	3' breakpoint	Neoepitope expected?	IAR	Count
21:41508081-21:38445621	122	5'UTR	Exon 2	No	NA	
21:41498119-21:38445621	48	Exon 2	Exon 2	Yes	DNSKMALNS EALSVVSED	37
21:41507950-21:38445621	35	Exon 1	Exon 2	Yes[3]	SGCEERGAA GSLISCE	22
21:41508081-21:38474121	18	5'UTR	Intron 1	No	NA	
21:41506445-21:38445621	18	Intron 1	Exon 2	Yes[1]	NA	
21:41508081-21:38584945	11	5'UTR	5'UTR	No	NA	
21:41498119-21:38474121	7	Exon2	Intron 1	Yes[2]	DNSKMALNS LNSIDDAQL	7
21:41508081-21:38423561	7	5'UTR	Exon 3	No	NA	
21:41498119-21:38423561	4	Exon2	Exon 3	Yes[3]	DNSKMALNS ELS	1
21:41494356-21:38445621	3	Exon 3	Exon 2	Yes[3]	SPSGTVCTS RSLISCE	3

Table 4.3: Recurrent TMPRSS2-ERG breakpoints in the cohort. IARs from 21:41498119-21:38445621 and 21:41507950-21:38445621 are recurrent suggesting their viability universal peptide vaccine candidates. We do not expect to see an IAR from fusions with 5'UTR breakpoints. [1] TransGene cannot handle de novo splice acceptors. [2] An Epitope will exist where the TMPRSS2 reads into the intron of ERG. [3] A frameshift is seen on the ERG side of the fusion.

We detected a number of recurrent mutations in the SPOP gene concordant with previous reports [81, 8, 11]. We detected 7 unique recurrent variants across 19 samples that map to 3 different amino acid substitutions in the SPOP protein, p.F133C/V/I/L, p.F102C/V, and p.W131G (Table 4.4). The mutation at position 133 might be of immunological interest since Leucine, Isoleucine and Valine have small hydrophobic side-chains and peptides from the 3 IARs may bind to the same MHC molecules in similar configurations. Table 4.5 shows similar predicted binding affinities of p.F133V/I/L variants to HLA-A*02:01 (as an example) using the IEDB suite. Samples containing SPOP mutations and samples harboring a TMPRSS2-ERG fusion are mutually exclusive of one another, suggesting that vaccine therapies against SPOP and the TMPRSS2-ERG fusion will target different populations of PRAD patients.

Variant	Count	Gene	Mutant	IAR	Frequency
chr17:49619064A>C	5	SPOP	p.F133V	RFVQGKDWG **V** KKFIRRDFL	4
chr17:49619063A>C	3	SPOP	p.F133C	RFVQGKDWG **C** KKFIRRDFL	2
chr17:49619064A>T	2	SPOP	p.F133I	RFVQGKDWG **I** KKFIRRDFL	1
chr17:49619062G>T	2	SPOP	p.F133L	RFVQGKDWG **L** KKFIRRDFL	2
chr17:49619281A>C	2	SPOP	p.F102C	CPKSEVRAK **C** KFSILNAKG	2
chr17:49619282A>C	3	SPOP	p.F102V	CPKSEVRAK **V** KFSILNAKG	3
chr17:49619070A>C	2	SPOP	p.W131G	AYRFVQGKD **G** GFKKFIRRD	2

Table 4.4: Recurrent mutants in the SPOP gene target 3 codons. The F133V/C/I/L mutant may be of interest as a universal neoepitope due to the similar chemical properties of Leucine, Isoleucine and Valine.

Epitope	p.F133V	p.F133I	p.F133L	WT
FVQGKDWG**X**	1.4	3.4	3	17
VQGKDWG**X**K	70	71	63	58
QGKDWG**X**KK	88	85	87	74
GKDWG**X**KKF	59	53	68	74
KDWG**X**KKFI	29	30	31	26
DWG**X**KKFIR	91	90	91	89
WG**X**KKFIRR	61	54	55	55
G**X**KKFIRRD	80	82	46	87
XKKFIRRDF	99	98	98	79

Table 4.5: Predicted binding affinities (better than *n* percent of a background set) of 9-mers arising from the SPOP mutants affecting p.F133 (FVQGKDWG **X** KKFIRRDF for X={V, I, L}) to HLA-A*02:01. The similar chemical properties of Valine, Leucine and Isoleucine lead to similar binding predictions of neoepitopes substituting them for Phenylalanine. Wildtype epitope affinity for reference.

We compared our HLA haplotyping results to the POLYSOLVER calls and consistent with prior work [63], we see the HLA-A*02:01/HLA-A*01:81 miscall in 34 samples. However, 29 of these samples are predicted to be homozygous HLA-A*02:01 by POLYSOLVER so the effect of this miscall will be to add information to the final ranked IARs from one additional alleles. Since most IARs contain peptides predicted to bind to more than one allele, this artifact

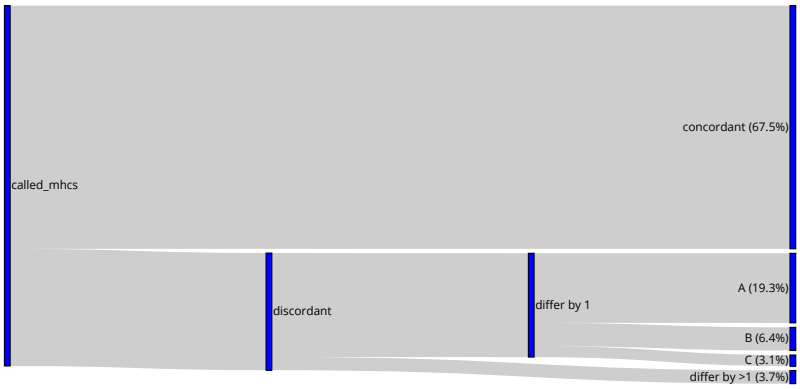


Figure 4.2: HLA Haplotypes called by ProTECT (using PHLAT) are fully concordant with POLYSOLVER haplotypes in only 67.5% of samples. 28.8% differ by 1 call and 3.7% by 2 calls. A majority of the miscalled HLA-A alleles are a documented PHLAT artifact.

should not skew the results by too much. The remaining 5 samples lacked an HLA-A*01:01 call. This means our results are lacking peptide binding affinity predictions for this allele. Overall, 67.5% of all samples had perfectly concordant haplotypes with POLYSOLVER, 28.8% differed by 1 allele and 3.7% differed by 2 (Figure 4.2). 28.8% of the second group consisted of the miscall mentioned above. Other highly frequent miscalls included HLA-A*26:01 to HLA-A*25:01 (10.17%) and HLA-B*27:02 to HLA-B*27:05 (8.5%) (Supplementary Tables 13, 14). This effect is out of the scope of this project and ProTECT allows users to provide pre-computed MHC haplotype calls if they trust another external caller over PHLAT, or if they have a biologically obtained haplotype.

The 326 samples were run on our cluster in different batch sizes to profile average sample runtime (Figure 4.3). As the number of samples increases, we see and expected increase in overall time, but the average time per sample decreases drastically. We processed samples

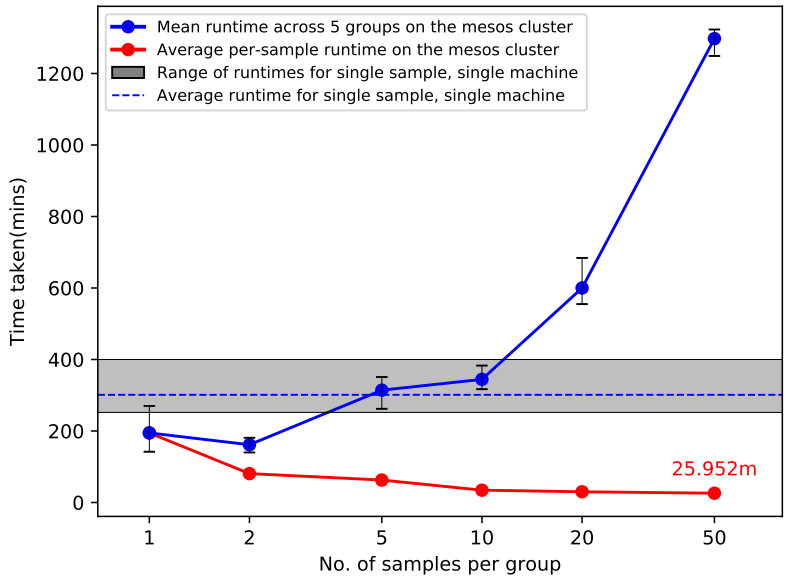


Figure 4.3: Average runtimes on our cluster when ProTECT is run in a batch of 'n' samples. Each batch size is run with 5 unique sample sets and the range of runtimes is described by the whiskers at each datapoint. The grey bar describes the result of running ProTECT on a single sample on one machine. ProTECT takes considerably less time on average when run in a large group.

from end-to-end at a rate of $\approx$ 26 minutes per sample when running in a batch of 50 samples.

4.2.6.2 Comparison with published callers

We ran ProTECT on three melanoma samples that were used to benchmark pVAC-Seq. We called the three expected immunogenic variants in each of the three samples. In some cases, we even predicted the expected variant in a sub-sample where the original paper missed it (E.g. The E153K variant in CDKN2A in the Lymph Node of Mel-21) (Table 4.6). Overall, we ranked IARs containing the expected variants relatively highly (in the top 15-20%) except in Mel218. One reason that could explain this result is that we are looking at all epitopes showing good predicted binding to every MHC alleles in the patient's haplotype, while the clinical trial focused on HLA-A*02:01-restricted peptides.

In addition to the expected variants, we also provided a larger ranked set of possible candidates that broaden the spectrum of testable epitopes. The data for all 7 tested peptides is provided in Supplementary Table 3 and the ProTECT results for all samples are provided in Supplementary Tables 4-11.

Sample and Source	Mel 21				Mel 38			Mel 218
	LN	Skin (2012)	Skin (2013)		Abdominal Wall	Axilla LN	Breast	LN
			RNA 1	RNA 2				
Total Variants	1532	2140	1681	1679	1213	1121	1259	2176
Actionable Variants	332	400	393	391	219	216	224	449
Total IARs	105	137	114	116	73	80	86	155
Vaccine Candidates with ProTECT ranks	NDC1:F169L (Reported as TMEM48 F169L)				SEC24A:P469L			EXOC8:Q656P
	4	5	10	3	8	9	2	152
	TKT:R438W				AKAP13:Q285K			PABPC1:R520Q
	6	6	4	4	14	80	17	140
	CDKN2A:E153K				OR8B3:T190I			MRPS5:P59L
	21	-	18	17	11	13	14	26

Table 4.6: ProTECT ranks for the validation data of PVAC-Seq. Some patients had biopsies taken at multiple time points. Green: Dominant epitope, Orange: Subdominant epitope, Red: Cryptic epitope. Non-highlighted ranks describe instances where PVAC-Seq failed to call the epitope.

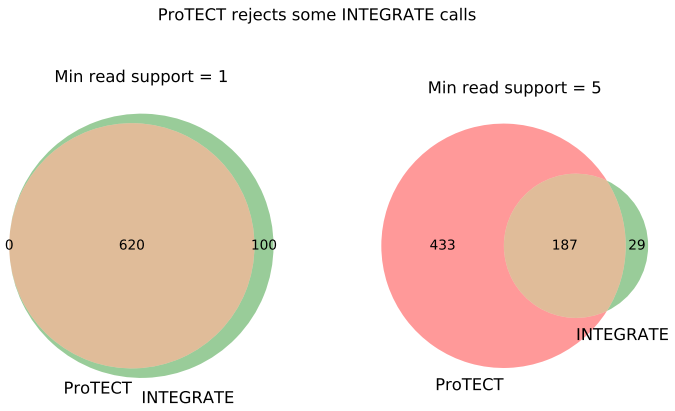


Figure 4.4: Fusion calls between ProTECT (STAR + Fusion-Inspector) are not concordant with INTEGRATE. A large number of INTEGRATE fusions have read support <5 (left) however some of these are called by ProTECT with >5 support (right).

We compared our fusion prediction accuracy with the published INTEGRATE-Neo. The test dataset consists of INTEGRATE-Neo-predicted epitopes from 240 samples that have not been validated using any biological experiments. We use the INTEGRATE-neo calls as a ground truth for or comparison. We first attempted to compare our ProTECT fusions with the fusion calls generated from INTEGRATE. The overlap between the two was very poor and a quick look at the reported read support for the INTEGRATE calls in the form of "spanning reads" showed that a lot of fusions were called with less than 5 read support. Figure 4.4 shows the overlap between ProTECT predicted fusions and INTEGRATE with a minimum of 1- and 5-read support. Furthermore, the MHC haplotypes called in the paper (using HLAMiner [117]) showed very little concordance with the POLYSOLVER and ProTECT-called alleles (Figure 4.5). 61 of the unique MHC alleles called did not match any of the other two callers and

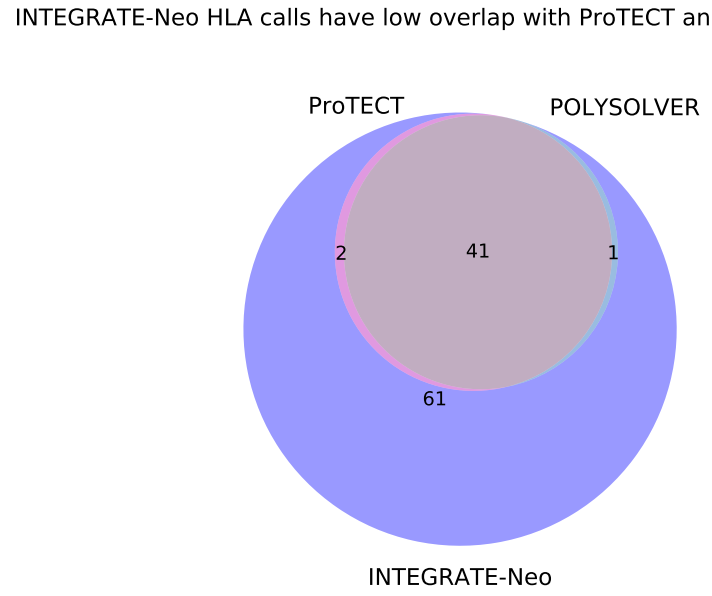


Figure 4.5: HLA haplotypes called by HLAMiner in the INTEGRATE-Neo paper have very low overlap with ProTECT and POLYSOLVER.

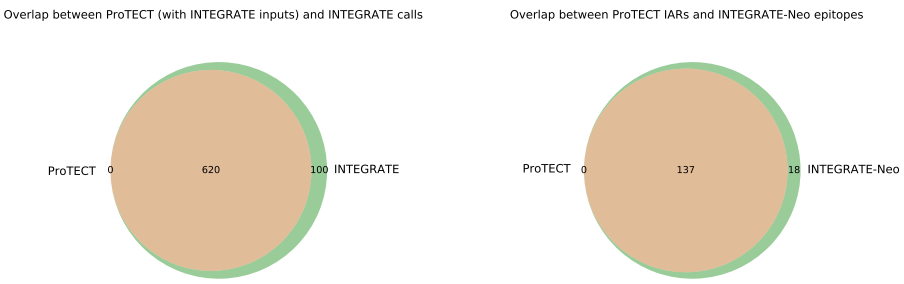


Figure 4.6: ProTECT rejects 100/720 INTEGRATE calls for being transcriptional readthroughs (92) or for having a 5' non-coding RNA partner (8) (Left). ProTECT predicts 137 of the expected 155 epitopes called by INTEGRATE-Neo (Right).

41 matched both. 2 alleles were common between ProTECT and INTEGRATE only, and 1 between INTEGRATE and POLYSOLVER. In order to conduct a more comparable analysis, we re-ran ProTECT with the INTEGRATE fusion calls and the scraped MHC haplotypes from the INTEGRATE-Neo supplementary. ProTECT rejected 100 of the 620 samples as transcriptional readthroughs (92) or for having a 5' non-coding RNA partner (8) (Figure 4.6). The remaining fusions were predicted by INTEGRATE-Neo to produce 155 neoepitopes. ProTECT correctly identified 137 of these neoepitopes as IARs and rejected the remaining 18 for scoring below the 5% threshold (13), having a 5' breakpoint in the UTR(4), or for not having a 5' non-coding partner (1) (Supplementary table 12, Figure 4.6). We believe that easing our 5% filter will increase the number of false positives called and stand by our decision to reject all these calls.

The epitopes predicted from the 4 fusions arising from the 5'UTR breakpoints align with IARs predicted from the same breakpoint but with a different 5'gene partner at the same locus (Figures 4.8 and 4.9). This artifact might be a "mis-annotation" arising due to differences between our GENCODE annotation and the annotation used by INTEGRATE. This highlights the importance of processing every aspect of a neoantigen prediction pipeline using a standard genomic reference.

4.2.7 Conclusion

We have described an efficient, automated, and portable workflow for the prediction of neoepitope candidates that can guide vaccine-based or adoptive cell therapy. We have shown that ProTECT scales well on a parallel processing environment and shows great efficiency gains as the number of samples processed in a batch increases. On average, we processed a sample from

end-to-end in 26 minutes when we ran 50 samples in a single batch on an 8-node cluster. We have shown that ProTECT is comparable to existing callers and provides a ranked list of neoepitopes arising from SNVs, INDELs and fusion genes. Positive results from a clinical trial were ranked highly in our results and we retrospectively picked up events missed by the caller used to guide the trial. We identified recurrent epitopes arising from the well-documented TMPRSS2-ERG fusion our cohort and theorize the possibility of a universal vaccine for one of the possible breakpoints. We suggest the possibility of targeting a recurrent SPOP point mutation that affects a completely different cohort of patients without a TMPRSS2-ERG fusion. ProTECT has immense potential in the rapidly growing cancer vaccine field and further improvements in the vaccine-design paradigm will ensure the viability of vaccines as a standard-of-care for cancer therapy in the future.

4.3 Supplementary information

4.3.1 Supplementary Note 1

In order to use the STAR-generated BAM files for fusion detection, we used the following parameters:

```
--twopassMode Basic
--outReadsUnmapped None
--chimSegmentMin 12
--chimJunctionOverhangMin 12
--alignSJDBoverhangMin 10
--alignMatesGapMax 200000
--alignIntronMax 200000
--chimSegmentReadGapMax parameter 3
--alignSJstitchMismatchNmax 5 -1 5 5
```

4.3.2 Supplementary Tables

All supplementary tables are provided as separate sheets in a Microsoft excel document and are named chapter4-supplementary_tableX where X is the number of the table.

- **Supplementary Table 1:** The overall metrics for the 326 samples PRAD ProTECT run. This table includes all metrics captured for each file including input file sizes, and numbers of mutations called and accepted, peptides generated, and IARs predicted.

- **Supplementary Table 2:** The recurrent SNVs and INDELs (n>1) detected in the PRAD cohort along with the predicted IAR and the frequency of the IAR within the subset of samples that contained the variant.

- **Supplementary Table 3:** The ProTECT ranks for all 7 vaccine candidates tested by Carreno et.al.

- **Supplementary Table 4:** Ranked IARs predicted by ProTECT using sequencing data obtained from tumor tissue in the Lymph Node of Mel 218.

- **Supplementary Table 5:** Ranked IARs predicted by ProTECT using sequencing data obtained from tumor tissue in the Breast of Mel 38.

- **Supplementary Table 6:** Ranked IARs predicted by ProTECT using sequencing data obtained from tumor tissue in the Axillary Lymph Node of Mel 38.

- **Supplementary Table 7:** Ranked IARs predicted by ProTECT using sequencing data obtained from tumor tissue in the Abdominal Wall of Mel 38.

- **Supplementary Table 8:** Ranked IARs predicted by ProTECT using sequencing data obtained from tumor tissue in the Skin of Mel 21 obtained in 2013 (with the second RNA FASTQ pair).

- **Supplementary Table 9:** Ranked IARs predicted by ProTECT using sequencing data obtained from tumor tissue in the Skin of Mel 21 obtained in 2013 (with the first RNA FASTQ pair).

- **Supplementary Table 10:** Ranked IARs predicted by ProTECT using sequencing data obtained from tumor tissue in the Skin of Mel 21 obtained in 2012.

- **Supplementary Table 11:** Ranked IARs predicted by ProTECT using sequencing data obtained from tumor tissue in the Lymph Node of Mel 21.

- **Supplementary Table 12:** Detailed reasons for rejecting 18 INTEGRATE-Neo-predicted neoepitopes. Table A describes 5 breakpoints that were missed due to issues with the 5' partner and Table B describes the INTEGRATE-Neo- and ProTECT-predicted binding affinity for the fusion epitope.

- **Supplementary Table 13:** A detailed list of all samples where PHLAT miscalled one MHC molecule compared to POLYSOLVER (Ground truth).

- **Supplementary Table 14:** A detailed list of all samples where PHLAT miscalled two MHC molecule compared to POLYSOLVER (Ground truth).

4.3.3 Supplementary Figures

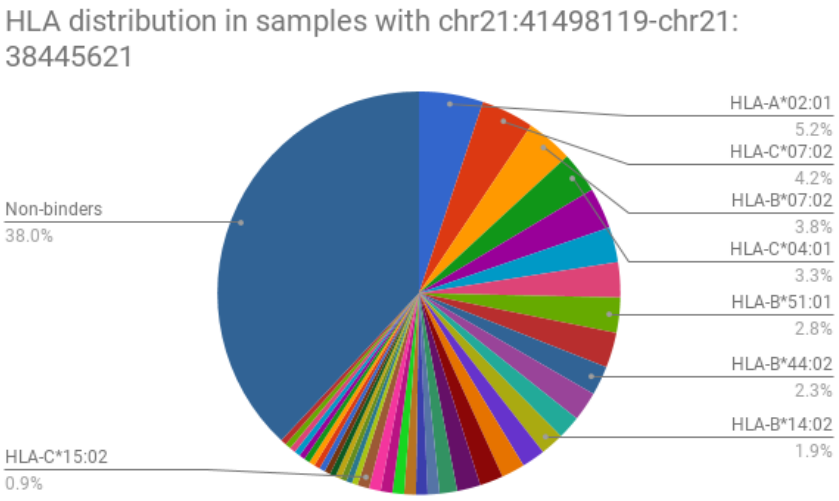


Figure 4.7: MHC alleles called in samples with the chr21:41498119-chr21:38445621 TMPRSS2-ERG breakpoint

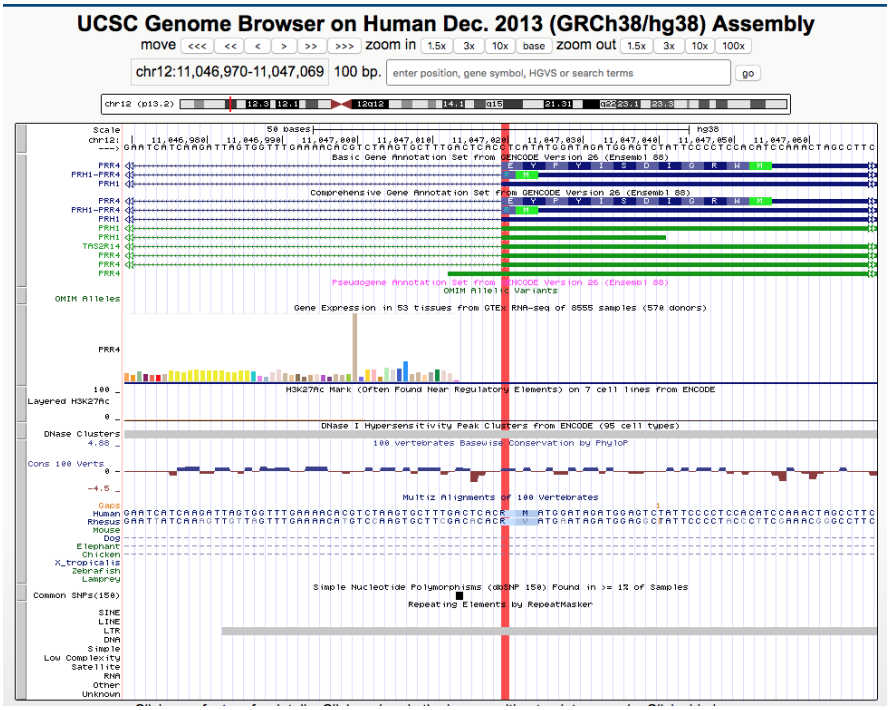


Figure 4.8: A UCSC genome browser showing the 5' breakpoint (highlighted) for the two missed epitopes from the ENSG00000231887-ENSG00000003056 fusion. The 5' partner was reported as PRH1 but the screenshot shows that the position is in the 5' UTR for PRH1. The overlapping PRR4 contains the epitope predicted by INTEGRATE-Neo.

