BIOINFORMATICS METHODS FOR PREDICTION OF SPLICE VARIANT NEOANTIGENS

Shengjie Chai

A dissertation submitted to the faculty at the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Curriculum in Bioinformatics and Computational Biology in the School of Medicine.

Chapel Hill
2019

Approved by:

Benjamin G. Vincent

Jan F. Prins

Jonathan S. Serody

James S. Marron

Terrence S. Furey

Kirk C. Wilhelmsen

ABSTRACT

Shengjie Chai: Bioinformatics Methods for Prediction of Splice Variant Neoantigens
(Under the direction of Benjamin Vincent, Jan Prins, and Jonathan Serody)

Tumor-specific peptide epitopes that are generated from mutated genes and presented on cell surface MHC molecules, known as neoantigens, are attractive targets for therapeutic vaccination given the lack of central tolerance and corresponding presence of endogenous T cells that recognize them. Currently most available neoantigen prediction methods focus on predicting neoantigens derived from missense mutations or indels. In acute myeloid leukemia (AML), there are markedly fewer mutations and predicted neoantigens in the cancer genome compared to other cancers, so it is less feasible to target neoantigens derived from missense mutations and indels in AML. However, mutations in spliceosomal genes and genome-wide aberrant splicing events are common in patients with AML. In work contributed to by our group, a small number of splice variant neoantigens have been found to exist in cancer.

Herein, we report the development of robust method, NeoSplice, to predict splice variant neoantigens from massively parallel RNA sequencing (RNA-Seq) data. One of the computational challenges for predicting splice variant neoantigens is to infer the novel transcript isoforms derived from tumor-specific splicing events. We utilized a Burrows Wheeler Transform (BWT) based algorithm to identify tumor specific k-mers and used a splice graph to determine whether such a k-mer represents a tumor-specific splice junction in a coding region and its corresponding amino-acid sequence. A frame-shift relative to the normal can easily lead to a novel peptide sequence that may be an actionable neoantigen.

Most current neoantigen calling algorithms primarily rely on epitope/MHC binding affinity predictions to rank and select for potential epitope targets. These algorithms do not predict for epitope immunogenicity using approaches modeled from tumor-specific antigen data. We developed an algorithm based on peptide-intrinsic biochemical features associated with neoantigen and minor histocompatibility mismatch antigen (mHA) immunogenicity and present a gradient–boosting algorithm for predicting tumor antigen immunogenicity.

In addition, as part of PhD training in bioinformatics analysis to complement training in methods development, we performed comprehensive genomic and immune characterizations of bladder tumors and triple-negative breast cancer brain metastases to gain novel insight about biomarkers that can be used with potential immunotherapies.

To my family, mentors, and friends, I could not have done this without you.
Thank you for all of your support along the way.

ACKNOWLEDGEMENTS

First of all, I would like to offer countless thanks to my advisors, Dr. Benjamin G. Vincent and Dr. Jan F. Prins for their insightful and tireless support and encouragement, and each of my committee members, Jonathan S. Serody, Terry Furey, James Stephen Marron, and Kirk Wilhelmsen for their thoughtful contributions to this research. I would also like to thank my collaborators from UNC-Chapel Hill for their assistance: Lisle Mose and Joel Parker for assisting me in running the NeoSplice pipeline on Google Cloud for chapter 2, Christof Smith for performing the murine vaccination experiments in Chapter 3, and Dr. Carey Anders for providing data for chapter 4. Finally, I would like to thank my co-workers, friends and family for their patience and support of this work.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1 MOTIVATION AND BIOLOGICAL JUSTIFICATION

Here we will discuss immunology concepts and bioinformatics methods for prediction of neoantigens and immunogenomic characterization of tumors. This section provides an overview of the biological problems addressed by our work.

1.1 Introduction to the immune system and neoantigens.

The immune system is a host defense system composed of biological molecules and processes that protect the body from pathogens. All eukaryotes are protected by a sophisticated mechanism of host defense[1]. Antigens are substances such as toxins, chemicals, bacteria, viruses, and peptides derived from tumor cells that can stimulate immune responses. Each antigen has distinct epitopes that can elicit an antigen specific immune response. The immune system can be divided into an innate immune system and adaptive immune system[1]. The innate immune system consists of phagocytes (neutrophils, Dendritic cells, macrophages), natural killer cells, and the complement protein system[1]. The targets of innate immune system include viruses, bacteria, parasites, and other foreign particles. Phagocytes patrol the body and engulf and digest pathogens or foreign particles. Neutrophils are responsible for killing pathogens that they phagocytose while dendritic cells and macrophages phagocytose pathogens to help clear cellular debris and present antigens to cells of the adaptive immune systems. Natural killer cells can recognize cells with low level of cell surface MHC I molecule expression such as tumor cells or virus-infected cells and release proteins such as perforin and proteases to promote apoptosis or cell lysis. The complement protein system consists of small proteins that are synthesized by the liver and circulate in the blood. Complement proteins bind to the cell surface of cellular pathogens (e.g. bacteria, fungi), with binding potentiated by antibodies targeting the same pathogens, leading to generation of a membrane attack complex that forms pores in target cell membranes to effect cell killing. Although the innate immune system can react rapidly against a broad set of targets via receptors that recognize pathogen associated molecular patterns, it cannot generate antigen-specific responses and lacks capacity for antigen-specific immunologic memory. Immunologic memory is the ability of the immune system to rapidly and specifically recognize and respond to an antigen that has been previously encountered. The adaptive immune system consists of T cells and B cells, which recognize specific antigens via highly diverse cell surface receptors and can generate and maintain immunologic memory by generating memory T cells and memory B cells. These memory cells of the adaptive immune system are induced stochastically following adaptive immune receptor engagement and signaling. They are relatively few in number and reside in the lymphatic tissues, however upon repeat antigen stimulation they clonally expand in large numbers yielding daughter effector cells of the same specificity that are strongly activated and able to respond to the cognate antigen. In this way, effective immunological memory is stored in a set of low frequency but high potency T and B cell populations.

The T-cell receptor (TCR) found on T cells and B-cell receptor (BCR) or immunoglobulin (Ig) produced by B cells play an important role in the adaptive immune system. T cell receptors are protein heterodimers of either an alpha chain + beta chain or gamma chain + delta chain. T cells recognize targets via TCR binding to peptide antigens presented by major histocompatibility complex molecules on the surface of the cell. Peptides of length 8-11 amino acids presented by MHC class I molecules are recognized by CD8+ T cells, whereas peptides of length 15-24 amino acids are recognized by MHC class II molecules. Once a T cell binds its cognate peptide/MHC through the TCR, signaling downstream of the TCR drives a program that includes activation, clonal expansion, cytokine production (CD4+ and CD8+ T cells), and capacity to effect cytotoxicity of targets via production of perforin, granzyme, and Fas ligand. Immunoglobulin (Ig) consist of two closely related forms: BCR that is attached to the surface of B cells and antibody that is secreted by B cells. The two immunoglobulin forms are structurally analogous. However, the BCR contains an additional transmembrane region that anchors it to the cell surface and allows it to participate in intracellular signaling. Antibody molecules, in contrast, have the same antigen-binding specificity as the BCR but are secreted by the cell and can act remotely to bind pathogens and stimulate clearance of these through complement-mediated lysis or phagocytosis by macrophages (with both mechanisms being cooperation between the adaptive (antibody) and innate (complement, macrophages) immune systems).

The adaptive immune system generates diversity of antigen-specific recognition in T cells and B cells via V(D)J recombination, a unique mechanism of genetic recombination that occurs during the maturation stage of T cells and B cells development and leads to a repertoire of antibodies/immunoglobulins (Igs) and T cell receptors (TCRs) found on B cells and T cells respectively. Specifically, The T-cell receptor (TCR) consists of variable alpha and beta or gamma and delta chains, and the BCR/antibody consists of variable light and heavy chains. During somatic recombination, the beta chains in TCR and heavy chains in BCR randomly select one copy of Variable (V), Joining (J), and Diversity (D) gene segments to form a unique variable region. The light chains of BCR and alpha chains of TCR undergo VJ recombination and do not contain a D region. Some diversity of TCR and BCR arises from the large number of V, D, and J gene segments that may be chosen from. For example, there are 44 Variable (V) gene segments, 27 Diversity (D), and 6 Joining (J) gene segments to be selected and combined for human immunoglobulin heavy chain during somatic recombination[2,3]. Additional diversity of the TCR and BCR repertoires is generated by the stochastic germline nucleotide deletion and addition of non-germline templated nucleotides at the V-D and D-J junctions. Thus, the total theoretical repertoire diversity for TCR and BCR is ~10^{15} unique sequences. The number of unique TCR/BCR sequences in any individual is < 10^{10}, the theoretical repertoires are vastly larger than the real repertoire in any person. This provides the population with a robust set of antigen receptors that can recognize diverse pathogen-derived antigens to drive productive immune responses.

T cells are widely recognized as the most important immune cell population for mediating anti-tumor immunity, including direct anti-tumor cytotoxicity in the case of CD8+ T cells and generation of pro-inflammatory cytokines and chemokines in the case of CD4+ T cells. A small number of B cells and T cells can proliferate and expand into a large oligoclonal population when a specific antigen is recognized. Typical targets of adaptive immune system include antigens derived from pathogens such as viruses and bacteria, however tumor-derived antigens can also be targeted. Tumor-specific variant peptides that are generated from mutated genes and can be presented by MHC molecules on the tumor cell surface are known as *neoantigens* (i.e. "new" antigens) since they are: (1) not present in normal non-cancerous cells, (2) derived from genomic variants that are present in the tumor cells, and (3) able to be recognized by the patient's T cells. Normally, T cells specific for "self" antigens (i.e. non-mutated self peptides) are deleted during T cell development in the thymus by a process of negative selection known as central tolerance. Neoantigens are attractive targets for therapeutic vaccination (i.e. vaccine given to induce anti-tumor immunity after a patient has been diagnosed, rather than as a preventative measure) given the lack of central tolerance and corresponding presence of endogenous T cells that recognize them[4]. The process of how neoantigens are generated and can induce cytotoxic T cell responses is shown in Figure 1.1. First, proteins synthesized in the tumor cell can be ubiquinated and degraded by the proteasome into shorter peptides with 8–11 amino acids[5]. The degraded small peptides will then enter the endoplasmic reticulum (ER) lumen through the transporter associated with antigen processing (TAP)[5]. Some of the degraded small peptides in the ER can bind to the binding groove of MHC I[5,6]. The peptide–MHC class I complexes will then be transported to the membrane surface where they can interact with $CD8^+$ cytotoxic T cells[5]. Exogenous peptides and membrane protein peptides with longer amino acid sequences (15–24 amino acids) will be processed and presented to CD4+ T helper cells via MHC class II on antigen presenting cells (APCs)[5]. Some populations of APCs are able to present peptides derived from intracellular proteins on MHC class II molecules in a process known as cross presentation.

The activation of T cells requires an antigen specific signal through TCR as well as an antigen nonspecific co-stimulatory signal through receptors such as CD28 that interact with molecules such as CD80 and CD86 on APCs. If the peptides presented by MHC molecules are derived from tumor-specific mutations (red peptides in Figure 1.1) and co-stimulatory signal is present, T cell-mediated cytotoxicity can be induced to destroy the cancer cell. Self-antigens (green peptides in Figure 1.1) might not be able to induce T cell-mediated cytotoxicity due to central tolerance processes that eliminate self-reactive T cells and/or peripheral tolerance processed that inactivate self peptide-reactive T cells.