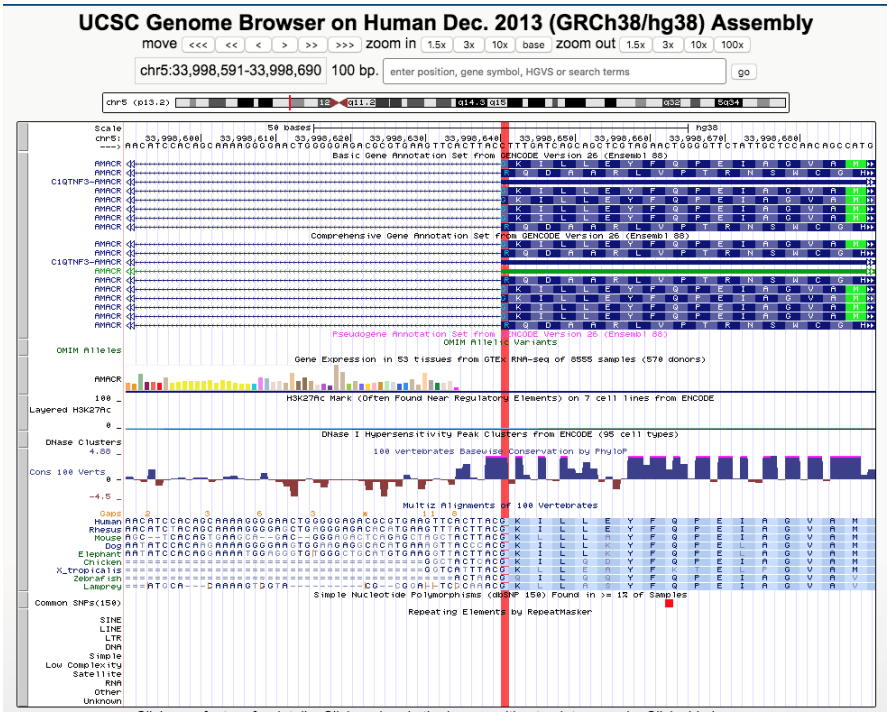


Figure 4.9: A UCSC genome browser showing the 5' breakpoint (highlighted) for the two missed epitopes from the ENSG00000273294-ENSG00000164182 fusion. The 5' partner was reported as the readthrough transcript C1QTNF3-AMACR but the screenshot shows that the position is in the 5' UTR for C1QTNF3-AMACR. The overlapping AMACR contains the epitope predicted by INTEGRATE-Neo.

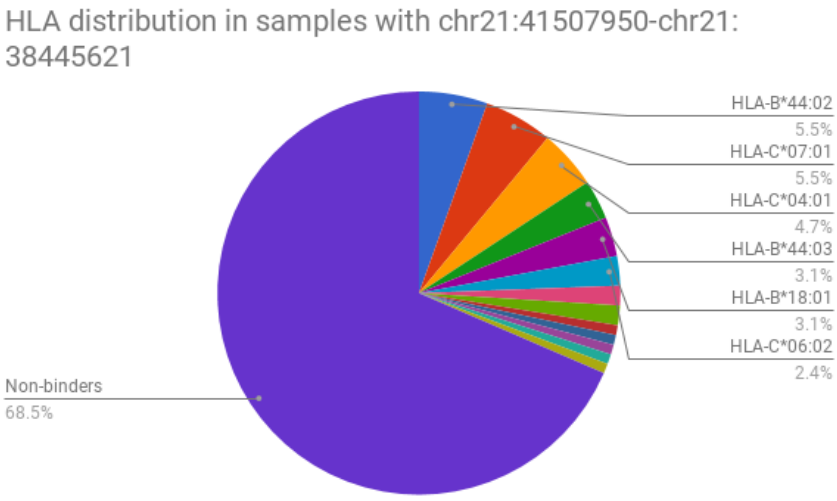


Figure 4.10: MHC alleles called in samples with the chr21:41507950-chr21:38445621 TMPRSS2-ERG breakpoint

Chapter 5

Identification of a potentially therapeutic hotspot neoepitope in Pediatric Neuroblastoma

5.1 Introduction

This work describes a collaborative project between the Haussler Lab at UCSC (Department of Biomolecular Engineering), the Sgourakis Lab at UCSC (Department of Chemistry and Biochemistry) and the Maris lab at The Children's Hospital of Philadelphia (Division of Oncology, Center for Childhood Cancer Research). We demonstrate the clinical utility of ProTECT by identifying 2 potentially therapeutic peptides arising from the hotspot ALK R1275Q mutation ($\approx$4% of all NBLs) in high-risk pediatric Neuroblastoma (NBL) from the Therapeutically Applicable Research to Generate Effective Treatments (TARGET) initiative [87]. The neoepitopes were predicted to bind strongly to the human HLA-B*15:01 MHC molecule ($\approx$ 8% of Caucasian individuals). A T-cell receptor receptive to this peptide and MHC combination (pMHC)

would theoretically be universal, and can be used to treat any new patient harboring this combination of variant and MHC.

The pMHC binding of the ALK R1275Q derived nonamer (AQDIYRASY) and decamer (ADQDIYRASYY) to HLA-B*15:01 were demonstrated biochemically in an *Escherichia coli* display system and the reactivity to $CD8^+$ T-cells was demonstrated using Peripheral Blood Mononuclear Cells (PBMCs) from healthy HLA-B*15:01 matched donors. The non-binding of the wild-type counterparts (ARDIYRASY and ARDIYRASYY) was demonstrated biochemically, and molecular modeling provided an explanation for the phenomenon. This work is significant because high-risk NBL is generally associated with extremely poor prognosis and has an overall survival rate of lower than 50% even after intensive chemotherapy, radiation therapy, or other approved treatments [73, 87]. This work was published in Frontiers in Immunology in January 2018.

The project involved 3 components – Bioinformatics for the identification of individual and shared neoepitopes in the TARGET NBL cohort; Biochemistry for the demonstration of pMHC binding, calculation of binding energy, and molecular modeling (including simulations); and Cell Biology for the demonstration of $CD8^+$ reactivity to the pMHC combination. The biochemistry work was conducted by the Sgourakis group and the Cell Biology work by the Maris Group. I conducted the Bioinformatics component with the help of my undergrad research assistant, Ada Madejska. This work included the identification and procurement of relevant and usable NBL data, analysis via ProTECT, and interpretation of results. I also wrote the sections concerning the Bioinformatics component in the "Results" and "Materials and Methods", and generated Figure 1, supplementary table 1, and supplementary data sheets 1, 2 and 3 (Excel

tables). The paper and supplementary information are shown below. Supplementary data sheets 1, 2, and 3 are too large to fit in this thesis and can be viewed on the Frontiers in Immunology website.

5.2 The pediatric NBL study paper

frontiers in Immunology

ORIGINAL RESEARCH
published: 30 January 2018
doi: 10.3389/fimmu.2018.00099

A Recurrent Mutation in Anaplastic Lymphoma Kinase with Distinct Neoepitope Conformations

Jugmohit S. Toor[1†], Arjun A. Rao[2†], Andrew C. McShan[1], Mark Yarmarkovich[3], Santrupti Nerli[1,4], Karissa Yamaguchi[1], Ada A. Madejska[5], Son Nguyen[6], Sarvind Tripathi[1], John M. Maris[3], Sofie R. Salama[2,7], David Haussler[2,7] and Nikolaos G. Sgourakis[1]**

[1] Department of Chemistry and Biochemistry, University of California, Santa Cruz, Santa Cruz, CA, United States, [2] Department of Biomolecular Engineering, University of California, Santa Cruz, Santa Cruz, CA, United States, [3] Division of Oncology, Center for Childhood Cancer Research, Children's Hospital of Philadelphia, Philadelphia, PA, United States, [4] Department of Computer Science, University of California, Santa Cruz, Santa Cruz, CA, United States, [5] Department of Molecular, Cell, and Developmental Biology, University of California, Santa Cruz, Santa Cruz, CA, United States, [6] Department of Microbiology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, United States, [7] Howard Hughes Medical Institute, University of California, Santa Cruz, Santa Cruz, CA, United States*

OPEN ACCESS

Edited by:
Jin S. Im,
University of Texas MD Anderson Cancer Center, United States

Reviewed by:
Brian M. Baker,
University of Notre Dame, United States
Sébastien Wälchli,
Oslo University Hospital, Norway
Evan W. Newell,
*Singapore Immunology Network (A*STAR), Singapore*

***Correspondence:**
David Haussler
haussler@soe.ucsc.edu;
Nikolaos G. Sgourakis
nsgourak@ucsc.edu

†These authors have contributed equally to this work.

Specialty section:
This article was submitted to Cancer Immunity and Immunotherapy, a section of the journal Frontiers in Immunology

Received: *05 October 2017*
Accepted: *12 January 2018*
Published: *30 January 2018*

Citation:
Toor JS, Rao AA, McShan AC, Yarmarkovich M, Nerli S, Yamaguchi K, Madejska AA, Nguyen S, Tripathi S, Maris JM, Salama SR, Haussler D and Sgourakis NG (2018) A Recurrent Mutation in Anaplastic Lymphoma Kinase with Distinct Neoepitope Conformations. Front. Immunol. 9:99. doi: 10.3389/fimmu.2018.00099

The identification of recurrent human leukocyte antigen (HLA) neoepitopes driving T cell responses against tumors poses a significant bottleneck in the development of approaches for precision cancer therapeutics. Here, we employ a bioinformatics method, Prediction of T Cell Epitopes for Cancer Therapy, to analyze sequencing data from neuroblastoma patients and identify a recurrent anaplastic lymphoma kinase mutation (*ALK* R1275Q) that leads to two high affinity neoepitopes when expressed in complex with common HLA alleles. Analysis of the X-ray structures of the two peptides bound to HLA-B*15:01 reveals drastically different conformations with measurable changes in the stability of the protein complexes, while the self-epitope is excluded from binding due to steric hindrance in the MHC groove. To evaluate the range of HLA alleles that could display the *ALK* neoepitopes, we used structure-based *Rosetta* comparative modeling calculations, which accurately predict several additional high affinity interactions and compare our results with commonly used prediction tools. Subsequent determination of the X-ray structure of an HLA-A*01:01 bound neoepitope validates atomic features seen in our *Rosetta* models with respect to key residues relevant for MHC stability and T cell receptor recognition. Finally, MHC tetramer staining of peripheral blood mononuclear cells from HLA-matched donors shows that the two neoepitopes are recognized by CD8+ T cells. This work provides a rational approach toward high-throughput identification and further optimization of putative neoantigen/HLA targets with desired recognition features for cancer immunotherapy.

Keywords: neoepitopes, MHC class I, human leukocyte antigens, structural biology, computational biology, cancer, T cell receptor

INTRODUCTION

Cancer immunotherapy harnesses a patient's CD4+ and CD8+ T cell responses toward peptide neoantigens, which are displayed on the surface of tumor cells by major histocompatibility complex molecules [MHC, termed human leukocyte antigen (HLA) in humans] (1). In the endogenous presentation pathway (MHC class I), abundantly expressed intracellular proteins are processed by the immunoproteasome and proteasome to yield short peptide fragments that are transported

into the endoplasmic reticulum and assembled together with the MHC-I heavy chain and β2-microglobulin light chain (β_2m) by the peptide-loading complex (2). The resulting peptide/MHC complexes (p/MHC) are further trafficked through the Golgi and eventually displayed on the cell surface, where they are surveilled by CD8+ cytotoxic T cells (CTLs) through specific interactions with αβ T cell receptors (TCRs) (3). Through this process, a large and heterogeneous pool of p/MHC antigens is continuously generated in healthy, pathogen infected, or tumor cells as a means of displaying a cell's peptide repertoire to the immune system (4). The display of high affinity peptides expressed exclusively by the tumor (i.e., neoepitopes) on MHC molecules can elicit specific CTL responses, which forms the basis of several established immunotherapies against cancers (5, 6). One such therapy utilizes *in vitro*-activated, autologous CTLs to selectively target tumor cells (7). Alternatively, vaccines can be designed based on known antigens or CTLs can be engineered to introduce TCRs with desired specificities toward displayed tumor antigens (8). In all cases, neoepitopes derived from commonly mutated oncogenic proteins are well-suited immunotherapy targets if they have high affinity interactions with MHC alleles that are prevalent in the population (9).

Neuroblastoma (NBL) is a widely metastatic form of cancer that affects the development of nerve cells that comprise the sympathetic nervous system, primarily in patients younger than 10 years old (10). High-risk NBL has a survival rate of less than 50% after intensive chemotherapy, radiation therapy, and other approved treatments (11). In addition, patients responding positively to radiation treatments generally do not achieve long-term survival and suffer from cancer relapse, often with an increased rate of tumor mutations (12). Sequencing studies focusing on NBL of all stages indicate a wide spectrum of somatic mutations in tumors, which poses a significant challenge for the development of targeted therapeutics (13). Notably, mutations in the anaplastic lymphoma kinase gene (*ALK)* have been implicated in 9.2% of 240 NBL cases with available whole exome, genome, and transcriptome sequencing data from the TARGET (Therapeutically Applicable Research to Generate Effective Treatments) initiative (12). This and other sequencing data support *ALK* as the target with the highest mutation rate among high-risk NBL patients (10, 12, 14). Furthermore, genome sequencing of relapsed NBL tumors demonstrates retention of ALK mutations and/or acquisition of an *ALK* mutation in 14/54 (15) and 10/23 (16) samples. Such *ALK* mutations have been shown to hyperactivate the RAS–MAPK signaling pathway in NBL, driving cancer formation (17). More recent studies have also shown evidence of ALK overexpression in NBL tumors making it a viable target for CAR-mediated immunotherapy along with other targeted T cell therapies (18). Immunotherapy offers an attractive approach toward NBL treatment. However, despite significant progress in identifying recurrent mutations toward understanding the genetic basis of NBL, important molecular details regarding derived neoantigen/HLA interactions remain unknown, which further limits the development of targeted T cell therapies (11).

Here, we use our recently developed multilayered bioinformatics pipeline, Prediction of T Cell Epitopes for Cancer Therapy (ProTECT), to predict therapeutically relevant antigens in NBL tumors. ProTECT analysis of 106 patient samples from the NBL TARGET cohort identifies a recurring "hotspot" mutation in the *ALK* protein (R1275Q), together with its specificity toward common HLA alleles. Specifically, two putative peptide sequences with the R1275Q mutation, a nonamer and a decamer, are predicted to bind HLA-B*15:01 with high affinity according to consensus methods (19, 20). X-ray structures of the two neoepitopes in complex with HLA-B*15:01 reveal a drastic change in peptide conformation, which correlates with increased thermal stability of the decamer neoepitope/HLA complex. For the self-peptide, unfavorable interactions between the peptide and residues in the MHC-binding groove prevent the formation of a stable complex. To evaluate the potential of the two ALK neoepitopes to interact with additional HLA alleles and predict structural features relevant for recognition by TCRs, we develop a high-throughput comparative modeling approach using the program *Rosetta*. Independent crystallographic analysis of a decamer-bound HLA-A*01:01 complex reveals a peptide conformation, which falls extremely close to our *Rosetta* model (within 1.1 Å backbone RMSD). Finally, tetramer staining of peripheral blood mononuclear cells (PBMCs) from HLA-B*15:01-matched donors followed by flow cytometry analysis shows that the two different neoantigen conformations are recognized by CD8+ T cells. Taken together, our bioinformatics analysis, *in vitro* and structural characterization, computational modeling, and T cell recognition analysis illustrate a powerful approach toward high-throughput identification and optimization of broadly displayed putative neoantigen/HLA targets for further development toward cancer immunotherapy. Results from this approach provide strong evidence for broad HLA display of recurrent *ALK*-derived neoantigens expressed in NBL tumors and further suggest that the presentation of distinct neoepitope conformations in the HLA groove could drive specific CD8+ T cell responses in patients.

RESULTS

Identification of *ALK* R1275Q Neoepitopes Using ProTECT

A reduced version of our software, ProTECT (**Figure 1**), was initially run on a batch of six primary:relapsed NBL sample pairs from the TARGET cohort. We find at least one neoepitope-generating mutation persisting in the relapsed tumor for five of six patients (Table S1 and Supplementary Data S1 in Supplementary Material). Among these are two well-known hotspot mutations, *NRAS* Q61K and *ALK* R1275Q (Table S1 in Supplementary Material). We predicted two HLA-B*15:01-restricted decamer (MAQDIYRASY and AQDIYRASYY) and one nonamer (AQDIYRASY) neoepitopes arising from *ALK* R1275Q in sample TARGET-30-PARHAM. The predicted binding affinities are better than 0.55, 0.85, and 2.1%, respectively, relative to all peptides in a background training set (the top 5% ranked peptides are considered true binders by our method). While the peptide beginning at M1273 is predicted to be the top binder, the two epitopes beginning at A1274 are more promising from an immunological perspective since they are predicted to be significantly better binders to HLA-B*15:01 than their parental self-antigens ARDIYRASYY (10.75 percentile score) and ARDIYRASY (35 percentile score).

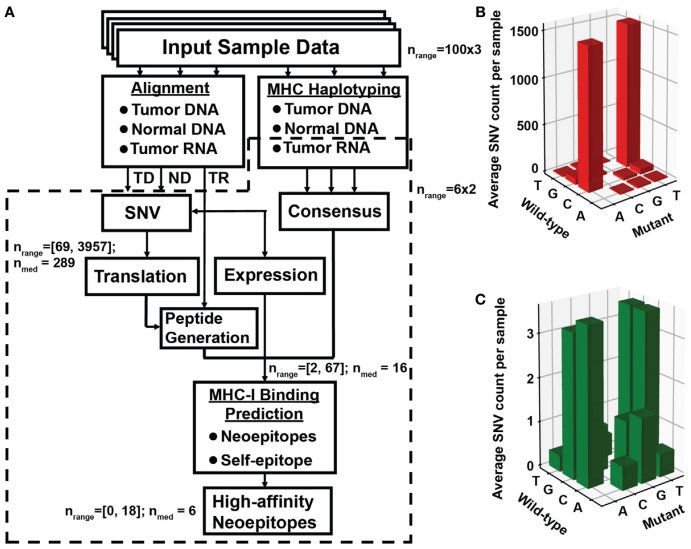


FIGURE 1 | Identification of neoantigen targets using the ProTECT pipeline. **(A)** Flowchart indicating each step of the ProTECT pipeline. Input FASTQs trios per sample ultimately give rise to MHC haplotyping and provide a list of candidate neoepitopes for each sample. Abbreviations: TD and ND, tumor and normal DNA, respectively; TR, tumor RNA. Predicted tumor–normal single-nucleotide variants (SNVs) are filtered during peptide generation, and again at neoepitope prediction. n_{range}: the range of SNV calls that make it past a certain step; n_{med}: median number of calls. The primary:relapse pairs were run through a smaller modified version of the pipeline that started directly from mutations curated from Eleveld et al. (16). Panels **(B,C)** show the TARGET neuroblastoma cohort OxoG mutation level. Before filtering for OxoG artifacts, we see a predominance of C>A/G>T mutants **(B)**, whereas after filtering we see a marked reduction in the total number of mutations and a more balanced nucleotide substitution rate **(C)**.

Using the full version of the ProTECT software (manuscript in preparation), we expanded our study to 100 primary NBL samples in the TARGET cohort to collect more complete statistics on *ALK* R1275Q-derived neoepitopes, and to identify other recurrent neoepitopes in NBL (Table S1 and Supplementary Data S2 and S3 in Supplementary Material). We identify four additional samples harboring *ALK* R1275Q (TARGET-30-PANWRR, -PANXJL, -PAPTFZ, and -PAPTLV). None of these samples express HLA-B*15:01, but sample PAPTFZ displays a close relative, HLA-B*15:03 that is predicted to bind ARDIYRASYY and ARDIYRASY with scores of 2.2 and 4.7%, respectively. Two samples (PANXJL and PAPTLV) express the high-frequency HLA-A*02:01 (20% in Caucasian populations), where an *ALK* R1275Q nonamer (GMAQDIYRA) is predicted to bind HLA-A*02:01 with a 1.4% score.

All but six of the 100 samples harbor one or more non-synonymous neoepitope with low percentile scores for at least one expressed HLA allele. Among these we identify other recurrent mutations, including the *ALK* mutation F1174L/I/C, present in 3/2/1 samples, respectively, and a *ZNF717* mutation (Q716), present in three samples (Table S1 in Supplementary Material). One sample in the cohort expresses *NRAS* Q61K, an activating mutation commonly found in melanoma, thyroid, and colorectal cancers. Finally, the *NRAS*-derived neoepitope ILDTAGKEEY arising from a single mutation (Q61K) is predicted to bind the common HLA-A*01:01 allele with a statistically significant score of 0.35%. Notably, the same HLA-A*01:01/ILDTAGKEEY interaction identified by our method has been previously shown to elicit a specific T cell response using a melanoma cell line (21).

ALK Tumor Neoantigens Form p/MHC Complexes with Distinct Stabilities *In Vitro*

The results obtained from ProTECT analysis provide a range of therapeutically relevant neoantigen/HLA interactions to validate and characterize using biophysical and structural methods. Given the extensive genetic evidence supporting a role for *ALK* mutations in NBL tumors (15, 17), we chose to pursue further the interaction between *ALK* R1275Q and HLA-B*15:01. We prepared recombinant HLA-B*15:01 bound to the two *ALK*-derived neoantigens, a nonamer (AQDIYRASY) and a decamer (AQDIYRASYY). As a control, we attempted to prepare HLA-B*15:01 with the self-antigen (ARDIYRASY), which is predicted to have a >10-fold reduced binding affinity for the HLA. Peptide/MHC samples were refolded from purified *Escherichia coli* inclusion bodies

in the presence of 10-fold molar excess peptide using standard methods and purified by size exclusion chromatography (SEC) (22). SEC traces of the nonamer and decamer samples show three distinct peaks corresponding to protein aggregate (22.8 min), p/MHC complex (29.5 min), and free β_2m (42.7 min) (**Figure 2A**). Notably, the sample refolded using the self-antigen peptide shows only two peaks in the chromatogram, none of which contain non-aggregated p/MHC molecules (**Figure 2A**, green trace), further suggesting that the affinity of the self-antigen is insufficient to promote the formation of a stable complex with the HLA.