1.2 Introduction to computational prediction of tumor-specific antigens

Recent studies have shown that neoantigen targeted therapy can yield improved *ex vivo* anti-tumor immune responses in patients with advanced and metastatic tumors, although there is a lack of evidence of *in vivo* anti-tumor activity due to limited numbers of patients and short follow up time in these phase I trials[7,8]. Typical neoantigen prediction pipelines involve first identifying highly expressed tumor-specific mutations by comparing mutations identified by whole-exome sequencing (WES) data of matched tumor- and normal-cell samples[6,9–16]. RNA sequencing (RNA-seq) data of tumor sample was used to filter lowly expressed mutations[6,9–16]. Highly expressed mutation derived neo-epitopes were prioritized using their predicted binding affinities to HLA I and HLA II molecules expressed by the patient[17–24]. The binding affinities are typically predicted using neural network model trained using peptide-MHC binding affinity measurements and mass spectrometry identified ligand data[17–21,23–25]. Currently most available neoantigen prediction methods such as pVAC-tools, MuPeXI, TSNAD, Neopepsee, and Antigen.garnish focus on predicting neoantigens derived from missense mutations or indels[6,9–16]. These tools all implement the neoantigen prediction pipeline described above with differences in neoantigen ranking and filtering. MuPeXI used a priority score calculated using information such as percent rank affinity of the mutant peptide and normal peptide, RNA-seq expression level, presence of mutant peptide in reference peptidome to rank peptides[14]. pVAC-tools also use a priority score calculated using RNA-seq expression level, binding affinity, and variant allele fraction to rank peptides[9,15]. Additionally, pVAC-tools contains a pVACfuse function that allows prediction of neoantigens derived from gene fusion and a pVACvector function that can extract peptide sequences around the predicted neoantigen to facilitate the design of design of DNA vector-based vaccines and long peptide vaccines[9,15]. In TSNAD, a TMHMM tool was used to predict topology of membrane proteins[13,26]. Additional neoantigens targets that are derived from extracellular mutations of membrane proteins so that they can be targeted by antibodies are included in the final results[13]. TSNAD can also predict neoantigen derived from gene fusions[13]. Neopepsee can classify predicted neoantigens into different immunogenicity categories using a locally weighted naïve Bayes algorithm with nine features including predicted binding affinity, hydrophobicity, polarity and charged score, and amino acid pairwise contact potentials as input[6]. Antigen.garnish predict neoantigen as well as neoantigen immunogenicity and clinical outcome through examining the characteristics such as hydrophobic sequences and dissimilarity of neoantigen to self-proteome[16]. **However, all the neoantigen prediction methods described require DNA sequencing data as well as RNA sequencing data. Furthermore, none of these tools can predict splice variant neoantigens.** The algorithmic details including method, input, algorithm, genomic source(s) of neoantigen prediction algorithms can be found in Table 1.1.

Method	**Input**	**Algorithm**	**Genomic Source(s)**
pVAC-tools	VCF file including DNA and RNA coverage information, BEDPE (gene fusion derived neoantigen prediction)	Enumerate mutant peptides that cover the mutation. Also predict neoantigens derived from gene fusion events. Use features such as RNA-seq expression level, binding affinity, and	DNA sequencing data and RNA sequencing data

		variant allele fraction to rank peptides.	
MuPeXI	VCF file and Expression file	Enumerate mutant peptides that cover the mutation. Use features such as percent rank affinity of the mutant peptide and normal peptide, RNA-seq expression level, presence of mutant peptide in reference peptidome to rank peptides	DNA sequencing data and RNA sequencing data
Neopepsee	VCF file and RNA sequencing FASTQ file	Enumerate mutant peptides that cover the mutation. Use features including predicted binding affinity, hydrophobicity, polarity and charged score, and amino acid pairwise contact	DNA sequencing data and RNA sequencing data

		potentials to rank peptides	
TSNAD	DNA sequencing and RNA sequencing FASTQ file	Enumerate mutant peptides that cover the mutation. Also predict neoantigens derived from gene fusion events. Additional neoantigens targets that are predicted to derive from extracellular mutations of membrane proteins so that can be targeted by antibodies are included in the final result	DNA sequencing data and RNA sequencing data
Antigen.garnish	VCF file and RNA count matrix	Enumerate mutant peptides that cover the mutation. Use the features such as hydrophobic	DNA sequencing data and RNA sequencing data

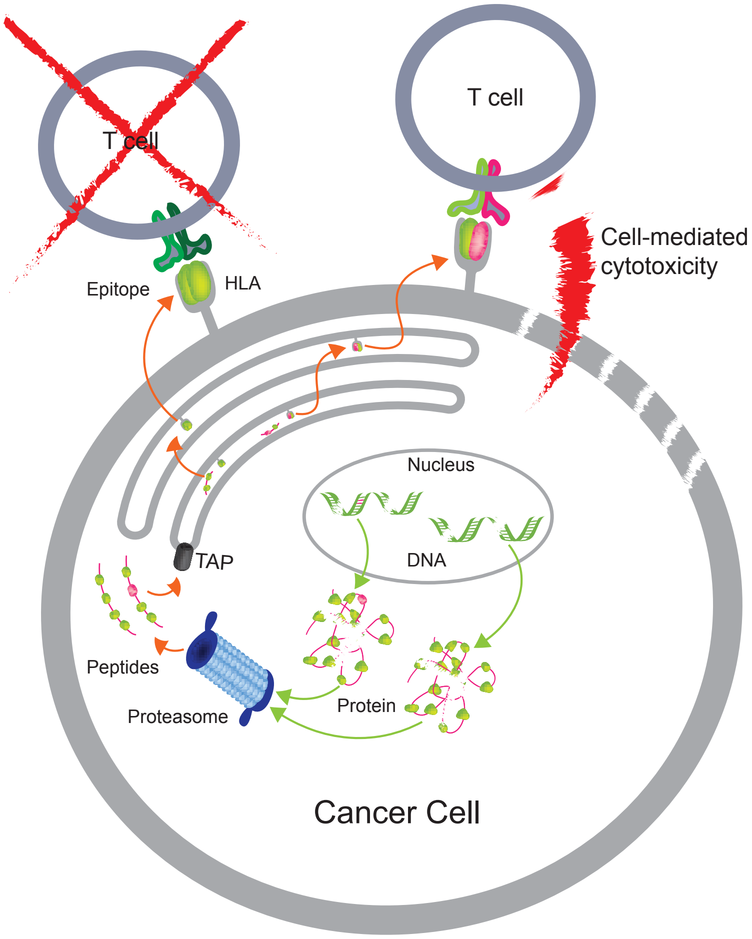
		sequences and dissimilarity of neoantigen to self-proteome to also predict neoantigen immunogenicity and clinical outcome	

Table 1.1 Algorithmic details of neoantigen prediction algorithms

There are markedly fewer mutations and predicted neoantigens in the cancer genome in acute myeloid leukemia (AML) compared to other cancers, possibly due to a low number of mutations and relatively low strength of association with exposure to common mutagens such as tobacco smoke, alcohol, and UV radiation[27,28]. Rather, AML is thought to result from accumulation of random mutations in hematopoietic stem cells that confer a selective advantage in the bone marrow microenvironment, and as such the ratio of driver to passenger mutations in AML is higher than that in epithelial tumors more associated with mutagen exposure (e.g. melanoma, lung cancer, bladder cancer). However, mutations in spliceosomal genes and genome-wide aberrant splicing events are more common in patients with AML compared to epithelial malignancies[29,30]. RNA splicing is a process that removes non-coding introns and concatenate protein coding exons in premature messenger RNA molecules to form mature messenger RNA molecules ready for translation into proteins. RNA splicing is catalyzed by the spliceosome which is a molecular machinery that consists of small nuclear RNAs (snRNA) and associated protein factors. U1, U2, U4, U5, and U6 are the major snRNAs that form the spliceosome. Multiple mutations involving the formation and regulation of spliceosome are implicated in

AML[31]. For example, mutations in DDX41 gene that affect spliceosome assembly, mutations in SRSF2 that affect mRNA stabilization and binding of U1 and U2 snRNAs to 5' splice site and branch site, mutations in U2AF1 and SF3B1 that affect binding of U2 snRNAs to branch site have been reported in many AML cases[31]. Neoantigens could arise from novel peptides sequences derived from neo-junctions directly as well as frameshifts due to neo-junctions. Targeting neoantigens derived from tumor specific alternative splicing events could expand the therapeutic target space in patients with AML and other type of tumors with a low frequency of neoantigens derived from single nucleotide variations or indels. However, none of the published methods to detect neoantigens from genomics data is able to predict splice variant neoantigens. Transcript isoform assembly methods that have been used to evaluate splicing in large tumor datasets that could be used to predict splice variant neoantigens are not specific in the prediction of splice variant peptides: splice variant peptide producing transcript isoforms with multiple mutated features were not tested to exist in the tumor transcriptome in these methods[32,33]. Critically, they require additional somatic variant information from whole exome sequencing data[32,33]. In chapter 2, we present the NeoSplice method, which perform **accurate prediction of splice variant neoantigens without requiring additional DNA sequencing data.**

Despite the ability of neoantigen therapeutic vaccines to promote tumor-specific T-cell responses in a number of pre-clinical models, clinical efficacy has yet to be demonstrated due to low patient numbers and short follow-up time in these phase I trials[7,8,34–36]. Though early in clinical development, there is tremendous excitement around therapeutic neoantigen vaccination with 24 clinical trials with plans to treat over a thousand patients registered at ClinicalTrials.gov. Challenges for translation of neoantigen therapies include manufacture of neoantigen peptide, delivery of neoantigen, identification of neoantigens derived from all genomic sources, and filtering of clinically relevant epitopes including by predicting *in vivo* immunogenicity. Chapter 2 of this thesis will describe our work to broaden neoantigen prediction by development of NeoSplice. We will focus on prediction of neoantigen immunogenicity in chapter 3, as an accurate selection of clinically relevant neoantigens could greatly reduce the cost of synthesizing large number of neoantigen candidates. The ability of neoantigen to induce anti-tumor immune response depends on the extent of T cells recognition of the neoantigen presented by major histocompatibility complex (MHC) proteins[37]. As described above, many neoantigen prediction algorithms ignored factors that can affect immunogenicity and rely heavily on peptide/MHC binding affinity predictions to rank epitopes[17–25]. Previous studies (Table 1.1) have shown that amino acid characteristics such as polarity, charge, hydrophobicity, and amino acid size can influence antigen immunogenicity[6,16,38–41]. For example, Chowell et al. discovered that immunogenic epitopes tend to contain more hydrophobic amino acids at T-cell receptor contact positions using an MHC class I peptide immunogenicity dataset[6,42]. One possible reason for this is that hydrophobic region will enhance peptide degradation and thus presentation[42,43]. Prior work by Cole et al. also indicated that alteration of antigen anchor residues which are residues bind to pocket of MHC can alter binding affinity of TCR to antigen using plasmon resonance and peptide–MHC tetramer binding experiments[40]. Moreover, the polarity and charge of amino acids in the center of a peptide can affect TCR recognition of peptide-MHC complex since it requires more energy and a higher structural precision of an engaging TCR to prevent the amino acid from interacting with water[41]. Hence, the development of an algorithm to predict the immunogenicity of neoantigens would be valuable for the selection of predicted neoantigens for clinical application. In chapter 3, we describe the development and validation of such an algorithm to predict neoantigen peptide immunogenicity.



Adopted from Paul Armistead

Figure 1.1 Tumor neoantigens induce T cell mediated cytotoxicity.

1.3 Immunogenomic characterization of the tumor immune microenvironment

Next-generation sequencing technology can be used to study the tumor immune microenvironment as well as genomics of tumor cells. Sequencing data can help elucidate molecular profiles and biomarkers that are associated with improved immune responses and patient outcomes and also guide the development of novel immunotherapies. For example, Hugo et al. identified an innately resistant transcriptional signature to PD-1 (termed IPRES) in melanoma, comprised of genes with functions related to angiogenesis, wound healing, regulation of mesenchymal transition, and cell adhesion using RNA-seq data[44]. Iglesia et al. found that high expression of T-cell and B-cell signatures are associated with improved survival in many tumor types including breast, lung, and melanoma while expression of Macrophage signatures and B-cell signatures predicted worse survival in GBM and renal tumors respectively using The Cancer Genome Atlas (TCGA) RNA-seq data from 11 tumor types[45]. As part of PhD training in bioinformatics analysis to complement training in methods development, we performed comprehensive genomic and immune characterizations of bladder tumors and triple-negative breast cancer brain metastases. Our goal in each of these studies was to understand the tumor immune microenvironment features that distinguish cancer subtypes, as well as to associate these with tissue of origin and survival outcomes. Results from these studies are shown in Chapter 4.

CHAPTER 2 NEOSPLICE: A BIOINFORMATICS METHOD FOR PREDICTION OF SPLICE VARIANT NEOANTIGENS

2.1 Introduction

Recent advances in immunotherapy drugs such as CTLA-4 and PD-1 blocking antibodies (called immune checkpoint inhibitors) have highlighted the capacity of induced anti-tumor immune responses to yield improved survival for cancer patients[4,46,47]. In parallel with the development of immune checkpoint inhibitors which broadly reverse effector T cell suppression, multiple groups are working on tumor-specific vaccines that selectively stimulate anti-tumor T cells to expand and kill tumor cells. Tumor- specific peptide epitopes that are generated from mutated genes and presented on cell surface MHC molecules, known as neoantigens, are attractive targets for therapeutic vaccination given the lack of central tolerance and corresponding presence of endogenous T cells that recognize them[4]. Recent studies have shown that neoantigen targeted therapy can yield improved anti-tumor immune responses in patients with advanced and metastatic tumors[7,8].

The vast majority of neoantigens in tumors are private rather than shared among patients, therefore it is essential for neoantigens to be predicted for each patient in a genome wide fashion for use in therapeutic vaccination[4]. Currently most available neoantigen prediction methods focus on predicting neoantigens derived from missense mutations or indels[6,9–16]. Typical neoantigen prediction pipelines involve identifying highly expressed tumor-specific mutations using whole-exome sequencing (WES) and RNA sequencing (RNA-seq) data of matched tumor-

and normal-cell samples and variant calling bioinformatics tools, and prioritizing mutation derived neo-epitopes using their predicted binding affinities to HLA I and HLA II molecules expressed by the patient[5–16,28].

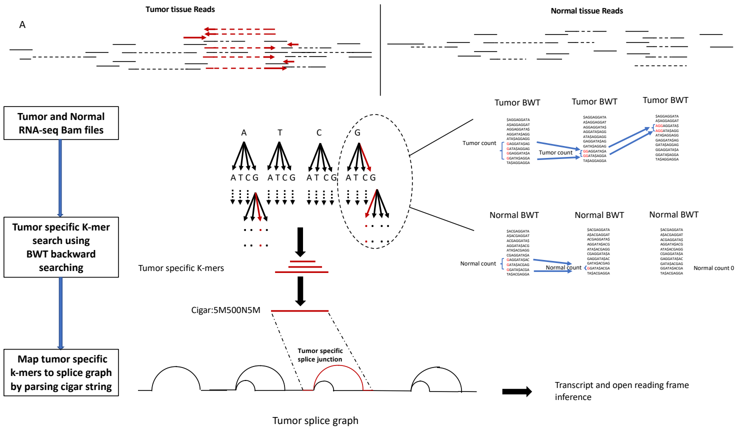
In acute myeloid leukemia (AML), there are markedly fewer mutations and predicted neoantigens in the cancer genome than in most other adult cancers, so it is less feasible to target neoantigens derived from missense mutations and indels in AML[27,28]. However, mutations in spliceosomal genes and genome-wide aberrant splicing events are common in patients with AML[29,30]. The prediction of neoantigens from tumor-specific splice variations presents significant computational challenges. Recently, two splice variant neoantigen prediction methods were reported[32,33]. However, these tools are also not specific in the prediction of splice variant peptides: splice variant peptide producing transcript isoforms with multiple mutated features were not tested to exist in the tumor transcriptome in these methods[32,33]. Critically, they require additional somatic variant information from whole exome sequencing data[32,33]. We report the development of NeoSplice for specific and comprehensive prediction of splice variant neoantigens using tumor and matched normal RNA-seq data, without requiring matched DNA sequencing data.

2.2 Methods

The NeoSplice Method.

Multiple steps are needed to identify a novel splice that occurs specifically in tumor cell transcripts whose translation will result in a neopeptide that can be targeted by T cells (Figure 2.1).

1. Using one RNA-seq dataset **T** of tumor cells and one RNA-seq dataset **N** of normal cells, identify tumor-specific sequences present abundantly in the transcriptome of the tumor cell, but rarely if at all present in the normal cells.
2. Generate the splice graph **G** from the tumor cell RNA-seq data and locate tumor specific k-mers that correspond to a novel splice. Use annotations to determine whether the novel splice lies within in a protein coding region of **G**. Use paired-end reads (or long read sequencing) and annotations to link the tumor-specific splice to the start codon to determine whether the splice results in a frame shift.
3. Use further methods including an MHC binding affinity prediction tool and immunogenicity prediction tool to determine whether the coding sequence of the novel tumor transcript will yield a neopeptide detectable on the cell surface by T cells that will then be activated via signaling through the T cell receptor.



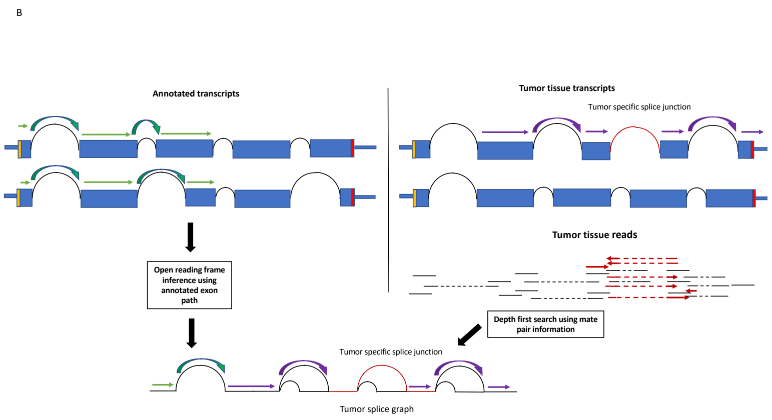

B
Annotated transcripts
Tumor tissue transcripts
Tumor specific splice junction
Tumor tissue reads
Open reading frame inference using annotated exon path
Depth first search using mate pair information
Tumor specific splice junction
Tumor splice graph

Figure 2.1 NeoSplice Overview: (A). The multi-string BWT tool based on a variant of the Burrows Wheeler transform (BWT) builds the multi-string BWT data structure for tumor and normal RNA-seq bam files. A depth-first search process operating in lockstep on the tumor and normal BWT data structures can find all tumor-specific k-mers. (B) The splice graphs are constructed from the tumor RNA-seq bam files. Graph traversal infers the tumor specific splice junction containing partial transcript isoforms within an open reading frame by taking advantage of using paired-end read information and annotated transcript information.

Step 1a. Construction of the BWT data structure and K-mer counting method. To efficiently determine the number of occurrences of an arbitrary sequence s in an RNA-seq dataset **D**, we organize the reads from **D** into a suffix array[48], a sorted array of all suffixes of every read in **D**. The suffix array enables the number of occurrences of an arbitrary sequence T to be determined in time O(|t|) where |t| is the sequence length. The time to create a suffix array is linear in the number of reads.

As a suffix array for an RNA-seq dataset would be prohibitively large, we use the Burrows-Wheeler Transform (BWT) method. It provides the same functionality using a compressed representation of the suffix array. We used the multi-string BWT tool (MSBWT 0.3.0)[49] to build separate multi-string BWT indexes for the tumor and normal RNA-seq reads. The RNA-seq reads were extracted from aligned bam files and soft clipped portions of reads were removed. All reads were represented on the same reference strand. The FM index[50] of the

BWT consists of counts information for the BWT data structure and rapidly locates the first index and last index of all occurrences of a given sequence (all occurrences appear consecutively in the suffix array). The number of occurrences of the sequence is the difference between the first and last index.

Step 1b. Tumor specific K-mer searching algorithm. A sequence whose occurrence count in the tumor RNA-seq **T** exceeds a threshold T_min (i.e. occurs sufficiently frequently) while its occurrence in the normal RNA-seq **N** is less than a maximum count N_max (i.e. is sufficiently rare) is a considered tumor specific k-mer. A typical value for T_min is 20 and a typical value for N_max is 3. Tumor-specific sequences are discovered by a depth-first search process operating in parallel on the tumor and normal BWT data structures.

The search is initialized starting from a 1-mer, e.g. "A". Clearly the sequence "A" will have a huge number of occurrences in both T and N, so will not satisfy the rarity condition in **N**. The search is then refined by adding a nucleotide from {A, T, C, G} in front of the current search sequence and recursively applying the search for the extended k-mer. The recursive search will either yield one or more tumor specific k-mers, or will backtrack from the refinement because the occurrence count of the search k-mer falls below T_min. In this fashion all possible tumor specific k-mers will be found. We may parallelize the algorithm by starting a separate search from each different 1-mer (4-way) or 2-mer (16 way) or 3-mer (64 way), etc. A sublinear speed up was observed when testing on typical illumina fastq files using 1 thread, 4-theads, and 16-threads of parallelism.

Tumor-specific k-mers of variable length are returned after the search terminates. An Aho–Corasick algorithm (pyahocorasick 1.4.0)[51] was used to search for the reads that contain tumor specific k-mers in the tumor RNA-seq bam file. This method runs in time linear in the size of the bam file. For each occurrence, the k-mer containing portion of the read along with corresponding quality scores and Cigar strings is written to a new bam file.

Step 2a. Splice graph construction. The splice graph is a weighted, directed graph. Nodes in the splice graph represent genomic coordinates. Edges in the splice graph represent transcribed intervals (exons) or splices. In addition, splice graphs may carry additional information about insertions, deletions, and single nucleotide variants, as well as annotations like translation start sites for coding regions. The splice graph was constructed using an RNA-seq BAM file and GENCODE GFF3 file[52]. The splice junction, insertion, and deletion information were detected from CIGAR strings of reads in the BAM file using pysam 0.14.1[53]. The exon and single-nucleotide polymorphism information were retrieved by examining aligned reads at each genomic coordinate using pysam 0.14.1[53]. Annotated translation initiation site information was retrieved from the GENCODE GFF3 file[52].