To confirm the presence of the neoepitopes in the two MHC samples, we performed liquid chromatography–mass spectroscopy (LC–MS). LC–MS reveals a high relative abundance of the correct peptide in each sample, with observed masses of 1,086.70 and 1,249.71 Da, which agree well with the expected masses of the nonamer and decamer, respectively (**Figures 2B,C**). Thus, we

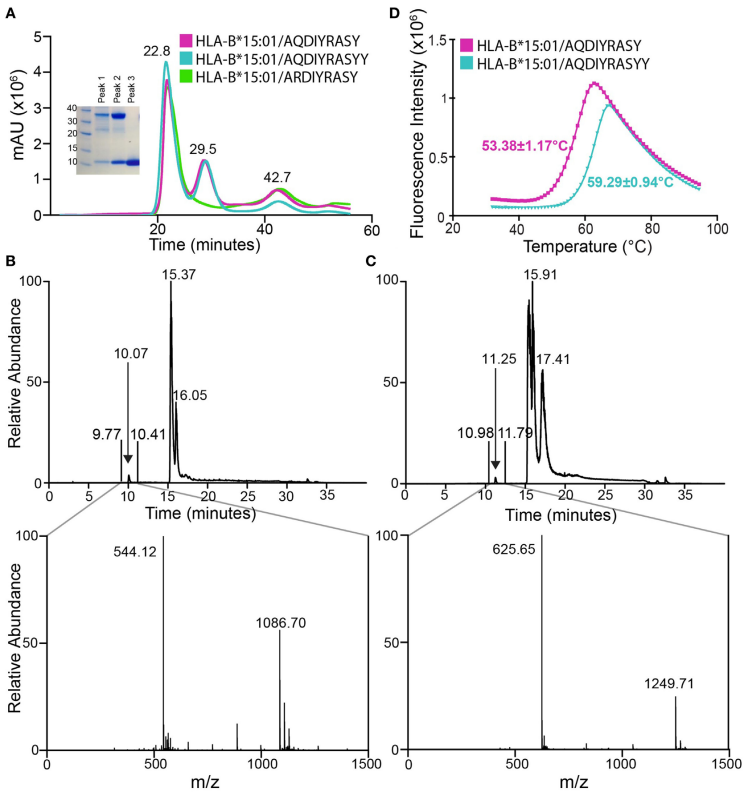


FIGURE 2 | Association of anaplastic lymphoma kinase neoepitopes with recombinant HLA-B*15:01 *in vitro*. **(A)** Size exclusion chromatography (SEC) traces of MHC samples refolded with nonamer (magenta), decamer (teal), or self (green) peptides. Purification was performed on a HiLoad 16/600 Superdex 75 pg column at a flow rate of 2 mL/min. Eluted fractions were probed using SDS-PAGE analysis followed by Coomassie staining (left) and show expected molecular weights for HLA-B*15:01 (32.4 kDa) and β_2m (11.8 kDa). Further analysis reveals SEC peak identities as protein aggregate (22.8 min), p/MHC complex (29.5 min), and free β_2m (42.7 min). Attempts at refolding HLA-B*15:01 with the self-peptide did not produce a p/MHC complex (green curve, lack of 29.5 min peak). LC–MS analysis of purified nonamer **(B)** and decamer **(C)** HLA-B*15:01 complex samples. The top panel shows the chromatogram trace of each complex, while the bottom panel is the average relative abundance for the time interval between 9 and 11 min, showing the presence of either the nonamer (observed mass 1,086.70 Da; expected mass 1,086.17 Da) or the decamer peptide (observed mass 1,249.71 Da; expected mass 1,249.35 Da) captured in the MHC peptide-binding groove. **(D)** Differential scanning fluorimetry shows that the decamer-bound MHC complex (teal) has an increased thermal stability of 59.3°C relative to the 53.4°C T_m observed for the nonamer-bound MHC complex (magenta).

confirm binding of the two tumor neoepitopes to recombinant HLA-B*15:01 prepared through *in vitro* refolding. To further characterize the resulting p/MHC molecules, we used a differential scanning fluorimetry (DSF) assay, which can accurately assess kinetic stability. According to this technique, properly folded class I p/MHC complexes show melting temperatures (T_m) from 37 to 63°C, which correlate with predicted IC_{50} values in the micromolar to nanomolar range (23). Here, both neoantigen p/MHC samples show a clear unfolding transition with a highly reproducible T_m of 53.4°C for the nonamer and 59.3°C for the decamer complex (**Figure 2D**), suggesting that the decamer forms a higher affinity complex with HLA-B*15:01. Such a difference in thermal stabilities of the p/MHC complexes together with previous observations that peptide length influences its conformation within a fixed-length groove is consistent with a hypothesis that the two peptides are displayed *via* distinct binding modes, as previously reported for nonamer and decamer peptides sampling unique conformations within an MHC groove (24).

Structural Plasticity within the MHC Peptide-Binding Groove Enables Distinct Neoantigen Conformations

To elucidate the structural basis underlying the distinct stabilities observed for the two *ALK* neoepitopes and to further characterize peptide features displayed to TCRs we solved the X-ray structures of the nonamer (HLA-B*15:01/β_2m/AQDIYRASY) (PDB ID 5TXS) and the decamer complex (HLA-B*15:01/β_2m/AQDIYRASYY) (PDB ID 5VZ5). The nonamer complex crystallized in the $P2_12_12_1$ space group at a resolution of 1.7 Å, while the decamer complex crystallized in the $P6_122$ space group at a resolution of 2.6 Å (Table S2 in Supplementary Material). The nonamer peptide adopts a canonical extended conformation promoted by the N-terminal (Ala1, Gln2) and C-terminal (Tyr9) anchors, which are deeply embedded within A/B, and F-pockets of the HLA groove (**Figures 3A,B**), respectively. This anchoring results in a "curved" conformation, where the backbone of residues from Asp3 to Ser8 is pushed toward the upper part of the groove while the remaining residues are maintained within the C, D, and E pockets (**Figure 3B**). A survey of previously deposited HLA-B*15:01-restricted antigens in the PDB (LEKARGSTY derived from Epstein–Barr virus, PDB ID 1XR8; ILGPPGSVY derived from human ubiquitin-conjugating enzyme-E2, PDB ID 1XR9; VQQESSFVM derived from SARS coronavirus, PDB ID 3C9N) reveals other nonamer epitopes consistently in extended conformations (25, 26), in agreement with the conformation of the *ALK* nonamer neoepitope in our X-ray structure (Figures S4A–D in Supplementary Material). Furthermore, the overall architecture of the B*15:01-binding groove is similar between the different structures with heavy atom backbone RMSDs of less than 1 Å (Figure S4E in Supplementary Material). Comparison between the peptide amino acid sequences reveals excellent agreement with the established HLA-B*15:01-binding motifs, where LMQ/AEISTV and FY/LM are preferred/tolerated in anchor positions 2 and 9, respectively (Figures S4F,G in Supplementary Material). Thus, the X-ray structure of our *ALK*-derived nonamer neoepitope is consistent with established structural features in the PDB, suggesting a trend where the peptide backbone conformation is defined by its length and anchor motifs.

Generally, peptides of length greater than nine amino acids either bulge further out of the binding groove or form a "zig-zag" conformation (27). However, in our decamer complex structure (**Figures 3C,D**), the peptide adopts a short 3_{10} helical backbone conformation from Ile4 to Ala7, as confirmed by an inspection of φ/ψ backbone dihedral angles (Figure S1A in Supplementary Material). Notably, while the N-terminal anchor residues are identical in the nonamer and decamer peptide, Tyr10 of the decamer replaces Tyr9 of the nonamer as the C-terminal anchor residue in a similar conformation (**Figures 3B,D**). The accommodation of a longer peptide sequence within the fixed-size MHC groove is thus achieved through the formation of a more compact 3_{10} helix for the decamer, relative to the extended nonamer backbone. In addition, the 3_{10} helix buries Arg6 further into the MHC groove and creates an amphipathic structure where Ile4, Tyr5, Ser8, and Tyr9 are oriented toward the solvent (**Figure 3D**). A structural superposition of the nonamer and decamer peptides (2.7 Å backbone heavy atom RMSD) highlights the changes in residues that are oriented toward the solvent, suggesting that the two epitopes display very different surface features for interactions with TCRs (**Figures 3E–G**).

The compaction of the peptide backbone in the decamer structure is accompanied by structural adaptations of MHC residues in the peptide-binding groove. In particular, in the decamer complex the HLA α_2 helix undergoes a significant widening involving a 5.1 Å displacement of the Cα atom of Arg151. This movement is driven by a change in orientation of Arg151, which points toward the solvent in the nonamer versus toward the groove in the decamer complex (Figures S2A,B in Supplementary Material), and the burying of Arg6 further toward the floor of the groove. Thus, the addition of a C-terminal Tyr in the peptide sequence drastically alters the tertiary structure of the HLA complex, driven by a widely different peptide conformation that can be accommodated through conformational plasticity within a malleable MHC groove.

Key structural parameters extracted from our crystallographic analysis provide insights into the increased stability of the decamer/HLA complex. Notably, the buried surface area (BSA) between HLA-B*15:01 and the decamer peptide is 1,986 $Å^2$, relative to 800 $Å^2$ in the nonamer structure. To further dissect different structural features for their contributions to p/MHC stability, we analyzed all polar (hydrogen bonds, salt bridges, and electrostatic interactions) and hydrophobic interactions involving HLA residues (Figure S5 and Table S3 in Supplementary Material). Specifically, the decamer peptide forms additional intra-peptide hydrogen bonds as a result of the more compacted 3_{10} helix conformation. In addition, the decamer participates in 25 polar and 21 hydrophobic interactions with the MHC residues, while the nonamer forms 26 polar but only 11 hydrophobic interactions with the groove (Figure S5A in Supplementary Material). Specifically, Asp3 and Arg6 of the decamer peptide extend further into the groove, forming additional contacts with HLA side chains (Figures S5B,C in Supplementary Material). Our structural analysis suggests that an increase in the total number of intra-peptide hydrogen bonds and hydrophobic packing

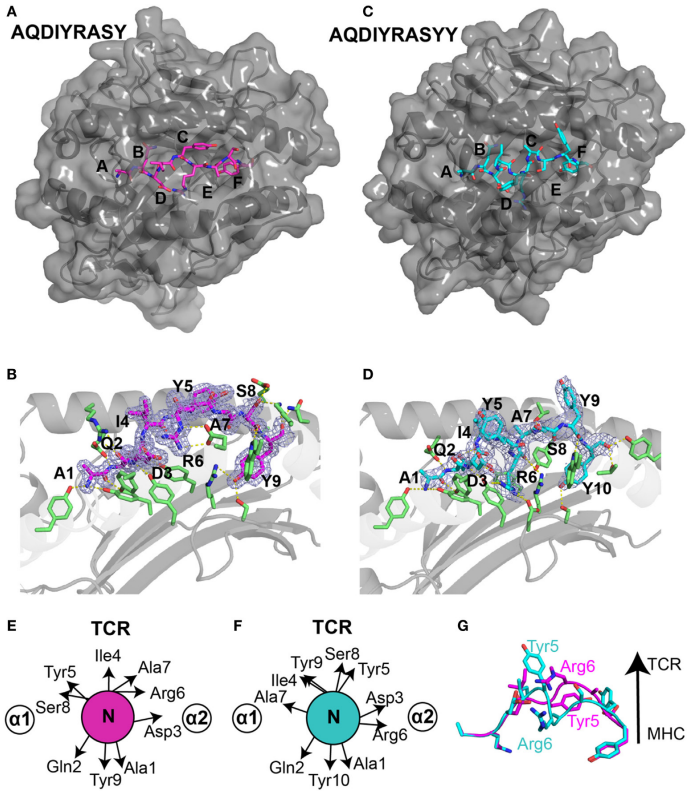


FIGURE 3 | Structural differences in *ALK* neoepitope displayed by HLA-B*15:01. X-ray structures of the **(A)** nonamer peptide (PDB ID 5TXS), shown as magenta sticks and **(C)** decamer peptide (PDB ID 5VZ5), shown as cyan sticks embedded into the groove of HLA-B*15:01 molecule. The canonical peptide-binding pockets in HLA groove are indicated with letters. **(B)** Nonamer peptide (magenta sticks) and **(D)** decamer peptide (cyan sticks) with $2F_o - F_c$ electron density maps contoured at 1.2 σ within the groove of HLA-B*15:01. Yellow dashes represent polar contacts between the peptide and selected MHC residues (green sticks). Side-chain orientation of the **(E)** nonamer peptide and **(F)** decamer peptide as viewed from the top axis of the peptide highlighting the placement of different residues. **(G)** Structural superposition heavy backbone atoms of the bound nonamer (magenta sticks) and decamer (cyan sticks) neoepitopes (all-atom RMSD of 4.0 Å) reveal distinct T cell receptor (TCR)-interacting residues between the two neoantigens.

interactions, consistently with an increase in BSA brought on by the more compact 3_{10} helical conformation, leads to an improved stability of the decamer complex, as confirmed independently by our DSF experiments (**Figure 2B**).

Structural Exclusion of the Self-antigen from the HLA-B*15:01 Groove

To further evaluate the potential immunogenicity of the *ALK* R1275Q neoepitopes, we compared their affinity for HLA-B*15:01 relative to the self-peptide (ARDIYRASY). Formation of a stable HLA complex displaying the self-peptide would compromise the therapeutic relevance of any related neoantigen, due to immune tolerance mechanisms that limit the repertoire of responsive T cells. Preliminary attempts to refold HLA-B*15:01 using a synthetic nonamer peptide with the parental *ALK* sequence did not result in efficient p/MHC formation, suggesting low binding affinity, likely in the micromolar range (**Figure 2A**, green trace). To further explore the basis of this exclusion we performed structural modeling of the self-peptide/HLA-B*15:01 complex,

using our solved X-ray structure of the nonamer complex as a template. We find that performing the reverse Gln to Arg substitution leads to steric hindrance between the longer Arg2 side chain and residues of the MHC-binding groove (Figure S6A in Supplementary Material). Despite a careful consideration of all possible Arg side-chain rotamers, significant clashes remain with Ser67 on the α_1 helix, as well as with Ala24, Met45 on the floor of the MHC groove (Figure S6A in Supplementary Material). As expected from the conservation of peptide residue anchors in the A- and B-pockets, we observe similar clashes when the self-decamer is modeled with HLA-B*15:01. By contrast, the neoepitope Gln2 side chain fits well into the B-pocket, forming an additional hydrogen bond Tyr9 from the HLA heavy chain (Figure S6B in Supplementary Material, cyan dotted line). Finally, we performed detailed structure modeling calculations using simultaneous optimization of the peptide backbone in addition to the side-chain degrees of freedom and ranked the calculated affinities of the three peptides for HLA-B*15:01 according to a physically realistic energy function (28). The self-antigen complexes yield the least favorable binding energies, followed by the nonamer, and finally the decamer complex (Figure S3 in Supplementary Material). Thus, our structural analysis is highly consistent with our *in vitro* results, i.e., that the self-peptide is excluded from binding, in sharp contrast with the nonamer and decamer neoepitopes which form tight complexes with the HLA.

Evaluating the HLA-Binding Repertoire Using Comparative Modeling Calculations

A patient's HLA haplotype plays a major role in determining the outcome of targeted cancer immunotherapies. Therefore, toward expanding the range of individuals that could mount a T cell response to *ALK* R1275Q neoepitopes, we evaluated the potential of other HLAs to display the two peptides *in silico*. Here, we developed and applied a high-throughput approach which exploits the availability of our high-resolution X-ray structures for the two neoepitopes to simultaneously predict peptide/HLA interactions and surface features of peptide residues poised for interactions with TCRs. First, we selected a non-redundant set of 2,904 HLA alleles (885 HLA-A, 1,405 HLA-B, and 614 HLA-C unique sequences) from the EMBL-EBI database (29). We then carried out detailed *Rosetta* comparative modeling calculations for each allele, using our experimentally determined HLA-B*15:01 structures for the nonamer and decamer ALK peptides as templates (**Figure 4**). In contrast to previous structure-based peptide/HLA modeling methods which use a flexible peptide docking approach (30–32), we used a fixed-peptide backbone threading approach followed by energy minimization of the interacting peptide and HLA residues to drastically confine the docking degrees of freedom. Our approach was motivated the observation that the peptide backbone conformation shows minimal variance (less than 1.5 Å RMSD) in all nonamer/HLA-B*15:01 structures reported in the PDB (Figure S4E in Supplementary Material). Using this strategy, we extracted highly reproducible binding energies for both the nonamer and decamer peptides, which are maintained in extended and 3_{10} helical conformations, respectively, in the resulting models

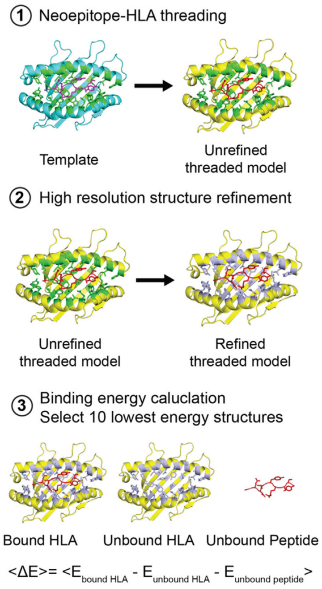


FIGURE 4 | Structure-based modeling of neoepitope/human leukocyte antigen (HLA) interactions. Step 1: A template (blue) peptide/HLA complex (X-ray structure) is provided to generate a threaded model with the same peptide and different HLA alleles (yellow). HLA residues in the groove within 3.5 Å of the peptide are colored green. Step 2: Models are refined by energy minimization and side-chain repacking of groove and peptide residues (gray). Step 3: The average peptide-binding energy is determined by subtracting the energy of the unbound HLA and unbound peptide from the energy of the peptide bound HLA. <E> represents the average binding energy. The top 10 lowest energy structures are compared with determine a consensus model.

(Figure S7 in Supplementary Material). As expected, the HLA-B*15 alleles rank systematically among the top binders, indicating a high degree of groove complementarity to both peptides (**Figure 5**; Figure S8A in Supplementary Material, purple). Among those, the HLA-B*15:84 allele shows the lowest binding energy for the decamer (**Figure 5A**, black circle), whereas the HLA-B*15:107 allele shows the lowest binding energy for the nonamer (Figure S8A in Supplementary Material, black circle). A total of 116 HLA alleles from all A, B, and C types exhibit lower binding energies for both the nonamer and decamer peptides than our initial HLA-B*15:01 structural templates (**Figure 5A**; Figure S8A in Supplementary Material, red square), suggesting the potential for a broader HLA display repertoire.

To elucidate a sequence bias for specific residues in the HLA-binding groove that consistently yield more favorable interactions with the two peptides, we analyzed the average binding energy as a function of sequence identity score (33), calculated relative to the best binding allele for each peptide (**Figure 4**). As a

negative control, we computed the binding energy for a mock HLA allele in which all residues in the MHC-binding groove are mutated to Ala. As expected, the mock polyAla HLA exhibits a low binding affinity (i.e., high-binding energy) to the peptide and is distant from the best binding allele (**Figure 5**; Figure S8A in Supplementary Material, green triangle, top left). We observe an

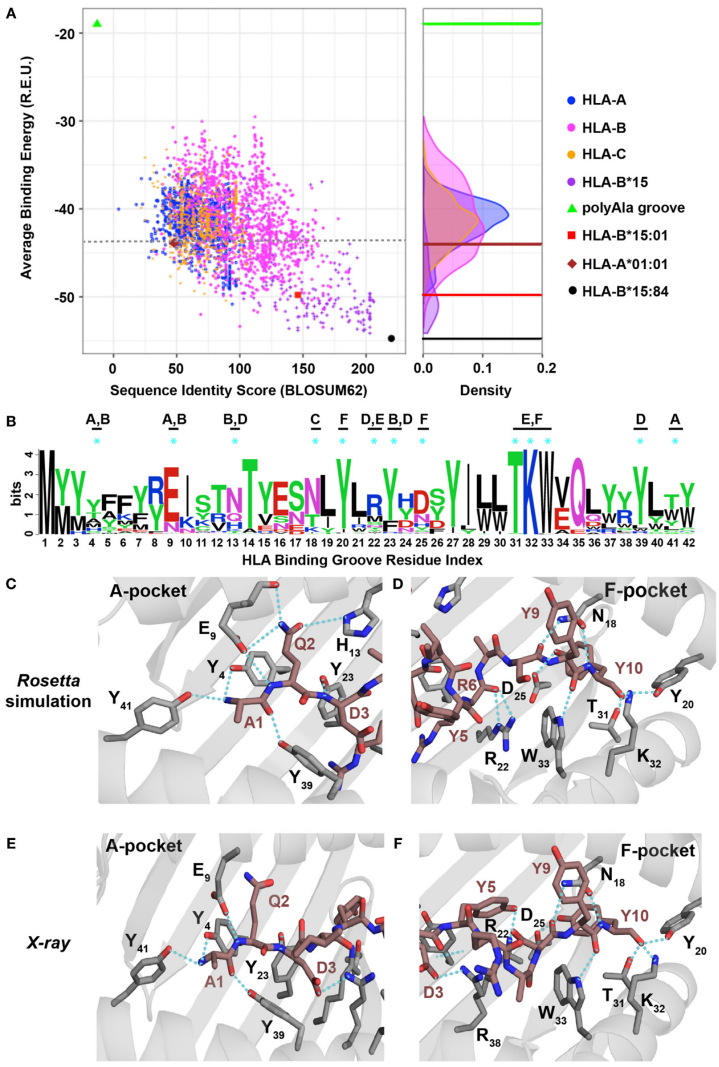


FIGURE 5 | Continued

FIGURE 5 | Evaluating the human leukocyte antigen (HLA)-binding repertoire of *ALK* decamer AQDIYRASYY using *Rosetta* structure-based modeling. **(A)** *Rosetta*-binding energies calculated from structure modeling of 2,904 unique HLA alleles from the IPD-IMGT/HLA Database (29), for the *ALK* neoepitope decamer (AQDIYRASYY) plotted as a function of sequence similarity to the top binding allele, HLA-B*15:84 (black circle). The binding energy of decamer in our HLA-B*15:01 X-ray structure is shown as a reference (red square). A negative control was performed with a mock HLA allele where all residues in the binding groove were replaced with Ala (polyAla groove, green triangle), which shows high-binding energy. The corresponding distribution of the HLA alleles on the binding energy landscape is captured in the density plot shown on the right. Sequence identity scores were calculated using the BLOSUM62 (33) matrix. Abbreviation: R.E.U., *Rosetta* energy units. **(B)** Kullback–Leibler sequence logo derived from multiple sequence alignment using ClustalOmega of peptide-binding groove residues from all the HLA alleles that exhibit better binding energies than HLA-A*01:01 (brown diamond), indicated with a gray dotted line in panel **(A)**. MHC residues with polar contacts to the peptide are denoted with a cyan asterisk with corresponding MHC pocket noted. **(C,D)** Threaded structural model of HLA-A*01:01 displaying decamer peptide. Polar contacts between the MHC groove (gray sticks) and peptide (brown sticks) are shown with cyan dotted lines in the A-, B-, and D-pockets **(C)** or C-, E-, and F-pockets **(D)**. The residue index for each interacting MHC residue is denoted with the corresponding number from panel **(B)** using subscripts. Peptide residues (non-indexed) are labeled without subscripts. Panels **(E,F)** show polar contacts observed in the A-pocket **(E)** and F-pocket in the X-ray structure of HLA-A*01:01/AQDIYRASYY (PDB ID 6AT9) between the peptide (brown sticks) and residues in the MHC groove (gray sticks). The residue index for each interacting MHC residue is denoted with the corresponding number from panel **(B)** using subscripts. Peptide residues (non-indexed) are labeled without subscripts.

evident correlation between the computed binding energies and sequence similarity to the top binder. Our approach additionally allows us to decompose residue specific contributions to overall binding energy for each peptide–HLA combination. We find a clear trend for both the nonamer and decamer peptides with a set of HLA alleles where a bulk of the binding energy is provided by the "anchor" positions (Figures S11A,B in Supplementary Material). By contrast, the mock polyAla HLA exhibits considerably higher binding energy across the entire peptide length (Figures S11A,B in Supplementary Material). To elucidate key sequence features that allow the peptides to be accommodated in the MHC groove, we derived a sequence profile among good binders for the two neoepitopes. Such features are highlighted in the Kullback–Leibler sequence logo, which reveals preferred residues in the HLA peptide-binding groove (**Figure 5B**; Figure S8B in Supplementary Material). According to this metric, highly invariant residues in the MHC-binding groove should play an essential role in mediating peptide/MHC interactions, as they are consistently observed in HLA alleles that exhibit high affinity binding. A close inspection of our structural models for the nonamer and decamer bound to a common allele in our data set, HLA-A*01:01, reveals similar polar contacts, primarily in the A-, B-, and F-pockets, that correlate well with the positions of invariant MHC residues (**Figures 5C,D**; Figures S8C,D in Supplementary Material). Specifically, both the nonamer and decamer C-terminal anchors employ a similar interaction pattern in the F-pocket with conserved Thr, Lys, Trp, and Tyr residues of the MHC (**Figures 5B,D**; Figure S8B,D in Supplementary Material).