Step 2b. Splice graph traversal algorithm for predicting splice variant neoantigens. Tumor specific splice junctions were identified by taking the intersection of splice junctions identified within tumor specific k-mers and splice junctions found by RNA-seq quantification to be highly expressed in tumor but lowly expressed in normal RNA-seq data (with tumor expression threshold 20 and normal expression thresholds 2). Tumor-specific k-mers that include tumor specific splice junctions were mapped to splice graph using cigar strings. If any end of the k-mer was mapped inside an exon edge, the k-mer graph path included the whole exon edge.

For each tumor-specific k-mer graph path supported by a sufficient number of k-mer containing reads, a depth first search algorithm was used for graph traversal upstream and downstream of the tumor-specific k-mer graph path. The depth first search was restricted to edges supported by sufficient number of paired-end reads that contain a tumor-specific k-mer. If the depth first search did not reach an annotated transcript's start codon, reference transcripts that cover the k-mer graph path were used for open reading frame inference. Specifically, if an annotated transcript is supported by a path in the splice graph, graph traversal will start from the annotated start codon, follow the annotated exon and splice junction path, and stop when it reaches the upstream depth first search stop position.

The transcript sequence identified by depth first search was then concatenated with the tumor specific k-mer sequence and translated into 8-11 mer peptides for MHC class I neoantigen prediction and 15 mer peptides MHC class II neoantigen prediction. Binding affinity to MHC molecules expressed by the tumor for in-silico generated peptides was predicted using NetMHCpan 4.0[20]. The reference peptidome was generated by translating protein coding transcripts present in GENCODE GFF3 file[52]. Peptides with an IC50 value of less than 500 nM for at least 1 MHC allele and not present in the reference peptidome were considered predicted neoantigens. The NeoSplice software is available on the GitLab page https://sc.unc.edu/benjamin-vincent-lab/tools/NeoSplice. The Docker image for NeoSplice is available on https://cloud.docker.com/u/max111/repository/docker/max111/neosplice

Performance comparison using simulated reads.

1. Read simulation. Protein coding regions of annotated transcripts were modified to generate tumor specific splice junctions resulting in novel transcripts. For each chromosome, 5 protein coding genes were selected. One tumor specific splice junction per selected gene was then simulated by randomly choosing a combination

of donor splice site and acceptor splice site that does not exist in any annotated transcripts for that gene and forms an exon skipping splice junction, a partial exon loss splice junction, or a partial intron gain splice junction compared with the set of reference transcripts. The selected splice junction was then applied to all annotated transcripts of that gene. Genes with no possible tumor specific splice junctions were excluded. A random proportion of exon skipping, partial exon loss, and partial intron retention splice junctions were simulated. These events cover the splicing abnormalities observed in AML. For example, splice acceptor site changes and 3' splicing changes due to mutations in SRSF2 gene and U2AF1 gene are included in partial exon loss events and partial intron retention events. Data sets with exon skipping only splice junctions, partial exon loss only splice junctions, and partial intron gain only splice junctions were also simulated. Reference transcripts corresponding to tumor transcripts were written to a separate GTF file and used as normal data. RNA-seq simulator Polyester 1.9.7[54] was used to simulate reads using the generated GTF file as input. 1000 100bp paired-end reads were simulated for each transcript in the GTF file. 20 random bootstrap Tumor-normal pair datasets were simulated in order to obtain more general results. Simulation code is available on GitLab page https://sc.unc.edu/benjamin-vincent-lab/tools/NeoSplice.

2. Read alignment and transcript inference. A STAR 2.7.0e and Neosplice pipeline, TopHat 2.1.1 and Cufflinks 2.2.1 pipeline, HISAT2 2.1.0 and StringTie 1.3.3 pipeline, and MapSplice 2.2.1, Trinity 2.8.5, and GMAP 2018-05-30 pipeline were used for alignment and transcript inference of the simulated data[55–62]. The hg19 reference genome was used for all pipelines. Gencode GFF3 file (gencode.v19.annotation.gff3)

was used by NeoSplice. Methods and parameters are provided in Table 2.1. A tumor threshold of T_min = 16 and normal threshold of N_max = 3 was used in the tumor-specific k-mer searching stage. Tumor specific k-mers length were restricted to be at most 90% of read length. Tumor specific k-mers that contained splice junction found by RNA-seq alignment to be supported by at least 20 reads in tumor RNA-seq data but supported by less than 3 reads in normal RNA-seq data were considered. K-mer graph paths supported by less than 10 tumor specific k-mer including reads were filtered. Only exon edges and splice junction edges were traversed using NeoSplice for ease of validation.

3. Performance comparison. Simulated tumor transcripts that contained tumor specific splice junctions with 10 nucleotides reference sequence upstream and downstream of the splice junction that occurred more than 15 times in simulated tumor RNA-seq data and less than 4 times in simulated normal RNA-seq data were used as ground truth for performance comparison. For each simulated tumor specific splice junction, transcript features of neoantigen producing regions including exon coordinates, open reading frame of last nucleotide before splice junction, and strandedness of inferred transcripts were compared between inferred transcripts and ground truth tumor transcripts for validation. The neoantigen producing region was defined as at most 50 nucleotides upstream and downstream of tumor specific splice junction. Open reading frames for Trinity, Cufflinks, and StringTie transcripts were inferred using annotated start codons. Sensitivity and precision were calculated for transcript isoforms inferred by NeoSplice, Trinity, Cufflinks, and StringTie.

Bioinformatics tools	Parameters
MapSplice 2.2.1	Default parameters
HISAT2 2.1.0	-f
StringTie 1.3.3	Default parameters
Trinity 2.8.5	--genome_guided_bam, --genome_guided_max_intron 10000
GMAP 2018-05-30	-B 5, -n 2, --gff3-add-separators=0, -f 2
TopHat 2.1.1	--library-type fr-unstranded
Cufflinks 2.2.1	--library-type fr-unstranded
STAR 2.7.0e	--outSAMunmapped Within, --twopassMode Basic, --outFilterScoreMinOverLread .45, --outFilterMatchNminOverLread .45, --outSAMattrRGline ID:1 LB:LB PL:PL SM:SIM PU:PU

Table 2.1 Parameters used in transcript inference pipelines for simulated data. Default parameters were used for parameters not listed

Performance comparison using long read sequencing data:

1. Data generation. Short read RNA-seq data for U937-A2 cell line and FACS-sorted CD34+ hematopoietic stem cells derived from healthy donor bone marrow biopsy specimens were generated by illumina sequencing (HiSeq2500). The long-read RNA-seq data for 15 genes predicted to have tumor-specific novel splicing events were

generated using PacBio RS and Oxford Nanopore MinION sequencing platforms. CD34+ hematopoietic stem cell RNA-seq data were used as reference normal data.

2. Read alignment and transcript inference. A STAR 2.7.0e, ABRA2 2.19, and NeoSplice pipeline, TopHat 2.1.1 and Cufflinks 2.2.1 pipeline, HISAT2 2.1.0 and StringTie 1.3.3 pipeline, and MapSplice 2.2.1, Trinity 2.8.5, and GMAP 2018-05-30 pipeline were used for alignment and transcript inference of the short read RNA-seq data[55–63]. Methods and parameters are provided in Table 2.2. The hg19 reference genome was used for all pipelines. Gencode GFF3 file (gencode.v19.annotation.gff3) was used by NeoSplice. In the tumor-specific k-mer searching stage, tumor thresholds of T_min = 21 and a normal threshold of N_max = 3 were used for U937 cell line. Tumor specific k-mers length were restricted to be at most 90% of read length. Tumor specific k-mer graph paths that contain splice junctions found by RNA-seq alignment to be supported by at least 20 reads in tumor RNA-seq data but supported by less than 3 reads in normal RNA-seq data were considered. Tumor specific k-mer graph paths supported by less than 10 tumor specific k-mer including reads were filtered in U937 cell line. Only exon edges and splice junction edges were traversed using NeoSplice for ease of validation. Long reads were aligned using the EDGAR tool developed by the Prins lab[64].

3. Performance comparison. Consensus tumor specific splice junctions determined by all four aligners (TopHat 2.1.1, STAR 2.7.0e, HISAT2 2.1.0, and MapSplice 2.2.1) to have at least 20 splice junction supporting reads in tumor and less than 3 splice junction supporting reads in normal and also discovered by tumor specific k-mer search were identified. Long read transcripts that contain consensus tumor specific

splice junctions were used as ground truth for performance comparison. Transcripts inferred by NeoSplice, Cufflinks, StringTie, and Trinity that contain consensus tumor specific splice junctions were validated for performance comparison. For each tumor specific splice junction, transcript features of neoantigen producing regions including exon coordinates, open reading frame of last nucleotide before splice junction, and strandedness were compared between inferred transcripts and ground truth tumor transcripts for validation. The neoantigen producing region was defined as above. Open reading frames for ground-truth long-read transcripts, Trinity, Cufflinks, and StringTie transcripts were inferred using annotated start codons. Sensitivity and precision were calculated for transcript isoforms as well as neoantigen producing regions inferred by NeoSplice, Trinity, Cufflinks, and StringTie.

Bioinformatics tools	Parameters
MapSplice 2.2.1	--qual-scale phred33 (CD34+ samples only)
HISAT2 2.1.0	-q, --phred64 (U937 sample only)
StringTie 1.3.3	Default parameters
Trinity 2.8.5	--genome_guided_bam, --genome_guided_max_intron 10000
GMAP 2018-05-30	-B 5, -n 2, --gff3-add-separators=0, -f 2
TopHat 2.1.1	--library-type fr-unstranded, --phred64-quals (U937 sample only)
Cufflinks 2.2.1	--library-type fr-unstranded
STAR 2.7.0e	--outSAMunmapped Within, --twopassMode Basic, --outFilterScoreMinOverLread .45, --

	outFilterMatchNminOverLread .45, --outSAMattrRGline ID:1 LB:LB PL:PL SM:SIM PU:PU
ABRA2 2.19	--junctions bam, --cl 1, --dist 500000, --sua

Table 2.2 Parameters used in transcript inference pipelines for short read RNA-seq data. Default parameters were used for parameters not listed.

Splice variant neoantigen prediction for U937 and K562 cell lines.

STAR 2.7.0e and ABRA2 2.19 were used for RNA-seq read alignment[55,63]. CD34+ hematopoietic stem cell RNA-seq data were used as reference normal data. NeoSplice was run on U937 and K562 cell lines RNA-seq bam file to predict splice variant neoantigens. The Hg19 reference genome was used for all pipelines. In the tumor-specific k-mer searching stage, tumor thresholds of T_min = 21 and T_min = 26 were used for U937 cell line and K562 cell line respectively, and a normal threshold of N_max = 3 was used for both U937 cell line and K562 cell line. Higher T_min was used for K562 cell line because the RNA-seq coverage of K562 cell line is higher. Tumor specific k-mers length were restricted to be at most 90% of read length. Tumor specific k-mer graph paths that contain splice junctions found by RNA-seq alignment to be supported by at least 20 reads in tumor RNA-seq data but supported by less than 3 reads in normal RNA-seq data were considered. Tumor specific k-mer graph paths supported by less than 10 and 15 tumor specific k-mer including reads were filtered in U937 and K562 cell line respectively. Peptides that were also present in hg19 reference peptidome generated by translating reference protein coding transcripts in Gencode GFF3 file (gencode.v19.annotation.gff3) were filtered.

Splice variant neoantigen prediction for Pan-TCGA data.

Hg38 STAR two-pass mode aligned RNA-seq bam files for TCGA tumor and adjacent normal pairs were downloaded from GDC (https://portal.gdc.cancer.gov) (N= 503). Hg38 Star two-pass mode aligned RNA-seq bam files of TCGA-LAML cohort were also downloaded from GDC (N=136). CD34+ hematopoietic stem cell RNA-seq data were used as reference normal data for TCGA-LAML cohort. The NeoSplice algorithm was run on pan-TCGA data to predict splice variant neoantigens. Gencode GFF3 file (gencode.v22.annotation.gff3) was used by NeoSplice. A tumor threshold of T_min = 21 and normal threshold of N_max = 3 was used in the tumor-specific k-mer searching stage. Tumor specific k-mers length were restricted to be at most 90% of read length. Tumor specific k-mer graph paths that contain splice junctions found by RNA-seq aligner to be supported by at least 20 reads in tumor RNA-seq data but supported by less than 3 reads in normal RNA-seq data were considered. k-mer graph paths supported by less than 10 tumor specific k-mer including reads were excluded. Peptides that were also present in hg38 reference peptidome generated by translating reference protein coding transcripts in Gencode GFF3 file (gencode.v22.annotation.gff3) were filtered. SNV derived neoantigens for TCGA LAML samples were predicted using method described in Chapter 4. INDEL derived neoantigens for TCGA LAML samples were not included in the analysis because a lack of data availability. INDEL derived neoantigen counts and SNV derived neoantigen counts for other TCGA samples were obtained from Thorsson et al[65]. Immune signatures for TCGA samples were calculated using method described in Chapter 4. Spearman correlation coefficients as well as benjamini-hochberg adjusted p-values were calculated for immune signatures and predicted splice variant neoantigen counts for TCGA LAML samples.

2.3 Results

Performance comparison using simulated RNA-seq data.

The main computational challenge for prediction of splice variant neoantigens is to infer tumor specific splice junction containing transcripts. To assess the performance of NeoSplice, we simulated 20 RNA-seq datasets to include novel, unannotated splice junctions for each of the splice junction types including exon skipping splice junctions, partial intron gain splice junctions, partial exon loss splice junction, and mixed random types of splice junctions. Five protein coding genes were selected for each chromosome. One tumor specific splice junction per was selected for each gene if tumor specific splice junction is supported. Genes with no possible tumor specific splice junctions were excluded. We then used Polyester to simulated 100bp paired-end reads from the generated transcript GTF files. 8000 reads were simulated per transcript. Performance comparison results were compared with StringTie, Trinity, and Cufflinks since they are the most widely used transcript inference tools. The inferred transcripts were validated by assessing prediction of the neoantigen producing region (defined in the method section). The comprehensive simulation result demonstrated that NeoSplice performed better in the simulated data that contains mixture of tumor specific splice junction types (Figure 2.2). NeoSplice performed better in terms of sensitivity for mixed type of splice junctions, partial intron gain splice junctions, and exon skipping splice junctions, and performed better in terms of precision for mixed type of splice junctions, partial intron gain splice junctions when compared with StringTie, Trinity, and Cufflinks. Notably, the sensitivity values and precision values of NeoSplice are almost always greater than 0.8 for all types of splice junctions while the sensitivity values and precision values for other transcript inference tools generally lower and varied for different types of splice junctions.

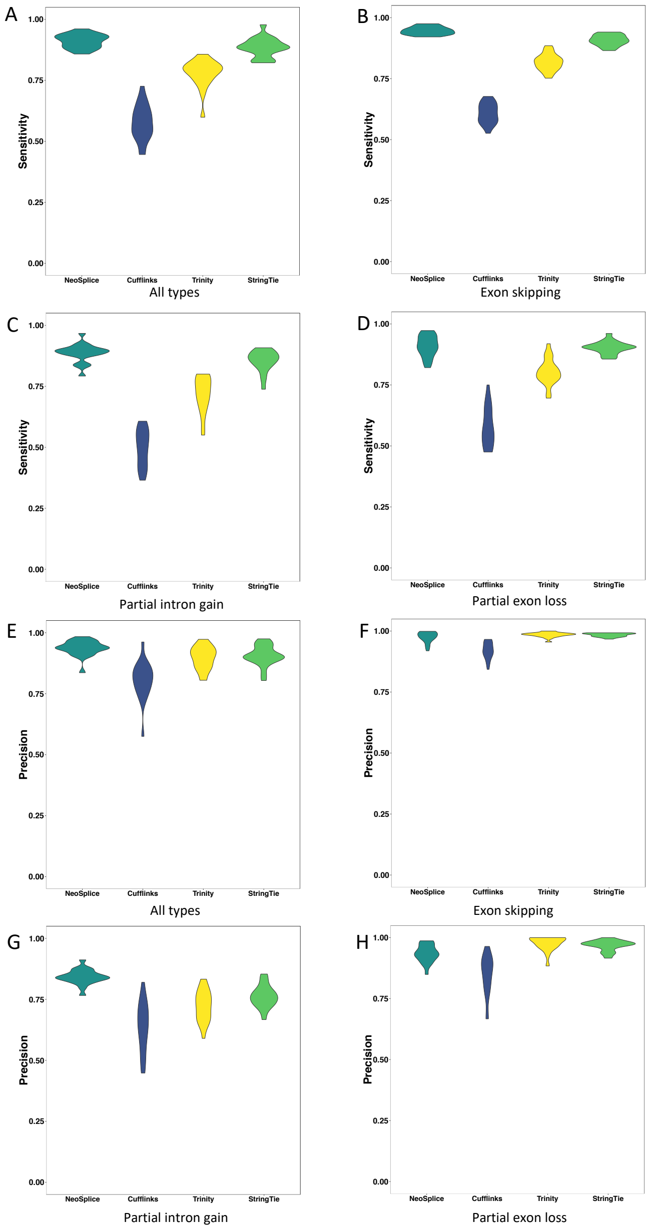


Figure 2.2 Performance comparison on simulated data. The precision values and sensitivity values for predicting splice variant neoantigen producing regions (50 nucleotides upstream and downstream of the tumor specific splice junction) are shown in the box plot for the NeoSplice (shown in Cyan), Cufflinks (shown in Blue), Trinity (shown in yellow), and StringTie (shown in Green). (A)-(D), The sensitivity values for simulated data sets with mixed splice junction type, exon skipping splice junction type, partial intron gain splice junction type, and partial exon loss splice junction type are shown. (E)-(H), The precision values for simulated data sets with mixed splice junction type, exon skipping splice junction type, partial intron gain splice junction type, and partial exon loss splice junction type are shown.

Performance comparison using short read RNA-seq data and long read RNA-seq data.

We also evaluated the performance of NeoSplice on real RNA-seq data. We generated short-read RNA-seq data of the U937-A2 cell line which is a human acute myeloid leukemia (AML) cell line which has been engineered to express HLA-A*0201, the most common HLA allele in the US population[66]. Long-read RNA-seq data that can be used for validation of transcript identification of 15 genes predicted to have tumor-specific novel splicing events were generated using PacBio and Oxford Nanopore sequencing platforms. NeoSplice, StringTie, Trinity, and Cufflinks made predictions for 7 genes. Since many long reads are highly similar due to coverage difference during sequencing, the count of full-length inferred transcripts, full-length long reads, as well as just the neoantigen producing regions of full-length inferred transcripts and full-length long reads are shown (Figure 2.3). The performance comparison results are similar to the results seen for the simulated data. NeoSplice inferred more splice variant neoantigen producing regions and had higher positive prediction values compared with StringTie, Trinity, and Cufflinks.

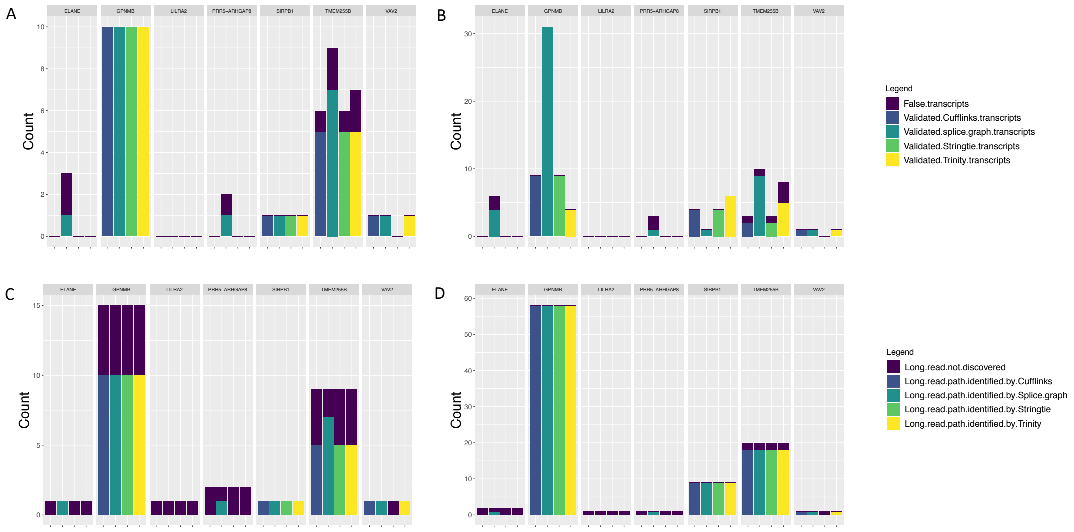


Figure 2.3 Performance comparison on experimental RNA-seq data in U937-A2 cell line. The counts of total and validated neoantigen producing regions of full-length inferred transcripts and full-length long (A, C), as well as just full-length inferred transcripts, full-length long reads (B, D) are shown for NeoSplice method, the TopHat2 + Cufflinks method, the MapSplice + Trinity + GMAP method, and Hisat2 + StringTie method.

Computation resource utilization of NeoSplice.

To measure CPU and memory requirements of NeoSplice, we measured the running time and memory consumption of NeoSplice using three simulated transcript sets generated in the previous performance comparison section. 1000, 4000, or 8000 reads per transcript were simulated for three transcript sets. The maximum Resident-Shared Size (RSS) and elapsed runtime were recorded using 'sacct' command. Runtime and memory consumption increase linearly in proportion to number of reads simulated per transcript (Fig 2.4), which is expected for NeoSplice.