To test the validity of our structure-based simulations, we performed *in vitro* refolding of the *ALK*-derived nonamer and decamer peptides with HLA-A*01:01. This allele was chosen because it is a high-frequency allele in multiple populations worldwide and has been previously shown to form stable recombinant p/MHC complexes for structural characterization (34). As observed in our previous experiments with HLA-B*15:01 (**Figure 2A**), refolding of HLA-A*01:01 with decamer or nonamer peptide results in a stable p/MHC complex (Figures S8E and S9A in Supplementary Material). Further characterization of the purified complex reveals a thermal stability of 47.9°C for the decamer (Figure S9B in Supplementary Material) and 46.7°C for the nonamer (Figure S8F in Supplementary Material), suggesting that both *ALK* neoepitopes have a lower affinity for HLA-A*01:01 compared with HLA-B*15:01 (**Figure 2D**, 59.3°C), consistently with our binding energy calculations (**Figure 5A**; Figure S8A in Supplementary Material). Although certain HLA and H2 MHC alleles have been previously reported to yield partially folded, peptide-free molecules with measurable thermal stabilities (35), control refolding experiments performed without peptide for each of our HLA alleles failed to yield a stable complex. Finally, to conclusively test the atomic features predicted by our simulations, we determined the X-ray structure of decamer complex HLA-A*01:01/β2m/AQDIYRASYY (PDB ID 6AT9). The decamer complex crystallized in the $P32_12$1space group at a resolution of 2.9 Å (Table S2 in Supplementary Material). Inspection of crystallographic φ/ψ dihedral angles reveals that the peptide backbone also adopts a short 3_{10} helix conformation when bound to HLA-A*01:01, suggesting that the peptide length is the main determinant of its conformation in the groove, and further justifying our fixed-backbone modeling approach (Figure S9F in Supplementary Material). The peptide conformation in the X-ray structure shows excellent agreement with our *Rosetta* model (1.1 backbone heavy atom RMSD), with several high-resolution features predicted by the model are confirmed by the X-ray, including polar contacts within both the A- and F-pockets of the MHC groove (**Figures 5C–F**; Figure S12 in Supplementary Material). Specifically, the side-chain hydroxyl group of the peptide Tyr10 is in contact with the same Tyr, Lys, and Trp side-chain atoms from the F-pocket (**Figures 5D,F**). Finally, in comparison with the X-ray structure of the same peptide bound to HLA-B*15:01, the side chain of Arg6 is flipped outwards from the groove when bound to HLA-A*01:01 altering the peptide surface displayed to TCRs (Figures S9D,E in Supplementary Material). Thus, our independent X-ray structure corroborates the trend observed in our structure-based binding energy simulations and further supports the potential for other HLA molecules to display the recurrent *ALK* neoepitopes with unique TCR interaction properties.

The Two *ALK*-Derived Neoepitopes Are Recognized by CD8+ T Cells

Given the unique conformations and surface features observed for the nonamer and decamer peptides, we sought to determine whether the two altered-self (i.e., mutated) neoantigens could be recognized by CD8+ T cells using a MHC tetramer staining assay followed by multichannel flow cytometry analysis. We hypothesized that an HLA-matched donor would be able to recognize

altered-self neoepitopes in the periphery, as long as the peptide adopts a conformation that can potentiate interactions with TCRs. To test this, we acquired PBMCs from two HLA-B*15:01-matched healthy donors. For each peptide, we performed a double staining experiment using HLA-B*15:01 tetramers conjugated with allophycocyanin (APC) or phycoerythrin (PE), toward identification of T cells that recognize each neoepitope. Final cell sorting using fluorescence-based detection results in identification of double positive populations with a total of 0.012% CD8+ T cells reactive to the nonamer (Figure S13A in Supplementary Material), and 0.024% reactive to the decamer epitope (Figure S13B in Supplementary Material). Notably, these findings were very similar between two independent staining experiments using PBMCs from individual donors (Figure S13C in Supplementary Material). As a control, we additionally performed staining experiments using tetramers made for HLA-B*15:01 complexed with an immunodominant SARS coronavirus-derived epitope (VQQESSFVM) (26). For double staining experiments with HLA-B*15:01/SARS tetramers, we observe double positive populations corresponding to 0.007 and 0.014% reactive CD8+ T cells (Figure S13C in Supplementary Material). Finally, simultaneous staining experiments using nonamer/HLA-B*15:01-PE and decamer/HLA-B*15:01-PE tetramers did not uncover populations of CD8+ T cells that recognized both epitopes (Figure S13C in Supplementary Material). Thus, CD8+ T cells are able to recognize both neoepitopes with nominal frequencies that are comparable to that of a known immunodominant epitope.

DISCUSSION

Immunotherapies that stimulate the immune system to attack tumors, including immune checkpoint blockade and adoptive T cell therapies, have achieved spectacular results in tumor types with high mutational burden, such as melanoma (36). However, their utility in tumors with lower mutational burden, such as those that occur in pediatric cancers, is less clear (37). The development of more targeted T cell based immunotherapies to treat cancer relies on understanding the molecular basis of neoepitope display on tumor cells, in addition to the initiation and regulation of cytotoxic CD8+ T cell responses (38). A current roadblock in the development of robust approaches across patients is that the HLA locus is extremely polymorphic, and an individual's exact HLA haplotype sculpts the repertoire of epitopes displayed to the immune system (5). Moreover, the identification of therapeutically relevant antigens in tumors remains extremely challenging and is further complicated by the fact that a single HLA allele can potentially bind 10^3–10^6 distinct peptide epitopes (39). Traditionally, *in vitro* measurements of affinities between an MHC and a potential antigen were achieved by equilibrium dialysis (35) and fluorescence polarization experiments (40). More recent approaches allow for a global evaluation of the entire peptide repertoire, using mass spectroscopy of MHC complexes extracted from cell lines expressing a single HLA allele followed by bioinformatics analysis (41). Robust alternative strategies to identify and characterize neoepitope/HLA complexes with desired T cell recognition features would significantly bolster the progress of targeted T cell therapies against cancer.

We have recently developed ProTECT, a fully automated and freely available tool for predicting expressed neoepitopes based on the somatic mutations present in tumor samples. In NBL, a common pediatric cancer, ProTECT analysis identifies a range of intriguing predicted high affinity neoepitope–HLA targets that should be examined in future studies, such as *NRAS*:Q61K—HLA-A*01:01 (Table S1 in Supplementary Material), including the *ALK* neoepitopes examined in detail here. Typically, Immune Epitope Database (IEDB)-based binding prediction methods are biased towards nonamer peptides due to limited number of datasets for peptide/MHC-binding affinity measurements of shorter or longer peptide lengths (19, 20). Moreover, affinity thresholds for binding based on IC_{50} values are HLA allele specific and range from 60 to 950 nM (42), which could result in false negative predictions where weak binding epitopes that may be immunogenic are not considered. We attempt to normalize for these limitations in ProTECT by using a suite of predictors trained on combined and/or allele-specific datasets that consider a range of epitope-binding affinities. In our analysis, we find that 90% of the NBL samples have one or more predicted high affinity neoepitope–HLA targets. We sought to characterize the nature of the p/MHC interactions resulting from the relatively common *ALK* R1275Q mutation, to lay the groundwork for developing a targeted immunotherapy for it, and to develop a pipeline for evaluating other promising tumor neoepitopes. Toward these goals, we have elucidated the structural characteristics underlying the *in vitro* stability and presentation of two *ALK* R1275Q-derived nonamer and decamer epitopes where the corresponding self-peptide does not bind to the same HLA groove. We additionally developed and applied a high-throughput comparative modeling approach to identify additional HLA alleles that could display the two neoepitopes and predict their structures with high accuracy, toward understanding the link between peptide surface features and interactions with TCRs. Finally, we examined the potential for the ALK-derived neoepitope/HLA complexes to activate an immune response by analyzing CD8+ T cell recognition from HLA-B*15:01-matched donors.

The exact conformation and dynamic features of the peptide within the MHC-I-binding groove are known to play pivotal roles in recognition by CD8+ T cells, by dictating MHC/peptide-binding affinity, stability on the cell surface and cross-reactivity of interactions with specific TCR molecules (43, 44). Our X-ray structures reveal an extreme case of such conformational plasticity, in which the addition of a single C-terminal Tyr in the neoantigen sequence (AQDIYRASY to AQDIYRASYY) significantly alters the peptide conformation (**Figures 3E–G**). This dramatic change relative to the canonical extended structure is highlighted by the formation of a 3_{10} helix spanning residues Ile4 to Ala7 (Figure S1 in Supplementary Material) and provides a link between peptide conformation and HLA complex stability. Specifically, the 3_{10} helix leads to an increase in BSA and number of molecular interactions between the peptide and HLA side chains (Figure S5 in Supplementary Material), in agreement with its increased thermal stability (**Figure 2D**). We additionally observed changes in the HLA groove, including a displacement of the α2 helix that undergoes a significant widening involving a 5.1 Å movement of the Cα atom of Arg151 (Figure S2A,B in Supplementary

Material). Our results further support the importance of the α2 helix, which participates in a myriad of immune processes, such as chaperone-mediated peptide loading/editing (45), allele-specific antigen presentation (46), and TCR recognition (47), in the context of conformational plasticity of the MHC groove to accommodate epitopes of varying length. Finally, structure modeling of the self-antigen sequence, in agreement with *in vitro* refolding experiments, shows a sharp contrast in stability relative to the neoantigens due to an Arg anchor that cannot be accommodated on either an extended or helical backbone conformation. Our results provide a rational approach for improving neoepitope/HLA complex stability and half-life on the cell surface, relative to unstable self-epitope/HLA complexes, through optimizing the peptide backbone conformation in addition to anchor residue interactions. This could ultimately lead to the selection of more efficient neoantigens, consistently with previous studies showing that the ability of tumor antigens to induce T cell responses that prevent tumor relapse correlates with p/MHC stability (6, 48).

As not all cancer patients who harbor a tumor-specific mutation that results in a neoepitope have the same HLA haplotype, it would be extremely beneficial to expand the repertoire of HLA molecules that bind and present a given therapeutic target. While sequence-based tools available at the IEDB (20) can provide highly reliable predictions of epitope binding for a range of HLA alleles, structural details of the predicted epitope/HLA complex relevant for interactions with TCRs are not provided by such methods. Complementary methods have been used to model interactions within peptide/HLA complexes by leveraging high-resolution structural data available in the PDB. These approaches employ flexible peptide docking to construct sequence specificity profiles by exploring different peptide/HLA combinations (30–32). Here, we utilize a comparative modeling approach with a fixed-peptide backbone while allowing for side-chain flexibility within the HLA groove to screen a large pool of HLA alleles for binding to our *ALK*-derived nonamer and decamer neoantigens (**Figure 5**; Figure S8 in Supplementary Material). High-ranking HLA alleles according to *Rosetta*'s binding energy consistently demonstrate a low percentile rank using the epitope prediction method recommended by IEDB, which further suggests a high probability of forming a tight complex with the neoepitopes (Figure S10 in Supplementary Material). We subsequently test our binding predictions and show that both the nonamer and decamer peptides form a stable complex with the common HLA-A*01:01 *in vitro*, albeit with decreased stability compared with the HLA-B*15:01 bound complex (**Figure 2D**; Figures S8 and S9 in Supplementary Material). The accuracy of the *Rosetta* models is highlighted by a comparison to our decamer/HLA-A*01:01 X-ray structure, which shows a backbone RMSD of 1.1 Å (Table S5 and Figure S12 in Supplementary Material). Our fixed-backbone approach is further supported by the observation that the conformation of the peptide backbone is maintained among X-ray structures containing different, high affinity nonamer peptides bound to HLA-B*15:01 (Figure S4E in Supplementary Material). Moreover, comparison of the decamer peptide conformation when bound to HLA-B*15:01 versus HLA-A*01:01, two alleles that share 51% of groove residues according to a pairwise sequence alignment, shows only a modest change (1.6 Å backbone RMSD) (Table S5 in Supplementary Material). In stark contrast, we observe a significant conformational change between the nonamer and decamer peptides in their crystallographic complexes with the same HLA-B*15:01 allele (2.7 Å backbone RMSD). These results suggest that peptide length defines the backbone conformation through the conservation of anchor residue interactions within a fixed-size class I MHC groove. This feature of peptide binding allows us to confidently model patient-specific neoepitope/HLA interactions in a high-throughput manner, using a single crystal structure containing the same peptide as template. Finally, our approach allows us to predict surface features of neoepitope/HLA complexes available for interactions with TCR molecules, toward further evaluating their immunogenicity. Within the current scope of our method, *Rosetta* accounts for conformational plasticity within the MHC groove by allowing for side-chain rotamer and limited backbone flexibility. Thus, accurate modeling of epitope binding is achieved given the template contains an MHC groove that is accommodated for a fixed-peptide length (i.e., to model a nonamer epitope, a template X-ray structure for a nonamer/HLA complex should be used). However, our current protocol cannot account for large changes in the backbone of the groove, which may be required to model peptides of shorter or longer length (49). Future improvements in our structure-based prediction procedure that account for this may be achieved using *Rosetta*'s Comparative Modeling (*RosettaCM*) hybridize (50) or *RosettaRemodel* (51).

To screen for CD8+ T cells that could recognize the tumor neoantigens, we focused our analysis on lymphocyte samples from healthy donors. We identify populations of CD8+ T cells which recognize our two *ALK* neoepitopes in a highly specific manner and with minimum cross-reactivity between them (Figures S13A–C in Supplementary Material). We observe approximately half the frequency (0.012 and 0.017%) of reactive CD8+ T cells for the HLA-B*15:01/nonamer tetramers relative to the frequency (0.024 and 0.028%) observed for HLA-B*15:01/decamer tetramers (Figure S13C in Supplementary Material), which may suggest differences in T cell recognition between the two epitopes. In addition, the percentage of reactive T cells against our two neoepitopes is comparable to values observed for the immunodominant SARS epitope (Figure S13C in Supplementary Material). While the nominal frequency of T cells specific for most p/MHC molecules ranges from 0.00005 to 0.01% (52, 53), our observed values for HLA-B*15 tetramers are within the range of specific T cells identified in previous reports of PBMC staining of healthy donors using HLA-B*15 tetramers (54). Our staining results support the recognition of our putative nonamer and decamer neoepitopes by CD8+ T cells, potentiating the ability for the epitopes to drive specific immune responses. Engagement of TCR molecules and triggering of signaling of CD8+ T cells are driven by interactions between the TCR complementarity-determining regions and specific peptide/HLA structural motifs (44, 55). Our detailed structural characterization provides further insight in the unique features that give rise to very distinct interface chemistries displayed by the two neoepitopes. It is likely the interplay between HLA complex stability and peptide surface features guides the engagement of CD8+ pools by the two neoantigens. Future studies in our group aim to identify the TCR(s)

that can recognize our HLA displayed *ALK* neoepitopes toward the goal of characterizing the interface of the p/MHC–TCR complexes. Structural characterization of *ALK* p/MHC–TCR complexes will allow us to understand how the conformational plasticity observed in our nonamer and decamer neoepitopes dictates CD8+ T cell recognition (56) toward fostering the development of p/MHC–TCR complexes with improved stability in the immunological synapse (57).

In summary, we outline a novel approach toward robust, high-throughput identification and detailed characterization of highly stable putative neoantigen/HLA targets with desired T cell recognition features for cancer immunotherapy. Recently established technologies have enabled high-throughput, parallel detection of T cell specificities for a wide spectrum of epitopes through the combinatorial encoding of p/MHC multimers (57). Such methods have already been applied to monitor the prevalence of T cells that are reactive for established tumor epitopes (58). In addition, vaccination of cancer patients that display neoantigens can elicit a broad T cell response, both in terms of specificity and clonal diversity (59–61). The success of future cancer immunotherapies based on these technologies will depend on the ability to fine-tune the desired T cell responses toward specific tumor epitopes. Our data suggest that malleable structural features of the target neoepitope/MHC complex can be harnessed to achieve such a fine-tuning. Thus, our characterization of recurring, T cell-reactive neoepitopes together with their HLA specificities and molecular determinants of stability provide new screening tools and therapeutic targets to enable the development of personalized immunotherapies against NBL tumors.

MATERIALS AND METHODS

NBL Sample Data Collection and ProTECT Analysis

One hundred NBL sequencing trios (normal and tumor DNA-seq, and tumor RNA-seq) were downloaded from the National Cancer Institute Genomics Data Commons (NCI-GDC) using the GDC Data Transfer Tool. Samples were all downloaded in BAM format and then converted back to the native paired FASTQ format using the Picard SamToFastq module. Some of the RNA-seq BAM files had reads in the pair mapped with separate read groups. These files were converted to FASTQ using an in-house python script.[1] We processed the samples from raw FASTQ trios to neoepitopes prediction at a rate of ~6 h/sample on four Microsoft Azure machines (Supplementary Data S1 in Supplementary Material). MHC haplotypes for MHC class I and MHC class II are called from the sequencing data using PHLAT (62). The haplotype for a sample is decided based on a consensus decision of the three input haplotypes. Somatic point mutations were called using a panel of five mutation callers, MuTECT (63), MuSE (64), RADIA (65), SomaticSniper (66), and Strelka (67). Since most mutation callers are DNA centric, we allow mutations rejected by up to two of the callers through this first filter. The vcf of first-pass mutants is subjected to SNPEff (68) using indexes generated from the GENCODE v19 annotations for GRCh37 (69). The accepted mutations are further filtered more stringently using an in-house tool, Transgene,[2] before being translated into mutant peptides. Library construction for sequencing can induce artificial oxidation of guanine bases (OxoG) (70) caused by high-energy sonication. These OxoG bases pair with thymine during PCR instead of their regular pairing partner, cytosine. This results in low allele fraction G>T or C>A substitutions seen predominantly in read 1 or read 2, respectively, in the FASTQ. Transgene filters variants arising solely form read 1 or read 2 in the alignment, and low allele-fraction mutants (<0.1 allele fraction) with no RNA-seq coverage. Since non-expressed proteins will never be picked up by the adaptive immune system, we filter events having low RNA-seq coverage. A mutation is filtered if the position has no evidence in the RNA (unexpressed ALT allele), there are reads spanning across, but none covering the position (splice variant), or if the gene is unexpressed. Filtered mutants are translated into peptides of length $2n - 1$ for $n = (9, 10, 15)$ using the GENCODE protein coding translations corresponding to the annotation used. Transcript-specific peptides are generated to account for known splice variants. The peptides generated by transgene are tested for binding against the inferred HLA haplotypes using the IEDB suite of MHC-I and MHC-II epitope predictors.[3] Each $2n - 1$-mer input peptide yields n calls for each allele in the HLA haplotype, for each $n = (9, 10)$ for MHCI and $n = 15$ for MHC-II. Each call represents a combined consensus percent score of the peptide from a number of IEDB algorithms that have been trained on that MHC allele. These methods include an artificial neural network, a stabilized matrix method, a method that uses binding motif obtained from Combinatorial libraries, etc., and each method returns the percent rank of the input peptide:MHC combination versus a background set generated by the IEDB. The consensus score for a call is the median of the scores across all methods for that call. Peptides having a consensus percent score of greater than 5% (i.e., binders worse than the top 5% of the background set) are filtered as non-binders. Peptides having a consensus percent rank of greater than 5% (i.e., binders worse than the top 5% of a background set) are filtered as non-binders. The rank of the self-peptide for each filtered mutant is calculated using the same method. Peptides are grouped by the mutation and transcript(s) of origin into ImmunoActive Regions (IARs), i.e., regions likely to produce a peptide that will stimulate the immune system. IARs are ranked based on the affinity of the best contained binder, expression of the transcript(s) of origin, the promiscuity of the region (the predicted number of MHCs stimulated by peptides in the IAR), and the number of 10-mers in the IAR overlapping a 9-mer that binds to the same MHC as the 10-mer, with similar affinity. In the initial pilot, RNA-seq BAMs from six primary:relapsed pairs of samples were downloaded from the GDC and run through a reduced version of the pipeline using VCF files generated from the supplementary data from Eleveld et al. (16) containing predicted mutations. MHC haplotypes for these samples were decided based on the consensus calls from the primary and relapsed RNA-seq. All samples were run

[1] https://github.com/arkal/random/process_rgs.py.

[2] https://github.com/arkal/Transgene.

[3] http://tools.iedb.org/mhci/.

through version 2.3.2 of the ProTECT pipeline (freely available Docker version at https://quay.io/repository/ucsc_cgl/protect) on Microsoft Azure standard_G5 (32 CPUs, 448GB RAM, 6TB disk) or standard_D15_v2 (20 CPUS, 140GB RAM, 1TB disk).