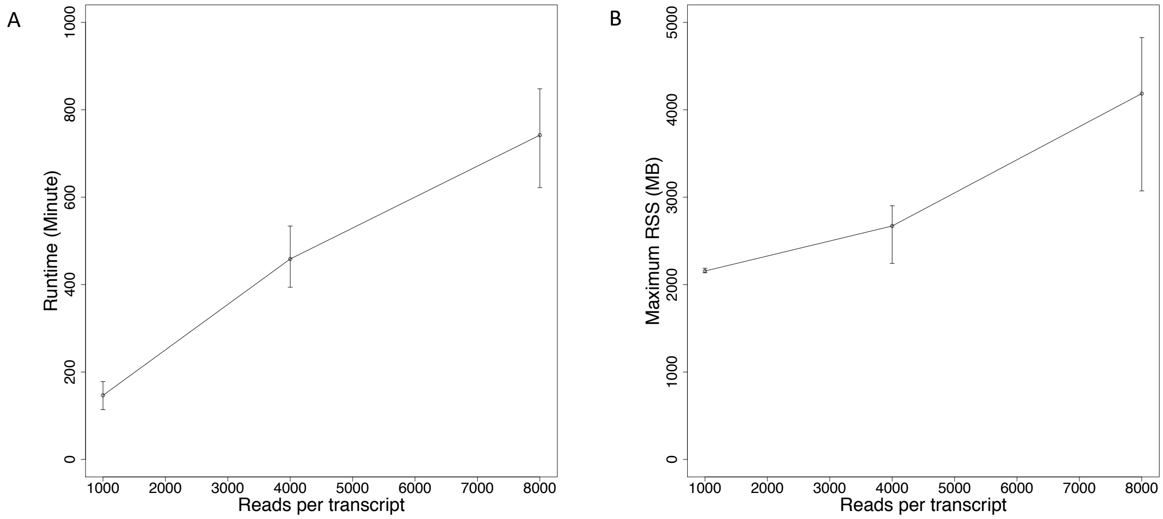


Figure 2.4 Computation resource utilization of NeoSplice. (A) Average time to run NeoSplice on simulated data. Error bars represent the maximum and minimum times across 3 simulated samples. Runtime is defined as the elapsed time as reported by the 'sacct' command as measured on an Intel Xenon ES-E5620 2.4GHz CPU and Intel Xenon ES-E5520 2.27GHz CPU. (B) Average memory required to run NeoSplice on simulated data. RSS is amount of memory requested by NeoSplice from the operating system as reported by the 'sacct' command.

Pan-TCGA analysis of splice variant neoantigens.

We hypothesized that splice variant neoantigens may be common in many types of cancer. To access the distribution of splice variant neoantigens in different types of cancer, NeoSplice was run on pan-TCGA data with available tumor and matched normal pairs and also TCGA LAML samples using CD34+ hematopoietic stem cell as reference normal. There is abundant predicted splice variant neoantigens for all type of cancers and the splice variant neoantigen burden is generally comparable with SNV derived neoantigen burden. Notably, the splice variant neoantigen count is much higher than SNV derived neoantigen count in AML, consistent with our hypothesis that splice variant neoantigen will expand therapeutic target space in patients with AML (Figure 2.5A). We further investigated the association of immune signatures in TCGA LAML samples with predicted splice variant neoantigen burden. 17 immune signatures including B_cells_naive immune signature, Monocytes immune signature, CD68 immune signature, and IFNG_score immune signature are found to be significantly associated with predicted splice variant neoantigen count in TCGA LAML samples (Figure 2.5B-E). We also assessed the frequency of neoantigens producing tumor splice junctions and predicted splice variant neoantigens across TCGA samples (Figure 2.6). Many neoantigens producing tumor splice junctions and predicted splice variant neoantigens are shared in about 10% of all TCGA samples (N=639) included in the analysis, suggesting the feasibility to develop an off-the-shelf splice variant neoantigen therapeutic.

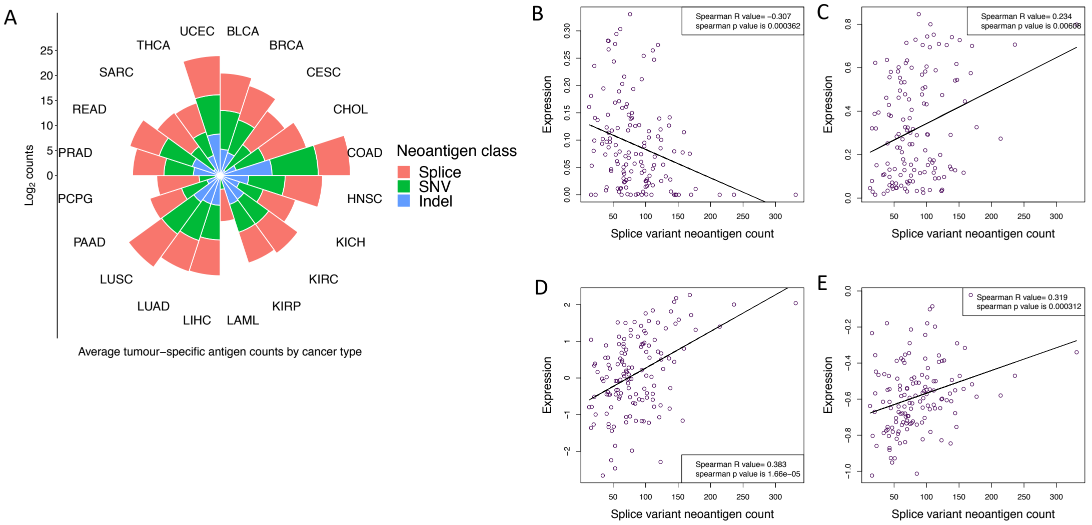


Figure 2.5 Pan-TCGA analysis of splice variant neoantigens. (A) Radial plot showing average Log$_2$ count of predicted splice variant neoantigen (pink), SNV-derived neoantigen (green), and INDEL derived neoantigen (blue) for 20 cancer types in TCGA data. SNV and INDEL neoantigens are derived from Thorsson et al. (Immunity, 2018). INDEL neoantigens for LAML samples are not shown. (B-E) Scatter plots showing correlation of predicted splice variant neoantigen count with expression level of B_cells_naive (B), Monocytes (C), CD68 (D), and IFNG_score (E) immune signatures.

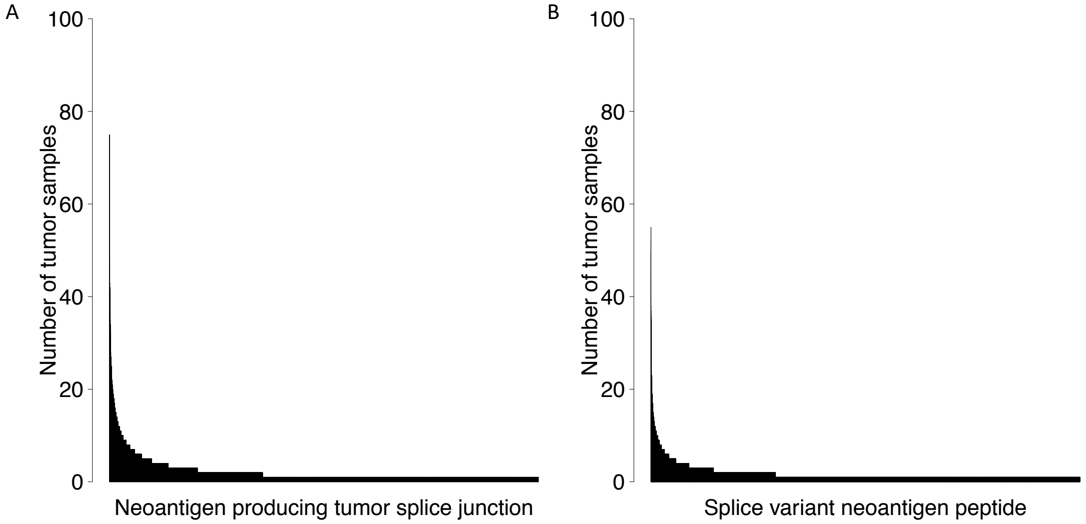


Figure 2.6 Frequency of neoantigen producing tumor splice junctions and predicted splice variant neoantigens across TCGA samples. (A) Number of TCGA samples expressing shared neoantigen producing tumor splice junctions. (B) Number of TCGA samples expressing shared splice variant neoantigens.

2.4 Discussion

In summary, we developed NeoSplice, a bioinformatics tool that predicts splice variant neoantigens from RNA-seq data, without need of additional DNA sequencing data. Although currently two splice variant neoantigen prediction tools have been reported, neither of them utilizes transcript level information in the RNA-seq data and both require DNA sequencing data along with RNA-seq data. The performance comparison results from both the simulated data and U937-A2 long read RNA-seq data indicate that predictions made by NeoSplice are highly comprehensive and specific. Moreover, NeoSplice is a time and space efficient algorithm. Our novel BWT based tumor specific k-mer searching algorithm does not need to examine all possible k-mers in tumor and normal RNA-seq data, and the BWT method required less memory compared to hash-table based methods that usually require large memory to build an index. The splice graph building step requires time proportional to number of reads in the data set and splice variant peptide prediction step requires time proportional to number of edges and nodes in the graph. Moreover, NeoSplice can be expanded to predict neoantigens derived from other types of mutations such as SNVs, indels, and gene fusions as identification of tumor-specific k-mers should capture each of these and the splice graph traversal will identify the transcripts that contain tumor specific mutations. Apart from AML, splice variant neoantigens are also predicted to exist in many types of cancer as shown by Kahles et al, Jayasinghe et al[32,33] as well as our pan-TCGA splice variant neoantigens analysis. Unlike SNV-derived neoantigens, many splice variant neoantigens are found to be shared among different tumors in our pan-TCGA analysis, suggesting the feasibility to develop an off-the-shelf splice variant neoantigen therapeutic. Many immune signatures also correlate with predicted splice variant neoantigen count in TCGA LAML data. We expect that NeoSplice will expand the neoantigen therapeutic target space for cancer patients. Currently there are over 20 clinical trials of therapeutic neoantigen vaccines in cancer registered on *ClinicalTrials.gov*. Multiple companies have others, along with neoantigen specific adoptive cellular therapy approaches, in development. Accurate identification of splice variant neoantigens will become more important as additional neoantigen specific therapeutic platforms enter clinical trials, especially for those tumors like AML where neoantigens derived from SNVs and Indels are rare.

CHAPTER 3 MACHINE-LEARNING PREDICTION OF TUMOR ANTIGEN IMMUNOGENICITY IN THE SELECTION OF THERAPEUTIC EPITOPES[1]

3.1 Introduction

T cells can affect antitumor immune responses through recognition of tumor-specific antigens (TSAs) presented by major histocompatibility complex (MHC) proteins. These peptides include tumor neoantigens, which are classically thought of as derived from mutation-containing proteins that generate novel immunogenic epitopes[37]. Despite the ability of neoantigen therapeutic vaccines to promote tumor-specific T-cell responses in a number of pre-clinical models[34–36], clinical efficacy has yet to be demonstrated[7,8]. A significant challenge for translation of TSA therapies is the ability to select the subset of clinically relevant epitopes from all computationally predicted neoantigens. Many neoantigen prediction algorithms rely heavily on peptide/MHC binding affinity predictions to rank epitopes[17–20,22–25]. Unlike murine pre-clinical models, where in vivo/ex vivo methods to further screen for immunogenicity can be applied[35,67], no such benchtop prediction method for immunogenicity is currently available for humans. We have previously demonstrated in multiple murine models that the number of predicted neoantigens is much higher than the number of confirmed immunogenic neoantigens[67]. Studies demonstrate that in some tumors, the number of predicted neoantigens is far greater than the number of immunogenic neoantigens which have been identified in mouse models[65,68]. As such, the development of an algorithm to predict the immunogenicity of neoantigen peptides (i.e.

[1] Chapter 3 was originally published as Smith, C. C. et al. Machine-Learning Prediction of Tumor Antigen Immunogenicity in the Selection of Therapeutic Epitopes. Cancer Immunol. Res. (2019)

variant peptides predicted to bind MHC) would be valuable for screening predicted neoantigens for clinical application.

In addition to conventional single nucleotide variant (SNV) neoantigens, studies have suggested the presence of tumor-specific mRNA splice variants[32,33], expression of non-coding regions[69], and alternative ribosomal products[70–77], allowing an out-of-frame translation to occur outside the setting of an insertion/deletion (INDEL) mutation. An increasing need exists to define frequencies of predicted TSAs existing in an out-of-frame context, their clinical implications, and whether frame-filtering should be applied for computational neoantigen prediction. In the context of SNV tumor antigen prediction, allowing for out-of-frame calls may identify "pseudo-SNV" antigens (i.e. out-of-frame antigens that contain concurrent SNV mutations) with immunogenicity responses similar to what is observed in frameshift neoantigens. As a preliminary approach to identify both in- and out-of-frame neoantigens, we performed SNV tumor-antigen computational screening across all open reading frames, looking for: 1) the correlates of immunogenicity for these predicted neoantigens, and 2) the capacity for out-of-frame epitopes to drive antitumor immunity.

Features associated with neoantigen immunogenicity remain unclear. Here, we have elucidated peptide-intrinsic features significantly associated with vaccine/IFNγ ELISpot–derived immunogenicity scores of MHC class I and class II TSAs. Using gradient boosting with cross-validation, we developed an algorithm to predict MHC I TSA peptide immunogenicity based on peptide-intrinsic biochemical features. We modeled the immunogenicity of predicted neoantigens in the BBN963 basal-like bladder cancer model and demonstrated the capacity of epitopes with high predicted immunogenicity to control tumor growth significantly better than those with low predicted immunogenicity. This algorithm was additionally validated using graft-versus-leukemia (GvL) minor histocompatibility mismatch antigens (mHA) in the P815 mastocytoma allogeneic transplant model. Applying this algorithm to predicted MHC I neoantigens from a TCGA pan-cancer dataset, we observed significant positive association between highly immunogenic neoantigens (HINs; in the top 95th percentile of predicted immunogenicity score) and microsatellite instability (MSI) high–driven immune features in colon adenocarcinoma (COAD) and significant negative association between signatures of anti–PD-1 therapy responsiveness and HIN numbers in lung adenocarcinoma (LUAD) cancer types. Lastly, we provide evidence in favor of antitumor cytotoxic T-cell responses generated against a predicted out-of-frame neoantigen, suggesting a proportion of predicted out-of-frame SNV tumor antigens may be presented by the tumor to generate an immune response. Prediction of peptide immunogenicity on a framework of peptide/MHC binding should improve understanding of antitumor T-cell responses and neoantigen selection for therapeutic vaccine applications.

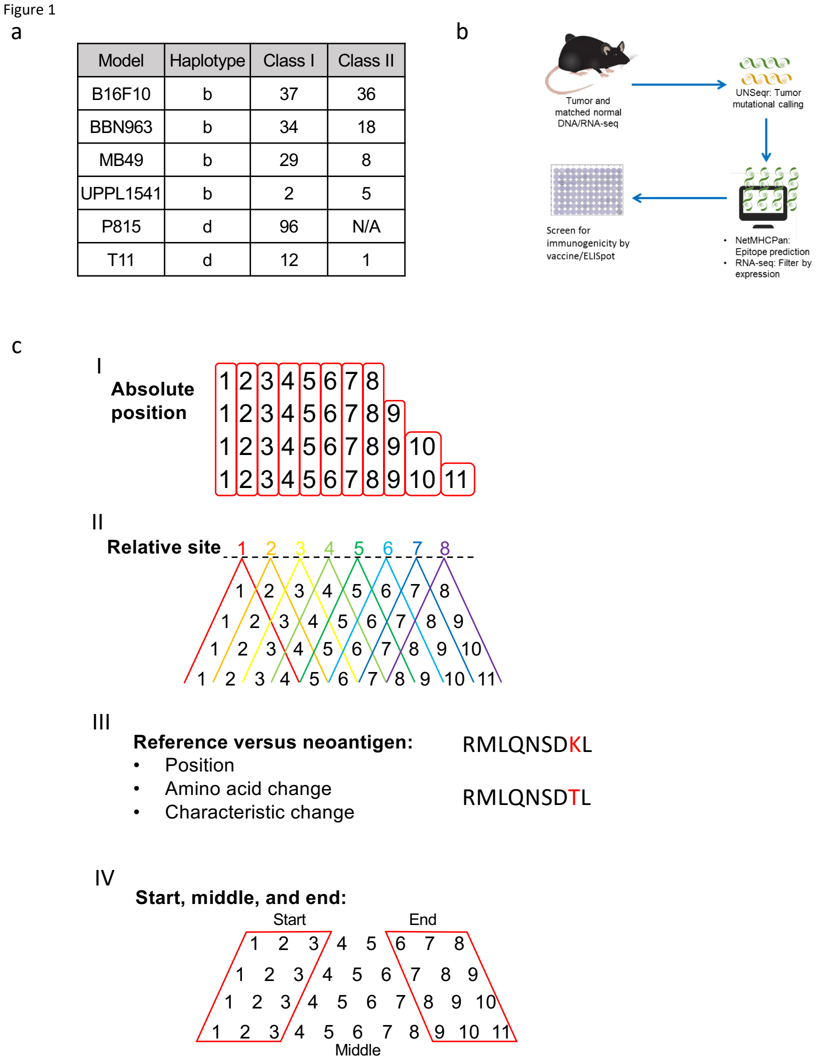
3.2 Results

Correlates of immunogenicity in class I MHC epitopes.

Neoantigens and mHA were predicted in six tumor models (B16F10, BBN963, MB49, UPPL1541, P815, and T11), allowing us to characterize neoantigens in the H2b and H2d haplotypes (Figure 3.1B). We predicted a total of 210 MHC I epitopes and 68 MHC II epitopes and determined their immunogenicity using a vaccine/ELISpot screening approach. Distribution of epitope ELISpot scores (SFC) varied by model, with MB49, B16F10, and P815 tumors including nine of the 10 most immunogenic epitopes, and BBN963 including seven of the 10 least immunogenic epitopes (Supplementary Fig. S1). MHC II epitopes were not predicted for

P815 GvL mHA. With the goal of identifying peptide-intrinsic features that associated with and predicted for immunogenicity, we derived a set of features for each peptide, including the amino acid sequence and characteristic at each I) absolute position, II) relative site, III) site of mutation and changes in amino acid sequence and characteristic at mutational site, and IV) presence of amino acid or characteristic at the beginning, middle, or end of each peptide (Figure 3.1C).

Figure 1



Model	Haplotype	Class I	Class II
B16F10	b	37	36
BBN963	b	34	18
MB49	b	29	8
UPPL1541	b	2	5
P815	d	96	N/A
T11	d	12	1

a
b
Tumor and matched normal DNA/RNA-seq
UNSeqr: Tumor mutational calling
• NetMHCPan: Epitope prediction
• RNA-seq: Filter by expression
Screen for immunogenicity by vaccine/ELISpot
c
I
Absolute position
II
Relative site
III
Reference versus neoantigen:
• Position
• Amino acid change
• Characteristic change
RMLQNSDKL
RMLQNSDTL
IV
Start, middle, and end:
Start
End
Middle

Figure 3.1 Summary of tumor antigen prediction and identification of peptide-intrinsic features. (A) Number of MHC class I and II neoantigens/mHA per tumor model contained within the

study. (B) Schematic of neoantigen/mHA prediction and ELISpot validation workflow. (C) Summary of the major classes of peptide-intrinsic features identified for each antigen, including amino acid sequence and characteristics at I) each absolute position, II) each relative site, III) the mutation position, and IV) the start, middle, and end of each peptide. Red boxes around columns in I) demonstrate each absolute position and in IV) demonstrate the start, middle, and end distinctions. Red lettering in III) provides an example for SNV mutation site between reference and antigenic sequences.

Univariable regression considering intrinsic peptide features as the predictor variable and immunogenicity (measured as IFNγ ELISpot values of T cells derived from vaccinated mice) of class I antigens demonstrated 38 significant features (q-value <0.05; Figure 3.2A) associated with ELISpot immunogenicity. Among these features, the most significant positive associations were changes in the mutation position to a small amino acid (Mutated_position_change_of_Small_feature), valine at relative site 2 ("Relative_site_2_V"), and basic amino acids of the reference sequence at the mutated position (Reference_AA_at_mutated_position_Basic). In contrast, the most signficant negative associations were small amino acids of the reference sequence at the mutated position (Reference_AA_at_mutated_position_Small), changes in the mutation position to a basic amino acid (Mutated_position_change_of_Basic_feature), and polar amino acids at position 6 (Absolute_position_6_Polar). We additionally sought to determine the correlation among the 38 significant features, observing relatively low correlation (Figure 3.2B). Features that demonstrated significant correlation were related, such as 1) the amino acid at the mutated position of the reference sequence with charged or basic features, 2) valine or small amino acids at absolute position 11, and 3) the presence of a valine or small amino acid at the last position and valine at relative site 8.

Next, we evaluated the independence of the variables identified using univariable analysis using multivariable regression. To increase confidence of our multivariable regression, we performed backward stepwise regression, optimizing on Akaike Information Criterion (AIC), as described in the Methods. Variables whose loss resulted in an insignificant change to the model performance (as measured by the AIC) were removed from the set of variables until no further variables could be removed without a significant decrease in model fit. Sixteen significant features from this step were inputted into multivariable regression. The resulting model of 8 features indicated 33.4% variation (p<0.0001) in immunogenicity was explained by the prediction (Supplementary Fig. S2A). Significant features of the multivariable model included valine at the last position (Last_position_V (p=0.0001)), tyrosine at position 3 (Absolute_position_3_Y (p=0.003)), changes in the mutated position to a small amino acid (Mutated_position_change_of_Small_feature (p=0.007)), cysteine at relative site 4 (Relative_site_4_C (p=0.012)), lysine at relative site 5 (Relative_site_5_K (p=0.015)), tiny amino acids at relative site 6 (Relative_site_6_Tiny (p=0.016)), basic amino acids of the reference sequence at the mutated position (Reference_AA_at_mutated_position_Basic (p=0.027)), and valine at relative site 2 (Relative_site_2_V (p=0.041) ; Supplementary Fig. S2B; Supplementary Table S1). To ensure this model was accurately representing both Hb and Hd haplotypes, we tested the immunogenicity for each of these five significant features, split categorically, which demonstrated similar trends between both haplotypes (Supplementary Fig. S3). We did not observe significant differences in ELISpot immunogenicity among predicted in-frame (n=131) and out-of-frame (n=79) epitopes, emphasizing that peptide-intrinsic features were the primary driver for immunogenicity (p>0.05; Supplementary Fig. S4).

Correlates of immunogenicity in class II MHC epitopes.