Recombinant Protein Expression and Purification

HLA-B*15:01 and HLA-A*01:01 genes containing a BirA tag were cloned into pET24+ plasmids and provided to us by the NIH Tetramer Core facility. For all *in vitro* experiments and the preparation of purified molecules for X-ray crystallography, we used soluble versions of the MHC heavy chain that lacks the BirA tag. Site directed mutagenesis to remove the BirA tag was performed using a QuikChange Lightning Multi-Site Kit (Agilent #210515) following the manufacturer's instructions. Resulting DNAs encoding HLA-B*15:01 (heavy chain), HLA-A*01:01 (heavy chain), and human β_2M (light chain) were transformed into *E. coli* BL21-DE3 (Novagen), expressed as inclusion bodies, and refolded using previously described methods (22). Briefly, *E. coli* growths with autoinduction (71) were pelleted by centrifugation and resuspended with 25 mL BugBuster (MilliporeSigma #70584) per liters of culture. Cell lysate was sonicated and subsequently pelleted by centrifugation (5,180 × *g* for 20 min at 4°C) to collect inclusion bodies. Inclusion bodies were resuspended with 25 mL of wash buffer (100 mM Tris pH 8, 2 mM EDTA, and 0.01% v/v deoxycholate), sonicated, and centrifuged again. Inclusion bodies were further resuspended in 25 mL of TE buffer (100 mM Tris pH 8, 2 mM EDTA) sonicated, and centrifuged. Following this, inclusion bodies are solubilized with 11 mL of resuspension buffer (100 mM Tris pH 8, 2 mM EDTA, 0.1 mM DTT, and 6 M guanidine–HCl). Solubilized inclusion bodies of heavy chain and light chain were mixed in a 1:3 M ratio and then added dropwise over 2 days to 1 L of refolding buffer (100 mM Tris pH 8, 2 mM EDTA, 0.4 M arginine HCl, 4.9 mM L-glutathione reduced, and 0.57 mM L-glutathione oxidized) containing 10 mg of synthetic peptide (Biopeptik). Refolding was performed for 4 days at 4°C without stirring then the sample was exhaustively dialyzed into SEC buffer (25 mM Tris pH 8 and 150 mM NaCl). Following this, the sample was concentrated with Labscale TFF system to 100 mL and further concentrated to a final volume of 5 mL using an Amicon Ultra-15 Centrifugal 10 kDa cutoff Filter Unit (Millapore Sigma). Purification was performed using SEC on a HiLoad 16/600 Superdex 75 pg with running buffer of 25 mM Tris pH 8 and 150 mM NaCl, followed by anion exchange chromatography using a mono Q 5/50 GL column and a 0–100% gradient of buffer A (25 mM Tris pH 8 and 50 mM NaCl) and buffer B (25 mM Tris pH 8 and 1 M NaCl). Finally, the purified protein was exhaustively buffer exchanged into 20 mM sodium phosphate pH 7.2 and 50 mM NaCl. The final sample was validated using LC–MS on an LTQ-Orbitrap Velos Pro MS instrument to confirm the presence of bound peptide.

MHC Tetramerization

HLA-B*15:01 containing BirA tag was refolded together with either synthetically produced AQDIYRASY or AQDIYRASYY peptide (Biopeptik) and purified following methods described earlier. Purified protein was concentrated to 0.5 mg/mL and 500 µg was biotinylated using a BirA biotin-protein ligase bulk reaction kit (Avidity Cat no. bulk BirA) following the manufacturer's instructions. An SDS-PAGE gel shift assay was performed to confirm the efficiency of the biotinylation reaction according to previously published protocols (72). The biotinylated protein sample was concentrated to 200 µL and split into two approximately 200 µg aliquots. For streptavidin–PE tetramers, 31.8 µL of 1 mg/mL of streptavidin-R-phycoerythrin (Prozyme cat no. PJRS25) was added 10 times in intervals of 10 min. For streptavidin–APC tetramers, 17.1 µL of 1 mg/mL streptavidin–allophycocyanin (Prozyme cat no. PJ27S) was added 10 times in intervals of 10 min. The final tetramer samples were stored at 4°C.

Protein Crystallization

Purified HLA-B*15:01/AQDIYRASY, HLA-B*15:01/AQDIYRASYY, and HLA-A*01:01/AQDIYRASYY complexes lacking a BirA biotinylation tag were used for crystallization. Proteins were concentrated to 10–12 mg/mL in 50 mM NaCl, 25 mM Tris pH 8.0, and crystal trays were set up using 1:1 protein-to-buffer ratio at room temperature. For HLA-B*15:01/AQDIYRASY, small crystals appeared in initial screening using molecular dimensions JCSG-plus screen after 3 days in 100 mM HEPES pH 6.5 and 20% PEG 6000 and they were further optimized. Diffraction quality crystals were harvested and incubated from above conditions plus Al's oil as a cryoprotectant and flash-frozen in liquid nitrogen before data collection. Diamond shaped diffraction quality crystals of HLA-B15:01/AQDIYRASYY were grown in crystallization buffer containing 100 mM HEPES, 2 M ammonium sulfate, and 2–4% PEG 400. Diffraction quality crystals of HLA-A*01:01/AQDIYRASYY were grown in 0.18 M magnesium chloride, 0.09 M sodium HEPES pH 7.5, 27% (v/v) PEG400, and 10% (v/v) glycerol. Crystals were flash-frozen in liquid nitrogen in a buffer containing the crystallization condition supplemented with 25% glycerol. All crystals used in this study were grown using the hanging drop vapor diffusion method. Data were collected from single crystals under cryogenic condition at Advanced Light Source (beam lines 8.3.1 and 5.0.1). Diffraction images were indexed, integrated, and scaled using Mosflm and Scala in the CCP4 package (73). Structures were determined by Phaser (74) using a previous structure of HLA-B*15:01 (PDB ID 1XR8) (25) and HLA-A*01:01 (PDB ID 1W72) (75) as search models. Model building and refinement were performed using COOT (76) and Phenix (77), respectively.

Differential Scanning Fluorimetry

All DSF experiments were performed using an Applied Biosystems ViiA qPCR machine with excitation and emission wavelengths at 470 nm and 569 nm respectively, according to previously described protocols (23). Each sample was run in triplicates of 50 µL total volume using a 96 well-plate format. Proteins were buffer exchanged into the assay buffer which was 20 mM sodium phosphate at pH 7.2 and 50 mM NaCl. Individual wells contained a final concentration of 7 µM of the respective proteins and 10× SYPRO orange dye (ThermoFisher). To determine thermal stability of each sample, the temperature incrementally increased at a scan rate of 1°C/min from 25 to 95°C. Data analysis was performed using GraphPad Prism. Melting temperatures (T_m) were determined by fitting the melting curves to a Boltzmann sigmoidal fit.

Modeling MHC Molecules and Extracting Peptide/MHC-Binding Energies

The solved X-ray structure of HLA-B*15:01/AQDIYRASY complex was used to generate a structural model for HLA-B*15:01/ARDIYRASY using single-point mutagenesis in Pymol (The PyMOL Molecular Graphics System, Version 1.8 Schrödinger, LLC.) All Dunbrack rotamers (78) for the Arg2 side chain were considered manually, and the rotamer giving the lowest strain was used in our final structural model in Figure S6A in Supplementary Material. Peptide/MHC-binding energies were computed using the *Rosetta* software suite.[4] Average binding energies of residue-specific interactions were calculated using the `residue_energy_breakdown` protocol in *Rosetta*.

To assess the ability of our *ALK* neoepitopes to bind to other HLA alleles, we performed homology-based structure simulations and computed p/MHC-binding energies *in silico*. An outline of our method is presented in **Figure 4**. Three-dimensional structural modeling and computation of p/MHC-binding affinities were performed using the *Rosetta* software suite (see text footnote 4). To carry out the modeling of homologous HLA alleles, we used *RosettaCM* protocol (50). The process of modeling high-resolution protein structures using *RosettaCM* primarily requires, that is, the sequence of the homolog is aligned with the sequence of a related known structure. It is subsequently followed by the generation of predicted 3D structures using restraints guided by a Monte Carlo sampling strategy. After performing the structure simulations of HLA alleles using our HLA-B*15:01 X-ray structure as a template, we carried out local refinement of the peptide and the MHC-binding groove. We kept backbone atoms fixed while allowing for conformational freedom of side-chain residues. The MHC-binding groove was defined by the HLA residues that were within 3.5 Å of the peptide. Local structure refinement allowed minimization of steric clashes introduced by the *RosettaCM* protocol. In addition, we refined only the peptide and the MHC-binding groove of the models to avoid noise that the full-atom refinement might introduce while trying to minimize the energy landscape at other regions and hence, making it difficult to extract accurate p/MHC-binding energies. At the local refinement stage, we generated a pool of refined structures from which we sampled low binding energy (or high-binding affinity) structures. Average binding energy was evaluated using the *Rosetta* energy function *talaris2014* (79, 80). The computation of binding energies was performed in the following steps: (1) we trimmed the MHC PDB file to remove the β_2m and α3 domains, such that only the α1/α2 domains that form the peptide-binding groove were retained. (2) We performed local refinement of the MHC-binding groove and the peptide using *Rosetta*'s relax protocol (81), which allows the region of focus to be in the local optimum of the *Rosetta* force field. Using the relax protocol, we obtained a pool of 100 locally refined models. (3) We computed the binding energies of the relaxed models using the *InterfaceAnalyzer* protocol (79, 82) by separating the MHC and the peptide energy contributions and subtracting them from the energy of the bound p/MHC (30). (4) We then selected the lowest 10 binding energy models and report their average binding energies. The sequence identity score was computed using the BLOSUM62 matrix (33) because most of the HLA alleles (68%) showed up to 62% sequence similarity. To perform the simulation, we obtained the HLA sequences from European Bioinformatics Institute's IPD-IMGT/HLA Database (29). We used ClustalOmega (83) to perform multiple sequence alignment of the HLA alleles before converting the alignment to *Rosetta*'s internal alignment format for homology modeling. Kullback–Leibler sequence logos were generated as previously described (84). Rosetta simulations were performed at the UCSC Baker cluster using 13 compute nodes with 32 cores per compute node (AMD Opteron(tm), 2.4 GHz Processor 6378). The total time used to model 2,904 HLA sequences was approximately 20,000 core hours.

[4]https://www.rosettacommons.org.

PBMC Staining

206 Cryopreserved PBMCs (CTL) from two healthy independent non-pooled HLA-B*15:01 donors were thawed and rested in phenol red free RPMI-1640 media supplemented with 10% FBS, 1% L-glut, and 1% Pen/Strep at 37°C for at least 1 h. Four independent PBMC staining experiments were run for each donor. 36 PBMCs were used for the nonamer/HLA-B*15:01, decamer/HLA-B*15:01, and SARS/HLA-B*15:01 double staining experiment. 11^6 PBMCs were used for the decamer/HLA-B*15:01-APC and nonamer/HLA-B*15:01-PE experiment. After the resting period, cells were washed with 1× PBS, followed by staining with 4 µL of each tetramer, 5% CO_2 for 10 min. An aqua amine-reactive dye (Invitrogen # L34957) was added for 10 min to assess cell viability, followed by the addition of an antibody cocktail (CD14, CD19, CD4, CD8) to stain for surface markers for an additional 20 min. The cells were washed with FACS buffer (PBS containing 0.1% sodium azide and 1% BSA) and sorted using an Aria C Flow Cytometer. Analysis of percentage of reactive CD8+ T cells was performed following gating on forward/side scattering for live lymphocytes (FSC+/SSC−), gating on Qdot− for live cells and gating on CD4−/CD8+ T cells.

DATA AVAILABILITY

The refined coordinates and structure factors for the X-ray structures of HLA-B*15:01/AQDIYRASY, HLA-B*15:01/AQDIYRASYY, and HLA-A*01:01/AQDIYRASYY complexes have been deposited in the Protein Data Bank (www.rcsb.org) with PDB IDs 5TXS, 5VZ5, and 6AT9, respectively. The ProTECT pipeline is available for use under the Apache License v2.0 for academic users (https://github.com/BD2KGenomics/protect).

ETHICS STATEMENT

All patients provided informed consent for analysis of health donor PBMCs according to specifications of the Children's Hospital of Philadelphia.

AUTHOR CONTRIBUTIONS

SS, JM, DH, and NS conceptualized and designed the research. AR and AAM performed ProTECT analysis of sequencing data. SN and NS performed *Rosetta* comparative modeling simulations and binding energy calculations. JT, ACM, and KY prepared

recombinant HLA samples and acquired LC–MS and DSF data. JT, ACM, and ST analyzed and interpreted X-ray crystallography data. MY, SN, ACM, and JT collected and analyzed PBMC tetramer staining data.

ACKNOWLEDGMENTS

The authors acknowledge Sarah Overall and David Margulies for helpful discussions as well as the NIH Tetramer Core Facility at Emory University for providing the initial HLA constructs. Additionally, we would like to acknowledge Noam Teyssier for some of the very early work in this project.

FUNDING

This research was supported by a K-22 Career development Award through NIAID (AI2573-01) and NIH (R35GM125034), in addition to a R35 Outstanding Investigator Award to NS through NIGMS (1R35GM125034-01), BD2K National Human Genome Research Institute of the National Institutes of Health (5U54HG007990), UCOP (UCSF) California Precision Medicine Initiative (OPR014149), Alex's Lemonade Stand Foundation for Childhood Cancer Innovation, and St. Baldrick's Foundation Consortium Research (OPR01419) grants to DH and SS, by a generous gift from Stephen R. and Catherine A. Shender and by the Office of the Director, NIH, under High End Instrumentation (HIE) Grant S10OD018455. DH is an investigator of the Howard Hughes Medical Institute. Diffraction data on the crystals were collected at beamlines 8.3.1 and 5.0.1 of the Advanced Light Source, which is supported by the Director, Office of Science, Office of Basic Energy Sciences, and U.S. Department of Energy, under contract DE-AC02-05CH11231. Beamline 8.3.1 at Advanced Light Source is operated by the University of California Office of the President, Multicampus Research Programs and Initiatives grant MR-15-328599 and Program for Breakthrough Biomedical Research, which is partially funded by the Sandler Foundation. Additional support was provided by a Stand Up To Cancer St. Baldrick's Pediatric Dream Team Translational Research Grant (SU2C-AACR-DT1113) to JM. Stand Up To Cancer is a program of the Entertainment Industry Foundation administered by the American Association for Cancer Research. KY was supported by NSF/DBI Award 1659649/REU Site: A Cyberlinked Program in Computational Biomolecular Structure & Design to Jeffrey J. Gray (Johns Hopkins University).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at http://www.frontiersin.org/articles/10.3389/fimmu.2018.00099/full#supplementary-material.

REFERENCES

1. Lu Y-C, Robbins PF. Cancer immunotherapy targeting neoantigens. *Semin Immunol* (2016) 28:22–7. doi:10.1016/j.smim.2015.11.002
2. Germain RN, Margulies DH. The biochemistry and cell biology of antigen processing and presentation. *Annu Rev Immunol* (1993) 11:403–50. doi:10.1146/annurev.iy.11.040193.002155
3. Blum JS, Wearsch PA, Cresswell P. Pathways of antigen processing. *Annu Rev Immunol* (2013) 31:443–73. doi:10.1146/annurev-immunol-032712-095910
4. Neefjes J, Jongsma MLM, Paul P, Bakke O. Towards a systems understanding of MHC class I and MHC class II antigen presentation. *Nat Rev Immunol* (2011) 11:823–36. doi:10.1038/nri3084
5. Schumacher TN, Hacohen N. Neoantigens encoded in the cancer genome. *Curr Opin Immunol* (2016) 41:98–103. doi:10.1016/j.coi.2016.07.005
6. Engels B, Engelhard VH, Sidney J, Sette A, Binder DC, Liu RB, et al. Relapse or eradication of cancer is predicted by peptide-major histocompatibility complex affinity. *Cancer Cell* (2013) 23:516–26. doi:10.1016/j.ccr.2013.03.018
7. Rosenberg SA, Restifo NP. Adoptive cell transfer as personalized immunotherapy for human cancer. *Science* (2015) 348:62–8. doi:10.1126/science.aaa4967
8. Oates J, Hassan NJ, Jakobsen BK. ImmTACs for targeted cancer therapy: why, what, how, and which. *Mol Immunol* (2015) 67:67–74. doi:10.1016/j.molimm.2015.01.024
9. Lim WA, June CH. The principles of engineering immune cells to treat cancer. *Cell* (2017) 168:724–40. doi:10.1016/j.cell.2017.01.016
10. Cheung N-KV, Zhang J, Lu C, Parker M, Bahrami A, Tickoo SK, et al. Association of age at diagnosis and genetic mutations in patients with neuroblastoma. *JAMA* (2012) 307:1062–71. doi:10.1001/jama.2012.228
11. Maris JM. Recent advances in neuroblastoma. *N Engl J Med* (2010) 362:2202–11. doi:10.1056/NEJMra0804577
12. Pugh TJ, Morozova O, Attiyeh EF, Asgharzadeh S, Wei JS, Auclair D, et al. The genetic landscape of high-risk neuroblastoma. *Nat Genet* (2013) 45:279–84. doi:10.1038/ng.2529
13. Sausen M, Leary RJ, Jones S, Wu J, Reynolds CP, Liu X, et al. Integrated genomic analyses identify ARID1A and ARID1B alterations in the childhood cancer neuroblastoma. *Nat Genet* (2013) 45:12–7. doi:10.1038/ng.2493
14. Molenaar JJ, Koster J, Zwijnenburg DA, van Sluis P, Valentijn LJ, van der Ploeg I, et al. Sequencing of neuroblastoma identifies chromothripsis and defects in neuritogenesis genes. *Nature* (2012) 483:589–93. doi:10.1038/nature10910
15. Schleiermacher G, Javanmardi N, Bernard V, Leroy Q, Cappo J, Rio Frio T, et al. Emergence of new ALK mutations at relapse of neuroblastoma. *J Clin Oncol* (2014) 32:2727–34. doi:10.1200/JCO.2013.54.0674
16. Eleveld TF, Oldridge DA, Bernard V, Koster J, Daage LC, Diskin SJ, et al. Relapsed neuroblastomas show frequent RAS-MAPK pathway mutations. *Nat Genet* (2015) 47:864–71. doi:10.1038/ng.3333
17. George RE, Sanda T, Hanna M, Fröhling S, Ii WL, Zhang J, et al. Activating mutations in ALK provide a therapeutic target in neuroblastoma. *Nature* (2008) 455:975–8. doi:10.1038/nature07397
18. Walker AJ, Majzner RG, Zhang L, Wanhainen K, Long AH, Nguyen SM, et al. Tumor antigen and receptor densities regulate efficacy of a chimeric antigen receptor targeting anaplastic lymphoma kinase. *Mol Ther* (2017) 25:2189–201. doi:10.1016/j.ymthe.2017.06.008
19. Lundegaard C, Lamberth K, Harndahl M, Buus S, Lund O, Nielsen M. 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* (2008) 36:W509–12. doi:10.1093/nar/gkn202
20. Vita R, Overton JA, Greenbaum JA, Ponomarenko J, Clark JD, Cantrell JR, et al. The immune epitope database (IEDB) 3.0. *Nucleic Acids Res* (2015) 43:D405–12. doi:10.1093/nar/gku938
21. Linard B, Bézieau S, Benlalam H, Labarrière N, Guilloux Y, Diez E, et al. A ras-mutated peptide targeted by CTL infiltrating a human melanoma lesion. *J Immunol* (2002) 168:4802–8. doi:10.4049/jimmunol.168.9.4802
22. Garboczi DN, Hung DT, Wiley DC. HLA-A2-peptide complexes: refolding and crystallization of molecules expressed in *Escherichia coli* and complexed with single antigenic peptides. *Proc Natl Acad Sci U S A* (1992) 89:3429–33. doi:10.1073/pnas.89.8.3429
23. Hellman LM, Yin L, Wang Y, Blevins SJ, Riley TP, Belden OS, et al. Differential scanning fluorimetry based assessments of the thermal and kinetic stability of peptide-MHC complexes. *J Immunol Methods* (2016) 432:95–101. doi:10.1016/j.jim.2016.02.016
24. Motozono C, Pearson JA, De Leenheer E, Rizkallah PJ, Beck K, Trimby A, et al. Distortion of the major histocompatibility complex class i binding groove

to accommodate an insulin-derived 10-Mer peptide. *J Biol Chem* (2015) 290: 18924–33. doi:10.1074/jbc.M114.622522
25. Røder G, Blicher T, Justesen S, Johannesen B, Kristensen O, Kastrup J, et al. Crystal structures of two peptide-HLA-B*1501 complexes; structural characterization of the HLA-B62 supertype. *Acta Crystallogr D Biol Crystallogr* (2006) 62:1300–10. doi:10.1107/S0907444906027636
26. Røder G, Kristensen O, Kastrup JS, Buus S, Gajhede M. Structure of a SARS coronavirus-derived peptide bound to the human major histocompatibility complex class I molecule HLA-B*1501. *Acta Crystallogr Sect F Struct Biol Cryst Commun* (2008) 64:459–62. doi:10.1107/S1744309108012396
27. Remesh SG, Andreatta M, Ying G, Kaever T, Nielsen M, McMurtrey C, et al. Unconventional peptide presentation by major histocompatibility complex (MHC) class I Allele HLA-A*02:01: BREAKING CONFINEMENT. *J Biol Chem* (2017) 292:5262–70. doi:10.1074/jbc.M117.776542
28. Leaver-Fay A, O'Meara MJ, Tyka M, Jacak R, Song Y, Kellogg EH, et al. Scientific benchmarks for guiding macromolecular energy function improvement. *Methods Enzymol* (2013) 523:109–43. doi:10.1016/B978-0-12-394292-0.00006-0
29. Maccari G, Robinson J, Ballingall K, Guethlein LA, Grimholt U, Kaufman J, et al. IPD-MHC 2.0: an improved inter-species database for the study of the major histocompatibility complex. *Nucleic Acids Res* (2017) 45:D860–4. doi:10.1093/nar/gkw1050
30. Yanover C, Bradley P. Large-scale characterization of peptide-MHC binding landscapes with structural simulations. *Proc Natl Acad Sci U S A* (2011) 108:6981–6. doi:10.1073/pnas.1018165108
31. Liu T, Pan X, Chao L, Tan W, Qu S, Yang L, et al. Subangstrom accuracy in pHLA-I modeling by Rosetta FlexPepDock refinement protocol. *J Chem Inf Model* (2014) 54:2233–42. doi:10.1021/ci500393h
32. London N, Raveh B, Cohen E, Fathi G, Schueler-Furman O. Rosetta FlexPepDock web server—high resolution modeling of peptide–protein interactions. *Nucleic Acids Res* (2011) 39:W249–53. doi:10.1093/nar/gkr431
33. Henikoff S, Henikoff JG. Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci U S A* (1992) 89:10915–9. doi:10.1073/pnas.89.22.10915
34. Maiers M, Gragert L, Klitz W. High-resolution HLA alleles and haplotypes in the United States population. *Hum Immunol* (2007) 68:779–88. doi:10.1016/j.humimm.2007.04.005
35. Fahnestock ML, Johnson JL, Feldman RM, Tsomides TJ, Mayer J, Narhi LO, et al. Effects of peptide length and composition on binding to an empty class I MHC heterodimer. *Biochemistry (Mosc)* (1994) 33:8149–58. doi:10.1021/bi00192a020
36. Rosenberg SA, Yang JC, Sherry RM, Kammula US, Hughes MS, Phan GQ, et al. Durable complete responses in heavily pretreated patients with metastatic melanoma using T-cell transfer immunotherapy. *Clin Cancer Res* (2011) 17:4550–7. doi:10.1158/1078-0432.CCR-11-0116
37. Mackall CL, Merchant MS, Fry TJ. Immune-based therapies for childhood cancer. *Nat Rev Clin Oncol* (2014) 11:693–703. doi:10.1038/nrclinonc.2014.177
38. Platten M, Offringa R. Cancer immunotherapy: exploiting neoepitopes. *Cell Res* (2015) 25:887–8. doi:10.1038/cr.2015.66
39. Hunt DF, Henderson RA, Shabanowitz J, Sakaguchi K, Michel H, Sevilir N, et al. Characterization of peptides bound to the class I MHC molecule HLA-A2.1 by mass spectrometry. *Science* (1992) 255:1261–3. doi:10.1126/science.1546328
40. Buchli R, VanGundy RS, Hickman-Miller HD, Giberson CF, Bardet W, Hildebrand WH. Real-time measurement of in vitro peptide binding to soluble HLA-A*0201 by fluorescence polarization. *Biochemistry (Mosc)* (2004) 43:14852–63. doi:10.1021/bi048580q
41. Abelin JG, Keskin DB, Sarkizova S, Hartigan CR, Zhang W, Sidney J, et al. Mass spectrometry profiling of HLA-associated peptidomes in mono-allelic cells enables more accurate epitope prediction. *Immunity* (2017) 46:315–26. doi:10.1016/j.immuni.2017.02.007
42. Paul S, Weiskopf D, Angelo MA, Sidney J, Peters B, Sette A. HLA class I alleles are associated with peptide binding repertoires of different size, affinity and immunogenicity. *J Immunol* (2013) 191:5831–9. doi:10.4049/jimmunol.1302101
43. Rudolph MG, Stanfield RL, Wilson IA. How TCRs bind MHCs, peptides, and coreceptors. *Annu Rev Immunol* (2006) 24:419–66. doi:10.1146/annurev.immunol.23.021704.115658
44. Ooi JD, Petersen J, Tan YH, Huynh M, Willett ZJ, Ramarathinam SH, et al. Dominant protection from HLA-linked autoimmunity by antigen-specific regulatory T cells. *Nature* (2017) 545:243–7. doi:10.1038/nature22329
45. Jiang J, Natarajan K, Boyd LF, Morozov GI, Mage MG, Margulies DH. Crystal structure of a TAPBPR–MHC-I complex reveals the mechanism of peptide editing in antigen presentation. *Science* (2017) 358(6366):1064–8. doi:10.1126/science.aao5154
46. Smith KJ, Reid SW, Stuart DI, McMichael AJ, Jones EY, Bell JI. An altered position of the alpha 2 helix of MHC class I is revealed by the crystal structure of HLA-B*3501. *Immunity* (1996) 4:203–13. doi:10.1016/S1074-7613(00)80429-X
47. Baker BM, Turner RV, Gagnon SJ, Wiley DC, Biddison WE. Identification of a crucial energetic footprint on the α1 helix of human histocompatibility leukocyte antigen (Hla)-A2 that provides functional interactions for recognition by tax peptide/Hla-A2–specific T cell receptors. *J Exp Med* (2001) 193:551–62. doi:10.1084/jem.193.5.551
48. Yu Z, Theoret MR, Touloukian CE, Surman DR, Garman SC, Feigenbaum L, et al. Poor immunogenicity of a self/tumor antigen derives from peptide-MHC-I instability and is independent of tolerance. *J Clin Invest* (2004) 114: 551–9. doi:10.1172/JCI21695
49. Wieczorek M, Abualrous ET, Sticht J, Álvaro-Benito M, Stolzenberg S, Noé F, et al. Major histocompatibility complex (MHC) class I and MHC class II proteins: conformational plasticity in antigen presentation. *Front Immunol* (2017) 8:292. doi:10.3389/fimmu.2017.00292
50. Song Y, DiMaio F, Wang RY-R, Kim D, Miles C, Brunette T, et al. High-resolution comparative modeling with RosettaCM. *Structure* (2013) 21:1735–42. doi:10.1016/j.str.2013.08.005
51. Huang P-S, Ban Y-EA, Richter F, Andre I, Vernon R, Schief WR, et al. RosettaRemodel: a generalized framework for flexible backbone protein design. *PLoS One* (2011) 6:e24109. doi:10.1371/journal.pone.0024109
52. Bacher P, Scheffold A. Flow-cytometric analysis of rare antigen-specific T cells. *Cytometry A* (2013) 83A:692–701. doi:10.1002/cyto.a.22317
53. Jenkins MK, Moon JJ. The role of naïve T cell precursor frequency and recruitment in dictating immune response magnitude. *J Immunol* (2012) 188:4135–40. doi:10.4049/jimmunol.1102661
54. Frøsig TM, Yap J, Seremet T, Lyngaa R, Svane IM, Thor Straten P, et al. Design and validation of conditional ligands for HLA-B*08:01, HLA-B*15:01, HLA-B*35:01, and HLA-B*44:05. *Cytometry A* (2015) 87:967–75. doi:10.1002/cyto.a.22689
55. Birnbaum ME, Mendoza JL, Sethi DK, Dong S, Glanville J, Dobbins J, et al. Deconstructing the peptide-MHC specificity of T cell recognition. *Cell* (2014) 157:1073–87. doi:10.1016/j.cell.2014.03.047
56. Miles JJ, McCluskey J, Rossjohn J, Gras S. Understanding the complexity and malleability of T-cell recognition. *Immunol Cell Biol* (2015) 93:433–41. doi:10.1038/icb.2014.112
57. Slansky JE, Rattis FM, Boyd LF, Fahmy T, Jaffee EM, Schneck JP, et al. Enhanced antigen-specific antitumor immunity with altered peptide ligands that stabilize the MHC-peptide-TCR complex. *Immunity* (2000) 13:529–38. doi:10.1016/S1074-7613(00)00052-2
58. Manzo T, Sturmheit T, Basso V, Petrozziello E, Hess Michelini R, Riba M, et al. T cells redirected to a minor histocompatibility antigen instruct intratumoral TNFα expression and empower adoptive cell therapy for solid tumors. *Cancer Res* (2017) 77:658–71. doi:10.1158/0008-5472.CAN-16-0725
59. Carreno BM, Magrini V, Becker-Hapak M, Kaabinejadian S, Hundal J, Petti AA, et al. Cancer immunotherapy. A dendritic cell vaccine increases the breadth and diversity of melanoma neoantigen-specific T cells. *Science* (2015) 348: 803–8. doi:10.1126/science.aaa3828
60. Ott PA, Hu Z, Keskin DB, Shukla SA, Sun J, Bozym DJ, et al. An immunogenic personal neoantigen vaccine for patients with melanoma. *Nature* (2017) 547:217–21. doi:10.1038/nature22991
61. Sahin U, Derhovanessian E, Miller M, Kloke B-P, Simon P, Löwer M, et al. Personalized RNA mutanome vaccines mobilize poly-specific therapeutic immunity against cancer. *Nature* (2017) 547:222–6. doi:10.1038/nature23003
62. Bai Y, Ni M, Cooper B, Wei Y, Fury W. Inference of high resolution HLA types using genome-wide RNA or DNA sequencing reads. *BMC Genomics* (2014) 15:325. doi:10.1186/1471-2164-15-325
63. Cibulskis K, Lawrence MS, Carter SL, Sivachenko A, Jaffe D, Sougnez C, et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat Biotechnol* (2013) 31:213–9. doi:10.1038/nbt.2514

64. Fan Y, Xi L, Hughes DST, Zhang J, Zhang J, Futreal PA, et al. MuSE: accounting for tumor heterogeneity using a sample-specific error model improves sensitivity and specificity in mutation calling from sequencing data. *Genome Biol* (2016) 17:178. doi:10.1186/s13059-016-1029-6
65. Radenbaugh AJ, Ma S, Ewing A, Stuart JM, Collisson EA, Zhu J, et al. RADIA: RNA and DNA integrated analysis for somatic mutation detection. *PLoS One* (2014) 9:e111516. doi:10.1371/journal.pone.0111516
66. Larson DE, Harris CC, Chen K, Koboldt DC, Abbott TE, Dooling DJ, et al. SomaticSniper: identification of somatic point mutations in whole genome sequencing data. *Bioinformatics* (2012) 28:311–7. doi:10.1093/bioinformatics/btr665
67. Saunders CT, Wong WSW, Swamy S, Becq J, Murray LJ, Cheetham RK. Strelka: accurate somatic small-variant calling from sequenced tumor-normal sample pairs. *Bioinformatics* (2012) 28:1811–7. doi:10.1093/bioinformatics/bts271
68. Orren A, Potter PC, Cooper RC, du Toit E. Deficiency of the sixth component of complement and susceptibility to *Neisseria meningitidis* infections: studies in 10 families and five isolated cases. *Immunology* (1987) 62:249–53.
69. Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, et al. GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res* (2012) 22:1760–74. doi:10.1101/gr.135350.111
70. Costello M, Pugh TJ, Fennell TJ, Stewart C, Lichtenstein L, Meldrim JC, et al. Discovery and characterization of artifactual mutations in deep coverage targeted capture sequencing data due to oxidative DNA damage during sample preparation. *Nucleic Acids Res* (2013) 41:e67. doi:10.1093/nar/gks1443
71. Studier FW. Stable expression clones and auto-induction for protein production in *E. coli*. *Methods Mol Biol* (2014) 1091:17–32. doi:10.1007/978-1-62703-691-7_2
72. Fairhead M, Howarth M. Site-specific biotinylation of purified proteins using BirA. *Methods Mol Biol Clifton NJ* (2015) 1266:171–84. doi:10.1007/978-1-4939-2272-7_12
73. Winn MD, Ballard CC, Cowtan KD, Dodson EJ, Emsley P, Evans PR, et al. Overview of the CCP4 suite and current developments. *Acta Crystallogr D Biol Crystallogr* (2011) 67:235–42. doi:10.1107/S0907444910045749
74. McCoy AJ, Grosse-Kunstleve RW, Adams PD, Winn MD, Storoni LC, Read RJ. Phaser crystallographic software. *J Appl Crystallogr* (2007) 40:658–74. doi:10.1107/S0021889807021206
75. Hülsmeyer M, Chames P, Hillig RC, Stanfield RL, Held G, Coulie PG, et al. A major histocompatibility complex-peptide-restricted antibody and t cell receptor molecules recognize their target by distinct binding modes: crystal structure of human leukocyte antigen (HLA)-A1-MAGE-A1 in complex with FAB-HYB3. *J Biol Chem* (2005) 280:2972–80. doi:10.1074/jbc.M411323200
76. Emsley P, Cowtan K. Coot: model-building tools for molecular graphics. *Acta Crystallogr D Biol Crystallogr* (2004) 60:2126–32. doi:10.1107/S0907444904019158
77. Adams PD, Afonine PV, Bunkóczi G, Chen VB, Davis IW, Echols N, et al. PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr D Biol Crystallogr* (2010) 66:213–21. doi:10.1107/S0907444909052925
78. Shapovalov MV, Dunbrack RL. A smoothed backbone-dependent rotamer library for proteins derived from adaptive kernel density estimates and regressions. *Structure* (2011) 19:844–58. doi:10.1016/j.str.2011.03.019
79. Bender BJ, Cisneros A, Duran AM, Finn JA, Fu D, Lokits AD, et al. Protocols for molecular modeling with Rosetta3 and RosettaScripts. *Biochemistry (Mosc)* (2016) 55:4748–63. doi:10.1021/acs.biochem.6b00444
80. Alford RF, Leaver-Fay A, Jeliazkov JR, O'Meara MJ, DiMaio FP, Park H, et al. The Rosetta all-atom energy function for macromolecular modeling and design. *J Chem Theory Comput* (2017) 13:3031–48. doi:10.1021/acs.jctc.7b00125
81. Khatib F, Cooper S, Tyka MD, Xu K, Makedon I, Popović Z, et al. Algorithm discovery by protein folding game players. *Proc Natl Acad Sci U S A* (2011) 108:18949–53. doi:10.1073/pnas.1115898108
82. Lewis SM, Kuhlman BA. Anchored design of protein-protein interfaces. *PLoS One* (2011) 6:e20872. doi:10.1371/journal.pone.0020872
83. Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol Syst Biol* (2011) 7:539. doi:10.1038/msb.2011.75
84. Thomsen MCF, Nielsen M. Seq2Logo: a method for construction and visualization of amino acid binding motifs and sequence profiles including sequence weighting, pseudo counts and two-sided representation of amino acid enrichment and depletion. *Nucleic Acids Res* (2012) 40:W281–7. doi:10.1093/nar/gks469

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2018 Toor, Rao, McShan, Yarmarkovich, Nerli, Yamaguchi, Madejska, Nguyen, Tripathi, Maris, Salama, Haussler and Sgourakis. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

5.3 The pediatric NBL study supplementary

The supplementary data pdf contains a lot of information that was conducted by the Sgourakis and Maris Groups. The relevant printable sections of the supplement are provided below. Supplementary data sheets 1, 2, and 3 are too large to add to this thesis as pdf formatted tables. The full supplementary (including supplementary data sheets 1, 2, and 3) can be obtained from the Frontiers in Immunology website.

5.3.1 Supplementary Table 1

Supplementary Tables

Protein:Mutation	Immuno Active Region	Samples with Expressed Mutation		Therapeutic MHC-I Allele (Frequency in Population)	Counts in COSMIC
		Primary	Relapse		
ALK:R1275Q*	AKIGDFGMA**[R/Q]**DIYRASYYR	5	1	HLA-B*15:01 (0.06) HLA-A*02:01 (0.20)	101
ALK:F1174L	LMEALIISK**[F/L]**NHQNIVRCI	3	0	HLA-B*44:03 (0.02)	144
ALK:F1174I	LMEALIISK**[F/I]**NHQNIVRCI	1	0	HLA-B*13:02 (0.01)	7
ZNF717:Q716E	NVENPFIRR**[Q/E]**IFRSIKVFT	3	0	HLA-B*07:02 (0.03)	0
KCNJ12:E378K	LPSANSFCY**[E/K]**NELAFLSRD	2	2	HLA-A*23:01 (0.01)	0
ITGB2:I712T	AIVGGTVAG**[I/T]**VLIGILLLV	2	0	HLA-A*02:01 (0.02)	0
USP6:R69W	TAREAKKIR**[R/W]**EMTRTSKWM	2	0	HLA-A*24:02 (0.38) HLA-C*04:01 (0.15)	0
TSC22D1:T961K	LFPLKVLPL**[T/K]**TPLVDGEDE	2	0	HLA-C*03:03 (0.04) HLA-C*04:01 (0.15)	0
NRAS:Q61K	LLDILDTAG**[Q/K]**EEYSAMRDQ	1	1	HLA-A*01:01 (0.04)	1287

Supplementary Table 1. ProTECT analysis of 100 primary and 6 primary:relapsed pairs of TARGET NBL samples. The first column shows identified proteins (in italics) with their respective mutation observed in tumors. The second column shows the ImmunoActive Region of each protein with the format [X /Y] in bold denoting wild-type and mutant amino acids, respectively. The third column shows the number of samples with the expressed mutation in primary and relapsed patients. The fourth column shows the best predicted therapeutic MHC allele, which is defined as either an allele shared by samples expressing the same mutation or an allele with the best predicted binding score. For each allele, the frequency (denoted in parenthesis) is in Caucasian populations. The fifth column shows counts in COSMIC that refer to the sum total of unique entries in the Catalog of Somatic Mutations in Cancer database (1) for each mutation.

5.3.2 Supplementary References

Supplementary References

1. Forbes SA, Beare D, Boutselakis H, Bamford S, Bindal N, Tate J, Cole CG, Ward S, Dawson E, Ponting L, et al. COSMIC: somatic cancer genetics at high-resolution. *Nucleic Acids Res* (2017) **45**:D777–D783. doi:10.1093/nar/gkw1121

2. Lovell SC, Davis IW, Arendall WB, de Bakker PIW, Word JM, Prisant MG, Richardson JS, Richardson DC. Structure validation by Calpha geometry: phi,psi and Cbeta deviation. *Proteins* (2003) **50**:437–450. doi:10.1002/prot.10286

3. Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H, Remmert M, Söding J, et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol Syst Biol* (2011) **7**:539. doi:10.1038/msb.2011.75

4. Robert X, Gouet P. Deciphering key features in protein structures with the new ENDscript server. *Nucleic Acids Res* (2014) **42**:W320–W324. doi:10.1093/nar/gku316

5. Thomsen MCF, Nielsen M. Seq2Logo: a method for construction and visualization of amino acid binding motifs and sequence profiles including sequence weighting, pseudo counts and two-sided representation of amino acid enrichment and depletion. *Nucleic Acids Res* (2012) **40**:W281–W287. doi:10.1093/nar/gks469

6. Maccari G, Robinson J, Ballingall K, Guethlein LA, Grimholt U, Kaufman J, Ho C-S, de Groot NG, Flicek P, Bontrop RE, et al. IPD-MHC 2.0: an improved inter-species database for the study of the major histocompatibility complex. *Nucleic Acids Res* (2017) **45**:D860–D864. doi:10.1093/nar/gkw1050

7. Henikoff S, Henikoff JG. Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci U S A* (1992) **89**:10915–10919.

8. Vita R, Overton JA, Greenbaum JA, Ponomarenko J, Clark JD, Cantrell JR, Wheeler DK, Gabbard JL, Hix D, Sette A, et al. The immune epitope database (IEDB) 3.0. *Nucleic Acids Res* (2015) **43**:D405-412. doi:10.1093/nar/gku938

9. Sidney J, Assarsson E, Moore C, Ngo S, Pinilla C, Sette A, Peters B. Quantitative peptide binding motifs for 19 human and mouse MHC class I molecules derived using positional scanning combinatorial peptide libraries. *Immunome Res* (2008) **4**:2. doi:10.1186/1745-7580-4-2

10. Hoof I, Peters B, Sidney J, Pedersen LE, Sette A, Lund O, Buus S, Nielsen M. NetMHCpan, a method for MHC class I binding prediction beyond humans. *Immunogenetics* (2009) **61**:1–13. doi:10.1007/s00251-008-0341-z

Chapter 6

Conclusion

6.1 Introduction

This thesis describes my doctoral work on the prediction of therapeutically significant neoepitopes for Adoptive Cell- and Vaccine-based immunotherapies. ProTECT has a lot of potential clinical utility, especially given the current traction of the entire immuno-oncology field.

6.2 Chapters

In Chapter 1, I introduced the basics of Immunology, the workings of the adaptive immune system, and I talked about how doctors can potentially treat cancers similar to bacterial and viral infections. I described the current state of immuno-oncology and talked about the methods that are being used to treat cancers using immunological agents. I highlight the importance of T-cell based targeted immunotherapies as being less harmful to the patient due to higher specificity of the tumor as a target. I talk about the limitations of these methods in terms of

time and complexity, explaining how shortlisting targets will lead to faster therapies for patients diagnosed with cancer.

In Chapter 2, I talk about my work on a workflow engine developed at UCSC. Running bioinformatics pipelines at scale can be resource intensive and using a smart, efficient workflow engine can allow a user to run pipelines faster, with fewer interruptions. Making large workflows easy to run enables researchers to broaden their sample sets and look at their data with a broader scope. Providing support for cloud deployment further broadens the horizons of researchers, making extremely large analyses possible at affordable costs. It also supports collaboration across institutions given the increase in cloud acceptance by Institutional Review Boards.

In Chapter 3, I describe a tool I wrote at UCSC, TransGene. TransGene allows a user to translate any combination of genomic variants (SNV, INDEL, or fusion-genes) into peptides that can be used to identify neoepitopes in a tumor. TransGene was developed since there was no available tool that could perform this task, limiting the automation of a neoepitope prediction workflow. TransGene served as the backbone of the ProTECT pipeline, translating the mutants called from a consensus of callers into peptides that could be processed by the IEDB suite of pMHC prediction tools.

Chapter 4 describes the bulk of my work as a Ph.D. candidate. The ProTECT pipeline is a one-of-a-kind end-to-end workflow to process sequencing data and produce a ranked list of neoepitopes of therapeutic significance. The workflow is entirely self-contained and automated, allowing the user to run the whole pipeline without any manual intervention. This allows clinical labs without a bioinformatics team to analyze in-house data that cannot be shared outside their

institution (A common problem faced by a number of hospitals worldwide). The pipeline is entirely reproducible through the magic of Docker, so the results from running ProTECT on a remote server are guaranteed to be the same as the results I would obtain on a local machine. This gives an external institution the confidence that their data is receiving the same treatment as that given by a trained bioinformatician. I demonstrate the efficiency and automation of ProTECT on a 326 sample cohort and compare ProTECT's performance with published tools. Using ProTECT, I predict potentially broadly therapeutic neoepitopes in the PRAD cohort that can benefit a large subpopulation of PRAD patients.

Chapter 5 describes a real-world application of the ProTECT pipeline. I analyzed the data from 106 pediatric Neuroblastoma samples and identified 2 peptides from a hotspot mutant that is predicted to bind to an MHC molecule that is sufficiently common in Caucasian populations. The predicted combinations were validated as good binders using Biochemical methods and were also predicted to be recognized by $CD8^+$ T-cells. This suggests the mutation is broadly immunotherapeutically significant.

6.3 Discussion

Vaccine-based and Adoptive cell-therapies are quickly demonstrating their superiority as the future of cancer therapies in terms of durable regression of treated cancers, and due to the reduced cytotoxicity associated with these therapies as compared to traditional chemo- and radiation-based therapies. Rapid administration of these therapies to terminally diagnosed cancer patients could save many lives, however these therapies are currently time consuming. Reduction

in the time taken to go from a patient biopsy to a cure is essential for these therapies to become clinically viable as the standard-of-care.

This thesis described an open-source, portable method to predict and rank neoepitopes in a tumor sample that will be relevant for existing and future neoepitope-driven immunotherapies. The method was used to identify a potentially therapeutic epitope in a high-risk cohort of pediatric cancers with poor prognosis. When it is published, ProTECT will be only the third published tool for neoepitope discovery. As the only fully portable and automated one, I hope ProTECT will be adopted by a larger audience in the field of cancer immunotherapy. Since ProTECT works with FASTQ reads straight off a sequencer, it serves as an easy-to use utility for labs with minimal bioinformatics support.

Chapter 7

Appendix

7.1 Figures

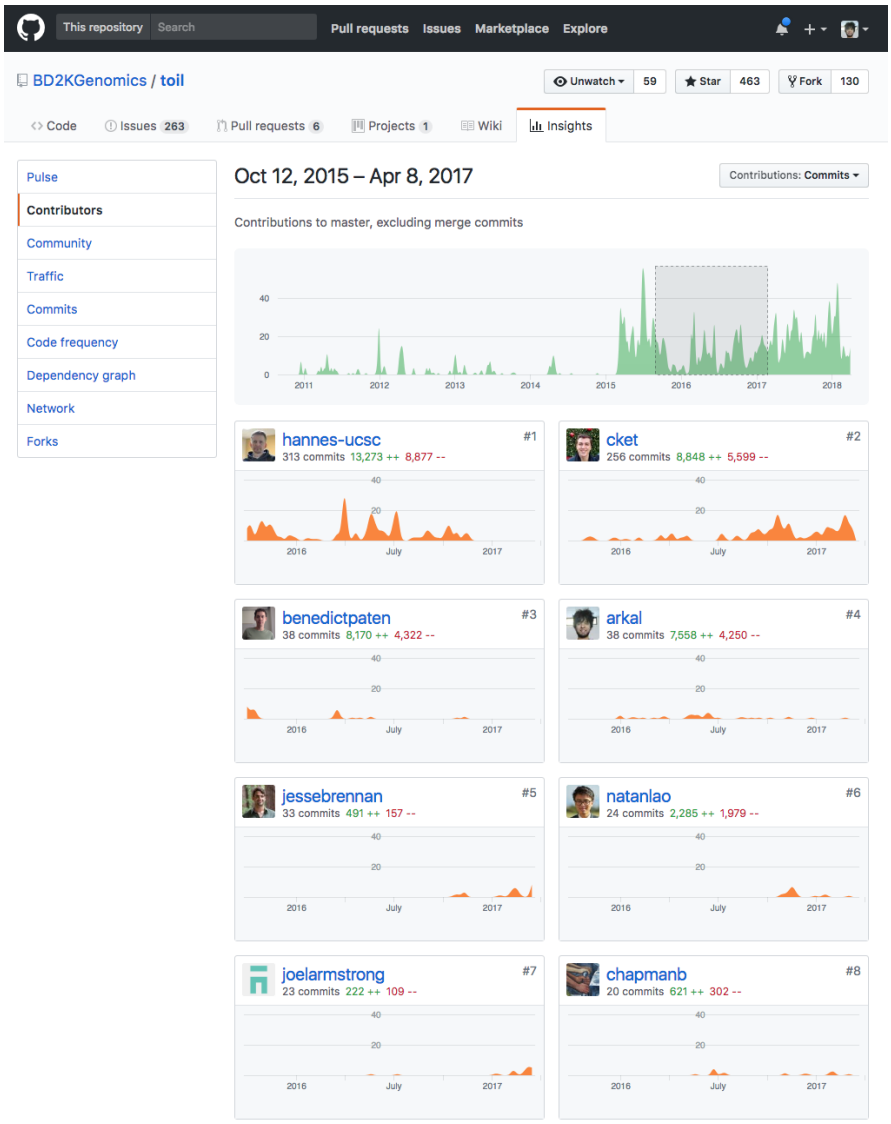


Figure 7.1: A screengrab of the 'Contributors' tab on the BD2KGenomics/toil repository from the date of my first commit to the time of publication showing my substantial contribution.

Bibliography

[1] Picard Tools - By Broad Institute.

[2] epidisco: Personalized cancer epitope discovery and peptide vaccine prediction pipeline, April 2018. original-date: 2016-08-08T19:59:58Z.

[3] FusionInspector code, May 2018. original-date: 2015-07-09T17:48:54Z.

[4] Rachel Airley. *Cancer Chemotherapy: Basic Science to the Clinic*. John Wiley & Sons, April 2009.

[5] Hashem O. Alsaab, Samaresh Sau, Rami Alzhrani, Katyayani Tatiparti, Ketki Bhise, Sushil K. Kashaw, and Arun K. Iyer. PD-1 and PD-L1 Checkpoint Signaling Inhibition for Cancer Immunotherapy: Mechanism, Combinations, and Clinical Outcome. *Frontiers in Pharmacology*, 8, August 2017.

[6] Yu Bai, Min Ni, Blerta Cooper, Yi Wei, and Wen Fury. Inference of high resolution HLA types using genome-wide RNA or DNA sequencing reads. *BMC Genomics*, 15(1):325, may 2014.

[7] Jacques Banchereau and Ralph M. Steinman. Dendritic cells and the control of immunity. *Nature*, 392(6673):245–252, March 1998.

[8] Christopher E. Barbieri, Sylvan C. Baca, Michael S. Lawrence, Francesca Demichelis, Mirjam Blattner, Jean-Philippe Theurillat, Thomas A. White, Petar Stojanov, Eliezer Van Allen, Nicolas Stransky, Elizabeth Nickerson, Sung-Suk Chae, Gunther Boysen, Daniel Auclair, Robert C. Onofrio, Kyung Park, Naoki Kitabayashi, Theresa Y. MacDonald, Karen Sheikh, Terry Vuong, Candace Guiducci, Kristian Cibulskis, Andrey Sivachenko, Scott L. Carter, Gordon Saksena, Douglas Voet, Wasay M. Hussain, Alex H. Ramos, Wendy Winckler, Michelle C. Redman, Kristin Ardlie, Ashutosh K. Tewari, Juan Miguel Mosquera, Niels Rupp, Peter J. Wild, Holger Moch, Colm Morrissey, Peter S. Nelson, Philip W. Kantoff, Stacey B. Gabriel, Todd R. Golub, Matthew Meyerson, Eric S. Lander, Gad Getz, Mark A. Rubin, and Levi A. Garraway. Exome sequencing identifies recurrent *SPOP*, *FOXA1* and *MED12* mutations in prostate cancer. *Nature Genetics*, 44(6):685–689, June 2012.

[9] Gregory L. Beatty, Andrew R. Haas, Marcela V. Maus, Drew A. Torigian, Michael C. Soulen, Gabriela Plesa, Anne Chew, Yangbing Zhao, Bruce L. Levine, Steven M. Albelda, Michael Kalos, and Carl H. June. Mesothelin-Specific Chimeric Antigen Receptor mRNA-Engineered T Cells Induce Antitumor Activity in Solid Malignancies. *Cancer Immunology Research*, 2(2):112–120, February 2014.

[10] G Berke. The binding and lysis of target cells by cytotoxic lymphocytes: Molecular and cellular aspects. *Annual Review of Immunology*, 12(1):735–773, 1994.

[11] Mirjam Blattner, Daniel J. Lee, Catherine O'Reilly, Kyung Park, Theresa Y. MacDonald, Francesca Khani, Kevin R. Turner, Ya-Lin Chiu, Peter J. Wild, Igor Dolgalev, Adriana Heguy, Andrea Sboner, Sinan Ramazangolu, Haley Hieronymus, Charles Sawyers, Ashutosh K. Tewari, Holger Moch, Ghil Suk Yoon, Yong Chul Known, Ove Andrén, Katja Fall, Francecsa Demichelis, Juan Miguel Mosquera, Brian D. Robinson, Christopher E. Barbieri, and Mark A. Rubin. SPOP Mutations in Prostate Cancer across Demographically Diverse Patient Cohorts. *Neoplasia*, 16(1):14–W10, January 2014.

[12] Vincenzo Bronte, Elisa Apolloni, Anna Cabrelle, Roberto Ronca, Paolo Serafini, Paola Zamboni, Nicholas P. Restifo, and Paola Zanovello. Identification of a CD11b+/Gr-1+/CD31+ myeloid progenitor capable of activating or suppressing CD8+ T cells. *Blood*, 96(12):3838–3846, December 2000.

[13] Vicki Brower. Checkpoint Blockade Immunotherapy for Cancer Comes of Age. *JNCI J Natl Cancer Inst*, 107(3):djv069, March 2015.

[14] Huynh-Hoa Bui, John Sidney, Bjoern Peters, Muthuraman Sathiamurthy, Asabe Sinichi, Kelly-Anne Purton, Bianca R. Mothé, Francis V. Chisari, David I. Watkins, and Alessandro Sette. Automated generation and evaluation of specific MHC binding predictive tools: ARB matrix applications. *Immunogenetics*, 57(5):304–314, June 2005.

[15] David P. Carbone, Martin Reck, Luis Paz-Ares, Benjamin Creelan, Leora Horn, Martin Steins, Enriqueta Felip, Michel M. van den Heuvel, Tudor-Eliade Ciuleanu, Firas Badin, Neal Ready, T. Jeroen N. Hiltermann, Suresh Nair, Rosalyn Juergens, Solange Peters, Elisa Minenza, John M. Wrangle, Delvys Rodriguez-Abreu, Hossein Borghaei, George R. Blumenschein, Liza C. Villaruz, Libor Havel, Jana Krejci, Jesus Corral Jaime, Han Chang, William J. Geese, Prabhu Bhagavatheeswaran, Allen C. Chen, and Mark A. Socinski. First-Line Nivolumab in Stage IV or Recurrent Non–Small-Cell Lung Cancer. *New England Journal of Medicine*, 376(25):2415–2426, June 2017.

[16] Beatriz M. Carreno, Vincent Magrini, Michelle Becker-Hapak, Saghar Kaabinejadian, Jasreet Hundal, Allegra A. Petti, Amy Ly, Wen-Rong Lie, William H. Hildebrand, Elaine R. Mardis, and Gerald P. Linette. A dendritic cell vaccine increases the breadth and diversity of melanoma neoantigen-specific T cells. *Science*, page aaa3828, April 2015.

[17] Cynthia A. Chambers, Michael S. Kuhns, Jackson G. Egen, and James P. Allison. CTLA-4-Mediated Inhibition in Regulation of T Cell Responses: Mechanisms and Manipulation in Tumor Immunotherapy. *Annual Review of Immunology*, 19(1):565–594, April 2001.

[18] Kristian Cibulskis, Michael S. Lawrence, Scott L. Carter, Andrey Sivachenko, David Jaffe, Carrie Sougnez, Stacey Gabriel, Matthew Meyerson, Eric S. Lander, and Gad Getz. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat Biotech*, 31(3):213–219, March 2013.

[19] P. Cingolani, A. Platts, M. Coon, T. Nguyen, L. Wang, S.J. Land, X. Lu, and D.M. Ruden. A program for annotating and predicting the effects of single nucleotide polymorphisms, snpeff: Snps in the genome of drosophila melanogaster strain w1118; iso-2; iso-3. *Fly*, 6(2):80–92, 2012.

[20] Pablo Cingolani, Adrian Platts, Le Lily Wang, Melissa Coon, Tung Nguyen, Luan Wang, Susan J. Land, Xiangyi Lu, and Douglas M. Ruden. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. *Fly*, 6(2):80–92, April 2012.

[21] Cyrille J. Cohen, Jared J. Gartner, Miryam Horovitz-Fried, Katerina Shamalov, Kasia Trebska-McGowan, Valery V. Bliskovsky, Maria R. Parkhurst, Chen Ankri, Todd D. Prickett, Jessica S. Crystal, Yong F. Li, Mona El-Gamil, Steven A. Rosenberg, and Paul F. Robbins. Isolation of neoantigen-specific T cells from tumor and peripheral lymphocytes. *The Journal of Clinical Investigation*, 125(10):3981–3991, October 2015.

[22] The GTEx Consortium. The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science*, 348(6235):648–660, May 2015.

[23] Maura Costello, Trevor J. Pugh, Timothy J. Fennell, Chip Stewart, Lee Lichtenstein, James C. Meldrim, Jennifer L. Fostel, Dennis C. Friedrich, Danielle Perrin, Danielle Dionne, Sharon Kim, Stacey B. Gabriel, Eric S. Lander, Sheila Fisher, and Gad Getz. Discovery and characterization of artifactual mutations in deep coverage targeted capture sequencing data due to oxidative DNA damage during sample preparation. *Nucleic Acids Research*, 41(6):e67–e67, April 2013.

[24] Jason N. Crosson, Philip W. Laird, Matthew Debiec, Chris S. Bergstrom, David H. Lawson, and Steven Yeh. Vogt-koyanagi-harada-like syndrome after CTLA-4 inhibition with ipilimumab for metastatic melanoma:. *Journal of Immunotherapy*, 38(2):80–84, 2015.

[25] Tyler J. Curiel, Shuang Wei, Haidong Dong, Xavier Alvarez, Pui Cheng, Peter Mottram, Roman Krzysiek, Keith L. Knutson, Ben Daniel, Maria Carla Zimmermann, Odile David, Matthew Burow, Alan Gordon, Nina Dhurandhar, Leann Myers, Ruth Berggren, Akseli Hemminki, Ronald D. Alvarez, Dominique Emilie, David T. Curiel, Lieping Chen, and Weiping Zou. Blockade of B7-H1 improves myeloid dendritic cell–mediated antitumor immunity. *Nat Med*, 9(5):562–567, May 2003.

[26] Jacques Deguine, Béatrice Breart, Fabrice Lemaître, James P. Di Santo, and Philippe Bousso. Intravital imaging reveals distinct dynamics for natural killer and CD8+ t cells during tumor regression. *Immunity*, 33(4):632–644, October 2010.

[27] Alexander Dobin, Carrie A. Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R. Gingeras. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, 29(1):15–21, January 2013.

[28] Fei Duan, Jorge Duitama, Sahar Al Seesi, Cory M. Ayres, Steven A. Corcelli, Arpita P. Pawashe, Tatiana Blanchard, David McMahon, John Sidney, Alessandro Sette, Brian M. Baker, Ion I. Mandoiu, and Pramod K. Srivastava. Genomic and bioinformatic profiling of mutational neoepitopes reveals new rules to predict anticancer immunogenicity. *The Journal of Experimental Medicine*, 211(11):2231–2248, October 2014.

[29] Gavin P. Dunn, Allen T. Bruce, Hiroaki Ikeda, Lloyd J. Old, and Robert D. Schreiber. Cancer immunoediting: from immunosurveillance to tumor escape. *Nat Immunol*, 3(11):991–998, November 2002.

[30] Gavin P. Dunn, Lloyd J. Old, and Robert D. Schreiber. The Three Es of Cancer Immunoediting. *Annual Review of Immunology*, 22(1):329–360, 2004.

[31] Andrea van Elsas, Roger P. M. Sutmuller, Arthur A. Hurwitz, Jennifer Ziskin, Jennifer Villasenor, Jan-Paul Medema, Willem W. Overwijk, Nicholas P. Restifo, Cornelis J. M. Melief, Rienk Offringa, and James P. Allison. Elucidating the Autoimmune and Antitumor Effector Mechanisms of a Treatment Based on Cytotoxic T Lymphocyte Antigen-4 Blockade in Combination with a B16 Melanoma Vaccine: Comparison of Prophylaxis and Therapy. *Journal of Experimental Medicine*, 194(4):481–490, August 2001.

[32] Z. Eshhar, T. Waks, G. Gross, and D. G. Schindler. Specific activation and targeting of cytotoxic lymphocytes through chimeric single chains consisting of antibody-binding domains and the gamma or zeta subunits of the immunoglobulin and T-cell receptors. *Proceedings of the National Academy of Sciences*, 90(2):720–724, January 1993.

[33] Adam D. Ewing, Kathleen E. Houlahan, Yin Hu, Kyle Ellrott, Cristian Caloian, Takafumi N. Yamaguchi, J. Christopher Bare, Christine P'ng, Daryl Waggott, Veronica Y. Sabelnykova, ICGC-TCGA DREAM Somatic Mutation Calling Challenge Participants, Liu Xi, Ninad Dewal, Yu Fan, Wenyi Wang, David Wheeler, Andreas Wilm, Grace Hui Ting, Chenhao Li, Denis Bertrand, Niranjan Nagarajan, Qing-Rong Chen, Chih-Hao Hsu, Ying Hu, Chunhua Yan, Warren Kibbe, Daoud Meerzaman, Kristian Cibulskis, Mara Rosenberg, Louis Bergelson, Adam Kiezun, Amie Radenbaugh, Anne-Sophie Sertier, Anthony Ferrari, Laurie Tonton, Kunal Bhutani, Nancy F. Hansen, Difei Wang, Lei Song, Zhongwu Lai, Yang Liao, Wei Shi, José Carbonell-Caballero, Joaquín Dopazo, Cheryl C. K. Lau, Justin Guinney, Michael R. Kellen, Thea C. Norman, David Haussler, Stephen H. Friend, Gustavo Stolovitzky, Adam A. Margolin, Joshua M. Stuart, and

Paul C. Boutros. Combining tumor genome simulation with crowdsourcing to benchmark somatic single-nucleotide-variant detection. *Nature Methods*, 12(7):623–630, July 2015.

[34] Andrea Facciabene, Gregory T. Motz, and George Coukos. T-Regulatory Cells: Key Players in Tumor Immune Escape and Angiogenesis. *Cancer Research*, 72(9):2162–2171, May 2012.

[35] Yu Fan, Liu Xi, Daniel S. T. Hughes, Jianjun Zhang, Jianhua Zhang, P. Andrew Futreal, David A. Wheeler, and Wenyi Wang. MuSE: accounting for tumor heterogeneity using a sample-specific error model improves sensitivity and specificity in mutation calling from sequencing data. *Genome Biology*, 17:178, August 2016.

[36] Robert L. Ferris, Jennifer L. Hunt, and Soldano Ferrone. Human leukocyte antigen (HLA) class I defects in head and neck cancer: molecular mechanisms and clinical significance. *Immunologic Research*, 33(2):113–133, 2005.

[37] Brian T. Fife and Kristen E. Pauken. The role of the PD-1 pathway in autoimmunity and peripheral tolerance. *Ann. N. Y. Acad. Sci.*, 1217:45–59, January 2011.

[38] Loise M. Francisco, Peter T. Sage, and Arlene H. Sharpe. The PD-1 pathway in tolerance and autoimmunity. *Immunol. Rev.*, 236:219–242, July 2010.

[39] François Ghiringhelli, Pierre E. Puig, Stephan Roux, Arnaud Parcellier, Elise Schmitt, Eric Solary, Guido Kroemer, François Martin, Bruno Chauffert, and Laurence Zitvogel. Tumor cells convert immature myeloid dendritic cells into TGF-β–secreting cells inducing CD4+CD25+ regulatory T cell proliferation. *J Exp Med*, 202(7):919–929, October 2005.

[40] Brian Haas, Alexander Dobin, Nicolas Stransky, Bo Li, Xiao Yang, Timothy Tickle, Asma Bankapur, Carrie Ganote, Thomas Doak, Natalie Pochet, Jing Sun, Catherine Wu, Thomas Gingeras, and Aviv Regev. STAR-Fusion: Fast and Accurate Fusion Transcript Detection from RNA-Seq. *bioRxiv*, January 2017.

[41] Brian J. Haas, Alexie Papanicolaou, Moran Yassour, Manfred Grabherr, Philip D. Blood, Joshua Bowden, Matthew Brian Couger, David Eccles, Bo Li, Matthias Lieber, Matthew D. MacManes, Michael Ott, Joshua Orvis, Nathalie Pochet, Francesco Strozzi, Nathan Weeks, Rick Westerman, Thomas William, Colin N. Dewey, Robert Henschel, Richard D. LeDuc, Nir Friedman, and Aviv Regev. *De novo* transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nature Protocols*, 8(8):1494–1512, August 2013.

[42] Yusuf A. Hannun. Apoptosis and the Dilemma of Cancer Chemotherapy. *Blood*, 89(6):1845–1853, March 1997.

[43] Jennifer Harrow, Adam Frankish, Jose M. Gonzalez, Electra Tapanari, Mark Diekhans, Felix Kokocinski, Bronwen L. Aken, Daniel Barrell, Amonida Zadissa, Stephen Searle,

If Barnes, Alexandra Bignell, Veronika Boychenko, Toby Hunt, Mike Kay, Gaurab Mukherjee, Jeena Rajan, Gloria Despacio-Reyes, Gary Saunders, Charles Steward, Rachel Harte, Michael Lin, Cédric Howald, Andrea Tanzer, Thomas Derrien, Jacqueline Chrast, Nathalie Walters, Suganthi Balasubramanian, Baikang Pei, Michael Tress, Jose Manuel Rodriguez, Iakes Ezkurdia, Jeltje van Baren, Michael Brent, David Haussler, Manolis Kellis, Alfonso Valencia, Alexandre Reymond, Mark Gerstein, Roderic Guigó, and Tim J. Hubbard. GENCODE: The reference human genome annotation for The ENCODE Project. *Genome Research*, 22(9):1760–1774, September 2012.

[44] Larry A. Harshyne, Michael I. Zimmer, Simon C. Watkins, and Simon M. Barratt-Boyes. A role for class a scavenger receptor in dendritic cell nibbling from live cells. *J Immunol*, 170(5):2302–2309, March 2003.

[45] William R. Heath and Francis R. Carbone. Cross-presentation in viral immunity and self-tolerance. *Nat Rev Immunol*, 1(2):126–134, November 2001.

[46] Ronald B. Herberman, Myrthel E. Nunn, and David H. Lavrin. Natural cytotoxic reactivity of mouse lymphoid cells against syngeneic and allogeneic tumors. I. Distribution of reactivity and specificity. *Int. J. Cancer*, 16(2):216–229, August 1975.

[47] Robin Hill, Brendan Healy, Lois Holloway, Zdenka Kuncic, David Thwaites, and Clive Baldock. Advances in kilovoltage x-ray beam dosimetry. *Phys. Med. Biol.*, 59(6):R183, March 2014.

[48] B. Hindman, A. Konwinski, M. Zaharia, A. Ghodsi, A. D. Joseph, R. H. Katz, S. Shenker, and Stoica I. Mesos: A platform for fine-grained resource sharing in the data center. *Technical Report UCB/EECS-2010-87, EECS Department, University of California, Berkeley*, May 2010.

[49] F. Stephen Hodi, Martin C. Mihm, Robert J. Soiffer, Frank G. Haluska, Marcus Butler, Michael V. Seiden, Thomas Davis, Rochele Henry-Spires, Suzanne MacRae, Ann Willman, Robert Padera, Michael T. Jaklitsch, Sridhar Shankar, Teresa C. Chen, Alan Korman, James P. Allison, and Glenn Dranoff. Biologic activity of cytotoxic T lymphocyte-associated antigen 4 antibody blockade in previously vaccinated metastatic melanoma and ovarian carcinoma patients. *PNAS*, 100(8):4712–4717, April 2003.

[50] F. Stephen Hodi, Steven J. O'Day, David F. McDermott, Robert W. Weber, Jeffrey A. Sosman, John B. Haanen, Rene Gonzalez, Caroline Robert, Dirk Schadendorf, Jessica C. Hassel, Wallace Akerley, Alfons J.M. van den Eertwegh, Jose Lutzky, Paul Lorigan, Julia M. Vaubel, Gerald P. Linette, David Hogg, Christian H. Ottensmeier, Celeste Lebbé, Christian Peschel, Ian Quirt, Joseph I. Clark, Jedd D. Wolchok, Jeffrey S. Weber, Jason Tian, Michael J. Yellin, Geoffrey M. Nichol, Axel Hoos, and Walter J. Urba. Improved Survival with Ipilimumab in Patients with Metastatic Melanoma. *New England Journal of Medicine*, 363(8):711–723, August 2010.

[51] Jasreet Hundal, Beatriz M. Carreno, Allegra A. Petti, Gerald P. Linette, Obi L. Griffith, Elaine R. Mardis, and Malachi Griffith. pVAC-Seq: A genome-guided in silico approach to identifying tumor neoantigens. *Genome Medicine*, 8:11, 2016.

[52] Bryan A. Irving and Arthur Weiss. The cytoplasmic domain of the T cell receptor ζ chain is sufficient to couple to receptor-associated signal transduction pathways. *Cell*, 64(5):891–901, March 1991.

[53] Liza B. John, Christel Devaud, Connie P. M. Duong, Carmen S. Yong, Paul A. Beavis, Nicole M. Haynes, Melvyn T. Chow, Mark J. Smyth, Michael H. Kershaw, and Phillip K. Darcy. Anti-PD-1 Antibody Therapy Potently Enhances the Eradication of Established Tumors By Gene-Modified T Cells. *Clin Cancer Res*, 19(20):5636–5646, October 2013.

[54] Laura A. Johnson, Richard A. Morgan, Mark E. Dudley, Lydie Cassard, James C. Yang, Marybeth S. Hughes, Udai S. Kammula, Richard E. Royal, Richard M. Sherry, John R. Wunderlich, Chyi-Chia R. Lee, Nicholas P. Restifo, Susan L. Schwarz, Alexandria P. Cogdill, Rachel J. Bishop, Hung Kim, Carmen C. Brewer, Susan F. Rudy, Carter VanWaes, Jeremy L. Davis, Aarti Mathur, Robert T. Ripley, Debbie A. Nathan, Carolyn M. Laurencot, and Steven A. Rosenberg. Gene therapy with human and mouse t-cell receptors mediates cancer regression and targets normal tissues expressing cognate antigen. *Blood*, 114(3):535–546, July 2009.

[55] Johanna A. Joyce and Douglas T. Fearon. T cell exclusion, immune privilege, and the tumor microenvironment. *Science*, 348(6230):74–80, April 2015.

[56] Sharon Stranford Judy Owen, Jenni Punt, editor. *Kuby Immunology 7th Edition*. Freeman, W. H. & Company, 2013.

[57] Michael Kalos, Bruce L. Levine, David L. Porter, Sharyn Katz, Stephan A. Grupp, Adam Bagg, and Carl H. June. T cells with chimeric antigen receptors have potent antitumor effects and can establish memory in patients with advanced leukemia. *Sci Transl Med*, 3(95):95ra73–95ra73, August 2011.

[58] Edita Karosiene, Claus Lundegaard, Ole Lund, and Morten Nielsen. NetMHCcons: a consensus method for the major histocompatibility complex class I predictions. *Immunogenetics*, 64(3):177–186, March 2012.

[59] M.A. Kenneth D. Kochanek, B.S. Sherry L. Murphy, M.D. Jiaquan Xu, and Ph.D Elizabeth Arias. Products - Data Briefs - Number 178 - December 2014.

[60] W. James Kent, Charles W. Sugnet, Terrence S. Furey, Krishna M. Roskin, Tom H. Pringle, Alan M. Zahler, and and David Haussler. The Human Genome Browser at UCSC. *Genome Res.*, 12(6):996–1006, June 2002.

[61] Michael H. Kershaw, Michele W. L. Teng, Mark J. Smyth, and Phillip K. Darcy. Supernatural T cells: genetic modification of T cells for cancer therapy. *Nature Reviews Immunology*, 5(12):928–940, December 2005.

[62] Ryungsa Kim, Manabu Emi, and Kazuaki Tanabe. Cancer immunoediting from immune surveillance to immune escape. *Immunology*, 121(1):1–14, May 2007.

[63] Kazuma Kiyotani, Tu H. Mai, and Yusuke Nakamura. Comparison of exome-based *HLA* class I genotyping tools: identification of platform-specific genotyping errors. *Journal of Human Genetics*, 62(3):397, March 2017.

[64] Mélanie Lambotin, Sukanya Raghuraman, Françoise Stoll-Keller, Thomas F. Baumert, and Heidi Barth. A look behind closed doors: interaction of persistent viruses with dendritic cells. *Nat Rev Micro*, 8(5):350–360, April 2010.

[65] David E. Larson, Christopher C. Harris, Ken Chen, Daniel C. Koboldt, Travis E. Abbott, David J. Dooling, Timothy J. Ley, Elaine R. Mardis, Richard K. Wilson, and Li Ding. SomaticSniper: identification of somatic point mutations in whole genome sequencing data. *Bioinformatics*, 28(3):311–317, February 2012.

[66] Michael S. Lawrence, Petar Stojanov, Paz Polak, Gregory V. Kryukov, Kristian Cibulskis, Andrey Sivachenko, Scott L. Carter, Chip Stewart, Craig H. Mermel, Steven A. Roberts, Adam Kiezun, Peter S. Hammerman, Aaron McKenna, Yotam Drier, Lihua Zou, Alex H. Ramos, Trevor J. Pugh, Nicolas Stransky, Elena Helman, Jaegil Kim, Carrie Sougnez, Lauren Ambrogio, Elizabeth Nickerson, Erica Shefler, Maria L. Cortés, Daniel Auclair, Gordon Saksena, Douglas Voet, Michael Noble, Daniel DiCara, Pei Lin, Lee Lichtenstein, David I. Heiman, Timothy Fennell, Marcin Imielinski, Bryan Hernandez, Eran Hodis, Sylvan Baca, Austin M. Dulak, Jens Lohr, Dan-Avi Landau, Catherine J. Wu, Jorge Melendez-Zajgla, Alfredo Hidalgo-Miranda, Amnon Koren, Steven A. McCarroll, Jaume Mora, Ryan S. Lee, Brian Crompton, Robert Onofrio, Melissa Parkin, Wendy Winckler, Kristin Ardlie, Stacey B. Gabriel, Charles W. M. Roberts, Jaclyn A. Biegel, Kimberly Stegmaier, Adam J. Bass, Levi A. Garraway, Matthew Meyerson, Todd R. Golub, Dmitry A. Gordenin, Shamil Sunyaev, Eric S. Lander, and Gad Getz. Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature*, 499(7457):214–218, July 2013.

[67] Rasko Leinonen, Hideaki Sugawara, and Martin Shumway. The Sequence Read Archive. *Nucleic Acids Research*, 39(Database issue):D19–D21, January 2011.

[68] Bo Li and Colin N. Dewey. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*, 12:323, August 2011.

[69] Heng Li and Richard Durbin. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*, 25(14):1754–1760, July 2009.

[70] Grazyna Lipowska-Bhalla, David E. Gilham, Robert E. Hawkins, and Dominic G. Rothwell. Targeted immunotherapy of cancer with CAR t cells: achievements and challenges. *Cancer Immunol Immunother*, 61(7):953–962, April 2012.

[71] Claus Lundegaard, Kasper Lamberth, Mikkel Harndahl, Søren Buus, Ole Lund, and Morten Nielsen. NetMHC-3.0: accurate web accessible predictions of human, mouse

and monkey MHC class I affinities for peptides of length 8–11. *Nucleic Acids Research*, 36(suppl_2):W509–W512, July 2008.

[72] Matthew Mansh. Ipilimumab and Cancer Immunotherapy: A New Hope for Advanced Stage Melanoma. *Yale J Biol Med*, 84(4):381–389, December 2011.

[73] John M. Maris. Recent Advances in Neuroblastoma. *New England Journal of Medicine*, 362(23):2202–2211, June 2010.

[74] Hirokazu Matsushita, Matthew D. Vesely, Daniel C. Koboldt, Charles G. Rickert, Ravindra Uppaluri, Vincent J. Magrini, Cora D. Arthur, J. Michael White, Yee-Shiuan Chen, Lauren K. Shea, Jasreet Hundal, Michael C. Wendl, Ryan Demeter, Todd Wylie, James P. Allison, Mark J. Smyth, Lloyd J. Old, Elaine R. Mardis, and Robert D. Schreiber. Cancer exome analysis reveals a T-cell-dependent mechanism of cancer immunoediting. *Nature*, 482(7385):400–404, February 2012.

[75] William McLaren, Laurent Gil, Sarah E. Hunt, Harpreet Singh Riat, Graham R. S. Ritchie, Anja Thormann, Paul Flicek, and Fiona Cunningham. The Ensembl Variant Effect Predictor. *Genome Biology*, 17:122, June 2016.

[76] Peter W. Mesner Jr., I. Imawati Budihardjo, and Scott H. Kaufmann. Chemotherapy-Induced Apoptosis. In Scott H. Kaufmann, editor, *Advances in Pharmacology*, volume 41 of *Apoptosls Pharmacological Implications and Therapeutic Opportunities*, pages 461–499. Academic Press, 1997.

[77] Oren Milstein, David Hagin, Assaf Lask, Shlomit Reich-Zeliger, Elias Shezan, Eran Ophir, Yaki Eidelstein, Ran Afik, Yaron E. Antebi, Michael L. Dustin, and Yair Reisner. CTLs respond with activation and granule secretion when serving target for t cell recognition. *Blood*, January 2010.

[78] Kenneth Murphy. *Janeway's Immunobiology 8th Edition*. Taylor & Francis Group, 2011.

[79] L. M. Muul, P. J. Spiess, E. P. Director, and S. A. Rosenberg. Identification of specific cytolytic immune responses against autologous tumor in humans bearing malignant melanoma. *J Immunol*, 138(3):989–995, February 1987.

[80] Srinivas Nagaraj and Dmitry I. Gabrilovich. Tumor Escape Mechanism Governed by Myeloid-Derived Suppressor Cells. *Cancer Res*, 68(8):2561–2563, April 2008.

[81] The Cancer Genome Atlas Research Network. The Molecular Taxonomy of Primary Prostate Cancer. *Cell*, 163(4):1011–1025, November 2015.

[82] Morten Nielsen, Claus Lundegaard, and Ole Lund. Prediction of MHC class II binding affinity using SMM-align, a novel stabilization matrix alignment method. *BMC Bioinformatics*, 8:238, July 2007.

[83] Morten Nielsen, Claus Lundegaard, Peder Worning, Sanne Lise Lauemøller, Kasper Lamberth, Søren Buus, Søren Brunak, and Ole Lund. Reliable prediction of T-cell epitopes using neural networks with novel sequence representations. *Protein Science*, 12(5):1007–1017, 2003.

[84] Masasuke Ohno, Takayuki Ohkuri, Akemi Kosaka, Kuniaki Tanahashi, Carl H June, Atsushi Natsume, and Hideho Okada. Expression of miR-17-92 enhances anti-tumor activity of t-cells transduced with the anti-EGFRvIII chimeric antigen receptor in mice bearing human GBM xenografts. *J Immunother Cancer*, 1:21, December 2013.

[85] Patrick A. Ott, Zhuting Hu, Derin B. Keskin, Sachet A. Shukla, Jing Sun, David J. Bozym, Wandi Zhang, Adrienne Luoma, Anita Giobbie-Hurder, Lauren Peter, Christina Chen, Oriol Olive, Todd A. Carter, Shuqiang Li, David J. Lieb, Thomas Eisenhaure, Evisa Gjini, Jonathan Stevens, William J. Lane, Indu Javeri, Kaliappanadar Nellaiappan, Andres M. Salazar, Heather Daley, Michael Seaman, Elizabeth I. Buchbinder, Charles H. Yoon, Maegan Harden, Niall Lennon, Stacey Gabriel, Scott J. Rodig, Dan H. Barouch, Jon C. Aster, Gad Getz, Kai Wucherpfennig, Donna Neuberg, Jerome Ritz, Eric S. Lander, Edward F. Fritsch, Nir Hacohen, and Catherine J. Wu. An immunogenic personal neoantigen vaccine for patients with melanoma. *Nature*, 547(7662):217–221, July 2017.

[86] Bjoern Peters and Alessandro Sette. Generating quantitative models describing the sequence specificity of biological processes with the stabilized matrix method. *BMC Bioinformatics*, 6:132, May 2005.

[87] Trevor J. Pugh, Olena Morozova, Edward F. Attiyeh, Shahab Asgharzadeh, Jun S. Wei, Daniel Auclair, Scott L. Carter, Kristian Cibulskis, Megan Hanna, Adam Kiezun, Jaegil Kim, Michael S. Lawrence, Lee Lichenstein, Aaron McKenna, Chandra Sekhar Pedamallu, Alex H. Ramos, Erica Shefler, Andrey Sivachenko, Carrie Sougnez, Chip Stewart, Adrian Ally, Inanc Birol, Readman Chiu, Richard D. Corbett, Martin Hirst, Shaun D. Jackman, Baljit Kamoh, Alireza Hadj Khodabakshi, Martin Krzywinski, Allan Lo, Richard A. Moore, Karen L. Mungall, Jenny Qian, Angela Tam, Nina Thiessen, Yongjun Zhao, Kristina A. Cole, Maura Diamond, Sharon J. Diskin, Yael P. Mosse, Andrew C. Wood, Lingyun Ji, Richard Sposto, Thomas Badgett, Wendy B. London, Yvonne Moyer, Julie M. Gastier-Foster, Malcolm A. Smith, Jaime M. Guidry Auvil, Daniela S. Gerhard, Michael D. Hogarty, Steven J. M. Jones, Eric S. Lander, Stacey B. Gabriel, Gad Getz, Robert C. Seeger, Javed Khan, Marco A. Marra, Matthew Meyerson, and John M. Maris. The genetic landscape of high-risk neuroblastoma. *Nature Genetics*, 45(3):279–284, March 2013.

[88] M. Pule, H. Finney, and A. Lawson. Artificial t-cell receptors. *Cytotherapy*, 5(3):211–226, 2003.

[89] B. B. Rad, H. J. Bhatti, and M. Ahmadi. An introduction to docker and analysis of its performance. *IJCSNS International Journal of Computer Science and Network Security*, 173, March 2017.

[90] Amie J. Radenbaugh, Singer Ma, Adam Ewing, Joshua M. Stuart, Eric A. Collisson, Jingchun Zhu, and David Haussler. RADIA: RNA and DNA Integrated Analysis for Somatic Mutation Detection. *PLoS ONE*, 9(11):e111516, November 2014.

[91] Martin Reck, Delvys Rodríguez-Abreu, Andrew G. Robinson, Rina Hui, Tibor Csőszi, Andrea Fülöp, Maya Gottfried, Nir Peled, Ali Tafreshi, Sinead Cuffe, Mary O'Brien, Suman Rao, Katsuyuki Hotta, Melanie A. Leiby, Gregory M. Lubiniecki, Yue Shentu, Reshma Rangwala, and Julie R. Brahmer. Pembrolizumab versus Chemotherapy for PD-L1–Positive Non–Small-Cell Lung Cancer. *New England Journal of Medicine*, 375(19):1823–1833, November 2016.

[92] Nicholas P. Restifo, Mark E. Dudley, and Steven A. Rosenberg. Adoptive immunotherapy for cancer: harnessing the t cell response. *Nat Rev Immunol*, 12(4):269–281, April 2012.

[93] Steven A. Rosenberg and Nicholas P. Restifo. Adoptive cell transfer as personalized immunotherapy for human cancer. *Science*, 348(6230):62–68, April 2015.

[94] Steven A. Rosenberg, James C. Yang, Richard M. Sherry, Udai S. Kammula, Marybeth S. Hughes, Giao Q. Phan, Deborah E. Citrin, Nicholas P. Restifo, Paul F. Robbins, John R. Wunderlich, Kathleen E. Morton, Carolyn M. Laurencot, Seth M. Steinberg, Donald E. White, and Mark E. Dudley. Durable complete responses in heavily pretreated patients with metastatic melanoma using t-cell transfer immunotherapy. *Clin Cancer Res*, 17(13):4550–4557, July 2011.

[95] Mabel Ryder, Margaret Callahan, Michael A. Postow, Jedd Wolchok, and James A. Fagin. Endocrine-related adverse events following ipilimumab in patients with advanced melanoma: a comprehensive retrospective review from a single institution. *Endocrine-Related Cancer*, 21(2):371–381, April 2014.

[96] Michel Sadelain, Renier Brentjens, and Isabelle Rivière. The promise and potential pitfalls of chimeric antigen receptors. *Current Opinion in Immunology*, 21(2):215–223, April 2009.

[97] Ugur Sahin, Evelyna Derhovanessian, Matthias Miller, Björn-Philipp Kloke, Petra Simon, Martin Löwer, Valesca Bukur, Arbel D. Tadmor, Ulrich Luxemburger, Barbara Schrörs, Tana Omokoko, Mathias Vormehr, Christian Albrecht, Anna Paruzynski, Andreas N. Kuhn, Janina Buck, Sandra Heesch, Katharina H. Schreeb, Felicitas Müller, Inga Ortseifer, Isabel Vogler, Eva Godehardt, Sebastian Attig, Richard Rae, Andrea Breitkreuz, Claudia Tolliver, Martin Suchan, Goran Martic, Alexander Hohberger, Patrick Sorn, Jan Diekmann, Janko Ciesla, Olga Waksmann, Alexandra-Kemmer Brück, Meike Witt, Martina Zillgen, Andree Rothermel, Barbara Kasemann, David Langer, Stefanie Bolte, Mustafa Diken, Sebastian Kreiter, Romina Nemecek, Christoffer Gebhardt, Stephan Grabbe, Christoph Höller, Jochen Utikal, Christoph Huber, Carmen Loquai, and Özlem Türeci. Personalized RNA mutanome vaccines mobilize poly-specific therapeutic immunity against cancer. *Nature*, 547(7662):222–226, July 2017.

[98] Ugur Sahin and Özlem Türeci. Personalized vaccines for cancer immunotherapy. *Science*, 359(6382):1355–1360, March 2018.

[99] S. Sakaguchi, N. Sakaguchi, M. Asano, M. Itoh, and M. Toda. Immunologic self-tolerance maintained by activated t cells expressing IL-2 receptor alpha-chains (CD25). breakdown of a single mechanism of self-tolerance causes various autoimmune diseases. *J Immunol*, 155(3):1151–1164, August 1995.

[100] Federica Sallusto and Antonio Lanzavecchia. The instructive role of dendritic cells on t-cell responses. *Arthritis Res*, 4(Suppl 3):S127–S132, 2002.

[101] Oliver Sartor. Overview of Samarium Sm 153 Lexidronam in the Treatment of Painful Metastatic Bone Disease. *Rev Urol*, 6(Suppl 10):S3–S12, 2004.

[102] Christopher T. Saunders, Wendy S. W. Wong, Sajani Swamy, Jennifer Becq, Lisa J. Murray, and R. Keira Cheetham. Strelka: accurate somatic small-variant calling from sequenced tumor–normal sample pairs. *Bioinformatics*, 28(14):1811–1817, July 2012.

[103] Andrew M. Scott, Jedd D. Wolchok, and Lloyd J. Old. Antibody therapy of cancer. *Nat Rev Cancer*, 12(4):278–287, April 2012.

[104] Padmanee Sharma and James P. Allison. The future of immune checkpoint therapy. *Science*, 348(6230):56–61, April 2015.

[105] Ethan M. Shevach. Fatal attraction: tumors beckon regulatory T cells. *Nature Medicine*, 10(9):900–901, September 2004.

[106] Sachet A. Shukla, Michael S. Rooney, Mohini Rajasagi, Grace Tiao, Philip M. Dixon, Michael S. Lawrence, Jonathan Stevens, William J. Lane, Jamie L. Dellagatta, Scott Steelman, Carrie Sougnez, Kristian Cibulskis, Adam Kiezun, Nir Hacohen, Vladimir Brusic, Catherine J. Wu, and Gad Getz. Comprehensive analysis of cancer-associated somatic mutations in class I HLA genes. *Nature Biotechnology*, 33(11):1152–1158, November 2015.

[107] Rebecca L. Siegel, Kimberly D. Miller, and Ahmedin Jemal. Cancer statistics, 2015. *CA: A Cancer Journal for Clinicians*, 65(1):5–29, January 2015.

[108] Alexandra Snyder, Vladimir Makarov, Taha Merghoub, Jianda Yuan, Jesse M. Zaretsky, Alexis Desrichard, Logan A. Walsh, Michael A. Postow, Phillip Wong, Teresa S. Ho, Travis J. Hollmann, Cameron Bruggeman, Kasthuri Kannan, Yanyun Li, Ceyhan Elipenahli, Cailian Liu, Christopher T. Harbison, Lisu Wang, Antoni Ribas, Jedd D. Wolchok, and Timothy A. Chan. Genetic Basis for Clinical Response to CTLA-4 Blockade in Melanoma. *New England Journal of Medicine*, 0(0):null, November 2014.

[109] Ralph M. Steinman and Zanvil A. Cohn. Identification of a novel cell type in peripheral lymphoid organs of mice. *J Exp Med*, 137(5):1142–1162, May 1973.

[110] Tiziana Sturniolo, Elisa Bono, Jiayi Ding, Laura Raddrizzani, Oezlem Tuereci, Ugur Sahin, Michael Braxenthaler, Fabio Gallazzi, Maria Pia Protti, Francesco Sinigaglia, and Juergen Hammer. Generation of tissue-specific and promiscuous HLA ligand databases using DNA microarrays and virtual HLA class II matrices. *Nature Biotechnology*, 17(6):555–561, June 1999.

[111] Jugmohit S. Toor, Arjun A. Rao, Andrew C. McShan, Mark Yarmarkovich, Santrupti Nerli, Karissa Yamaguchi, Ada A. Madejska, Son Nguyen, Sarvind Tripathi, John M. Maris, Sofie R. Salama, David Haussler, and Nikolaos G. Sgourakis. A Recurrent Mutation in Anaplastic Lymphoma Kinase with Distinct Neoepitope Conformations. *Frontiers in Immunology*, 9, 2018.

[112] A Townsend and H Bodmer. Antigen recognition by class i-restricted t lymphocytes. *Annual Review of Immunology*, 7(1):601–624, 1989.

[113] Eric Tran, Simon Turcotte, Alena Gros, Paul F. Robbins, Yong-Chen Lu, Mark E. Dudley, John R. Wunderlich, Robert P. Somerville, Katherine Hogan, Christian S. Hinrichs, Maria R. Parkhurst, James C. Yang, and Steven A. Rosenberg. Cancer immunotherapy based on mutation-specific CD4+ t cells in a patient with epithelial cancer. *Science*, 344(6184):641–645, May 2014.

[114] Nienke van Rooij, Marit M. van Buuren, Daisy Philips, Arno Velds, Mireille Toebes, Bianca Heemskerk, Laura J.A. van Dijk, Sam Behjati, Henk Hilkmann, Dris el Atmioui, Marja Nieuwland, Michael R. Stratton, Ron M. Kerkhoven, Can Keşmir, John B. Haanen, Pia Kvistborg, and Ton N. Schumacher. Tumor Exome Analysis Reveals Neoantigen-Specific T-Cell Reactivity in an Ipilimumab-Responsive Melanoma. *Journal of Clinical Oncology*, 31(32):e439–e442, November 2013.

[115] John Vivian, Arjun Arkal Rao, Frank Austin Nothaft, Christopher Ketchum, Joel Armstrong, Adam Novak, Jacob Pfeil, Jake Narkizian, Alden D. Deran, Audrey Musselman-Brown, Hannes Schmidt, Peter Amstutz, Brian Craft, Mary Goldman, Kate Rosenbloom, Melissa Cline, Brian O'Connor, Megan Hanna, Chet Birger, W. James Kent, David A. Patterson, Anthony D. Joseph, Jingchun Zhu, Sasha Zaranek, Gad Getz, David Haussler, and Benedict Paten. Toil enables reproducible, open source, big biomedical data analyses. *Nature Biotechnology*, 35(4):314–316, April 2017.

[116] Thomas A. Waldmann. Immunotherapy: past, present and future. *Nat Med*, 9(3):269–277, March 2003.

[117] René L. Warren, Gina Choe, Douglas J. Freeman, Mauro Castellarin, Sarah Munro, Richard Moore, and Robert A. Holt. Derivation of HLA types from shotgun sequence datasets. *Genome Medicine*, 4(12):95, December 2012.

[118] K. Yoshihara, Q. Wang, W. Torres-Garcia, S. Zheng, R. Vegesna, H. Kim, and R. G. W. Verhaak. The landscape and therapeutic relevance of cancer-associated transcript fusions. *Oncogene*, 34(37):4845–4854, September 2015.

[119] Nikolaos Zacharakis, Harshini Chinnasamy, Mary Black, Hui Xu, Yong-Chen Lu, Zhili Zheng, Anna Pasetto, Michelle Langhan, Thomas Shelton, Todd Prickett, Jared Gartner, Li Jia, Katarzyna Trebska-McGowan, Robert P. Somerville, Paul F. Robbins, Steven A. Rosenberg, Stephanie L. Goff, and Steven A. Feldman. Immune recognition of somatic mutations leading to complete durable regression in metastatic breast cancer. *Nature Medicine*, page 1, June 2018.

[120] Hao Zhang, Ole Lund, and Morten Nielsen. The PickPocket method for predicting binding specificities for receptors based on receptor pocket similarities: application to MHC-peptide binding. *Bioinformatics*, 25(10):1293–1299, May 2009.

[121] Jin Zhang, Elaine R. Mardis, and Christopher A. Maher. INTEGRATE-neo: a pipeline for personalized gene fusion neoantigen discovery. *Bioinformatics*, 33(4):555–557, February 2017.

[122] Wanding Zhou, Tenghui Chen, Zechen Chong, Mary A. Rohrdanz, James M. Melott, Chris Wakefield, Jia Zeng, John N. Weinstein, Funda Meric-Bernstam, Gordon B. Mills, and Ken Chen. TransVar: a multilevel variant annotator for precision genomics. *Nature Methods*, 12(11):1002–1003, November 2015.

[123] Jinfang Zhu and William E. Paul. CD4 t cells: fates, functions, and faults. *Blood*, 112(5):1557–1569, September 2008.