Among class II epitopes, 15 peptide-intrinsic features were significantly associated with ELISpot immunogenicity (GLM q-value <0.05; Figure 3.2C). Among the most significant positively associated included changes in the mutation position to a non-polar amino acid (Mutated_position_change_of_NonPolar_feature), valine at position 1, tyrosine at position 6, and basic amino acid at position 2. the more significant negatively correlated feature was a change in the mutation position into a small amino acid (Mutated_position_change_of_Small_feature), which was positively correlated in class I epitopes. Negatively correlated features also included changes in the mutation position to a polar amino acid (Mutated_position_change_of_Polar_feature), and small/tiny amino acids at the mutated site. Among the significant features (Figure 3.2D), we observed one cluster of closely inter-correlated features (Sig 1; n=7, right-hand side of dendrogram), as well as a second cluster of loosely inter-correlated features (Sig 2; n=8, left-hand side of dendrogram). With each respective tumor model defined as a binary variable (1 = true, 0 = false), the mean expression of cluster 1 features was significantly correlated with the B16F10 model and inversely correlated with MB49, whereas the mean expression of cluster 2 was significantly correlated with MB49 tumors (Supplementary Fig. S5; Spearman q-value <0.05). This corroborated with the greater burden of immunogenic class II neoantigens identified in these two models, suggesting relatively greater contribution of these two models (particularly MB49) in the regression outcomes. Using the same backwards AIC stepwise regression approach described above, multivariable GLM regression was performed on six features. The resulting model indicated that 50.7% variation (p<0.0001) in immunogenicity was explained by the prediction (Supplementary Fig. S6; Supplementary Table S2), with three significant features (tyrosine at positive 6 (Absolute_position_6_Y), valine at positive 1 (Absolute_position_1_V) and changes in the mutated position to a small amino acid

(Mutated_position_change_of_Small_feature)) primarily driving the fit (p=0.024, 0.002, and 1.6x10-5, respectively). As with class I epitopes, we did not observe significant differences in immunogenicity among predicted in-frame (n=59) and out-of-frame (n=9) epitopes (p>0.05; Supplementary Fig. S3).

Figure 2

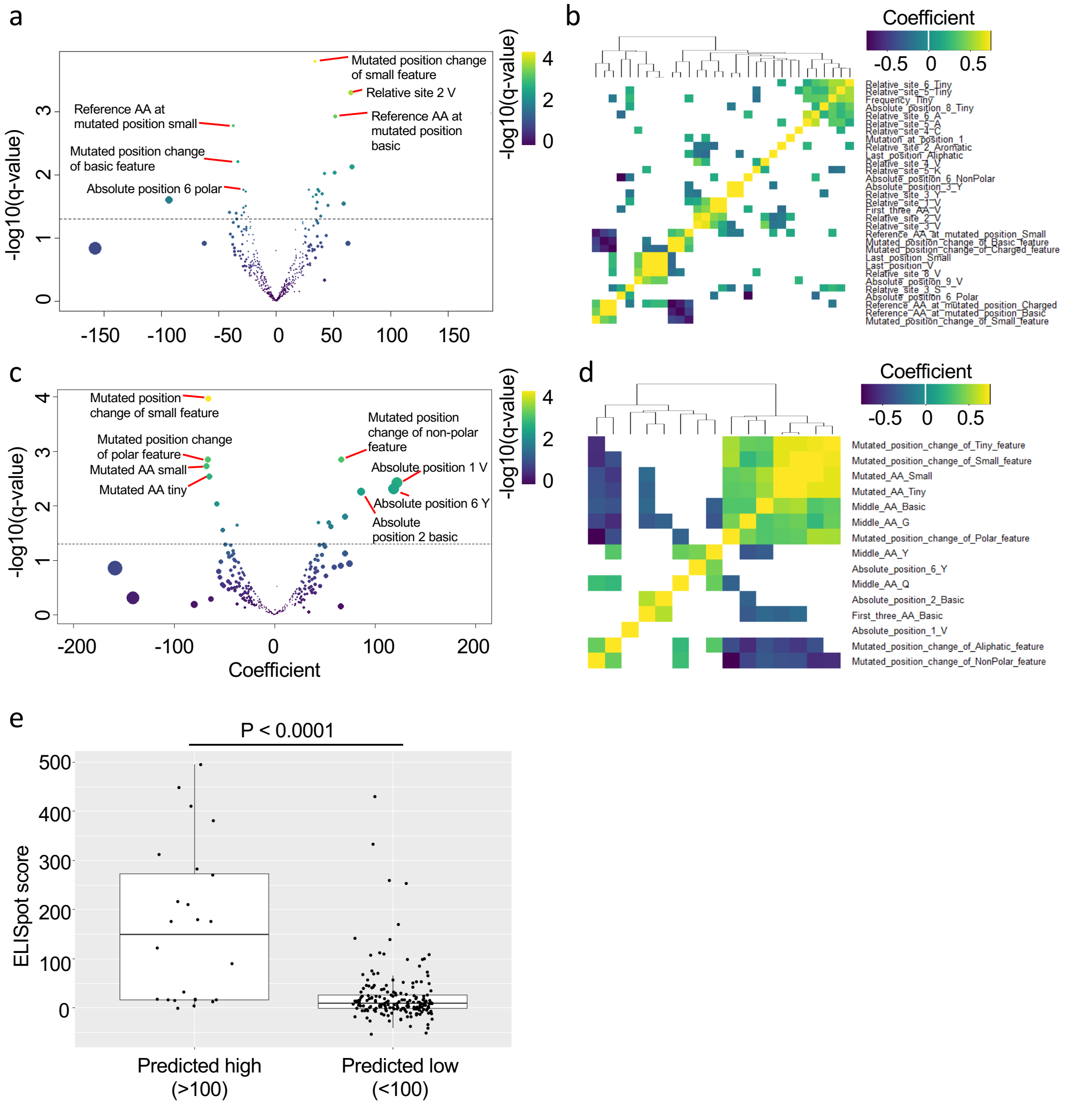


Figure 3.2 Linear regression analysis between peptide-intrinsic features and tumor antigen immunogenicity. (A and C) Volcano plots representing the generalized linear method (GLM) coefficient (x-axis) and –log10(q-value)(y-axis) for each peptide-intrinsic feature as a predictor for immunogenicity in (A) class I and (C) class II neoantigens/mHA. Dashed line represents q-value=0.05. Spot color represents –log10(q-value) magnitude and size represents magnitude of the coefficient. (B,D) Heatmap representing Spearman correlations between each significantly correlated feature from (A,C) for (B) class I and (D) class II neoantigens/mHA, respectively. Colored cells represent significantly correlated features (q<0.05), with magnitude of the correlation coefficient represented by color. (E) ELISpot-derived immunogenicity scores for class I neoantigens/mHA classified as predicted high (>100) or low (<100) immunogenic by multivariable GLM regression. Data represent median (middle line), with boxes encompassing the 25th to 75th percentile, whiskers encompassing 1.5× the interquartile range from the box, and

independent values shown by dots. Statistics performed with Mann-Whitney U-test, with significance defined as $p<0.05$.

Machine-learning algorithm for immunogenicity prediction in class I MHC epitopes.

Our analysis of class I epitopes using multivariable GLM suggested an optimized multivariable model may be able to discriminate between high- and low-immunogenicity peptides (Figure 3.2E). With the goal of designing a predictive model for neoantigen and mHA immunogenicity, we split our class I epitope database into an exploration (2/3 of epitopes, n=141) and validation (1/3 of epitopes, n=69) set (Figure 3.3A). Class II modeling was not attempted due to the low number of epitopes available within our database (n=68). In order to reduce noise within our model, we collapsed ELISpot scores with absolute values less than or equal to the absolute value of the most negative count (¬–53 spots) to zero. This was performed because we were not focused on the ability of the model to characterize exact immunogenicity values within the low immunogenicity range. Within the exploration set, we used a 10,000-fold cross-validation (2/3rd random resampling) approach, which demonstrated that only gradient boosting consistently performed better than chance. Our final gradient boosting algorithm contained seven predictive features: valine at position 1 (Absolute_position_1_V), valine at the last position (Last_position_V), small amino acids at the last position (Last_position_Small), basic amino acids of the reference sequence at the mutated position (Reference_AA_at_mutated_position_Basic), changes in the mutated position to a small amino acid (Mutated_position_change_of_Small_feature), lysine at relative site 1 ("Relative_site_1_K"), and presence of valine within the first 3 positions (First_three_AA_V). The class I validation set was run through this final gradient boosting algorithm, demonstrating significantly accurate performance when comparing the linear fit between the actual immunogenicity by ELISpot and the predicted immunogenicity by modeling (p=0.01893, coefficient=0.30, Figure 3.3B). This model provided a high negative predictive value (83.6% predicted values <53), ideal in the setting for filtering out a large pool of predicted tumor antigens in order to select epitopes for therapeutic targeting.

In vivo validation of the class I immunogenicity prediction model

To test whether our final algorithm increased the likelihood of identifying clinically relevant, immunogenic epitopes for antitumor vaccine responses, we used two tumor models within our validation set: BBN963 basal-like bladder cancer neoantigens (epithelial tumor) and P815 mastocytoma GvL mHA (hematopietic tumor). Epitopes were binned into predicted high immunogenicity (top quartile) and predicted low immunogenicity (bottom quartile) groups for comparison of relative efficacy (Supplementary Table S3). In BBN963 tumors, three predicted high (B2: VALLPSVML; C2: VSLTLFSSWL; A5: SNVMQLLL) and two predicted low (B5: ETLLNSATI; B12: MISRNRHTL) immunogenicity neoantigens were identified. Animals were vaccinated with 30 µg of one peptide (or no-peptide control) alongside 50 µg poly(I:C) as adjuvant, challenged with tumor 12 days after vaccination, then given a 30 µg peptide boost on day 21 after the initial vaccination (Figure 3.3C). Animals vaccinated with predicted a high immunogenicity peptide survived longer than those vaccinated with either predicted low immunogenicity peptide (p=0.0006) or no-peptide control (p=0.0031; Figure 3.3D ; Supplementary Fig. S7A). In contrast, no significant difference in survival was observed between predicted low immunogenicity peptide and no-peptide control groups (p=0.9674). We additionally observed better control of tumor size in animals vaccinated with predicted high immunogenicity peptide (Supplementary Fig. S7B–S7D).

In P815 tumors, two predicted high (AFQRVTCTTL and QYSSANDWTV) and three predicted low (HYAANEWI, KFFPNCIFL, and LYISPNPEVL) immunogenicity GvL mHAs were identified. BALB/c donor animals were vaccinated with a pool of predicted high or low immunogenicity peptides (100 µg each peptide) or no-peptide control, with 50 µg poly(I:C) as adjuvant on days 0 and 7. DBA/2 recipient animals were lethally irradiated on day 13; transplanted with 3x106 BALB/c T cells, 3x106 BALC/c bone marrow cells, and 3x105 P815 tumor cells on day 14; and finally given a 3rd booster vaccine on day 21 (Figure 3.3E). Animals given predicted high immunogenicity T cells survived longer than those given predicted low immunogenicity T cells (median survival 44.5 and 28 days, respectively), both of which survived longer than no-peptide control T cells (median survival 19 days, Figure 3.3F). Additionally, we observed significantly lower tumor burden in high immunogenicity versus low immunogenicity peptide vaccinated animals by luciferase imaging on day 26 ($p < 0.05$, Mann-Whitney u-test; Supplementary Figs. S8 and S9). All groups receiving donor T cells demonstrated measurable graft-versus-host disease (GvHD) clinically after transplant, without significant differences in weight loss or GvHD clinical scores between groups up to day 30 (Supplementary Fig. S10). In summary, these experiments demonstrated the in vivo biological relevance of our immunogenicity prediction model, with significant differences observed between predicted high and low immunogenicity epitopes.

Figure 3

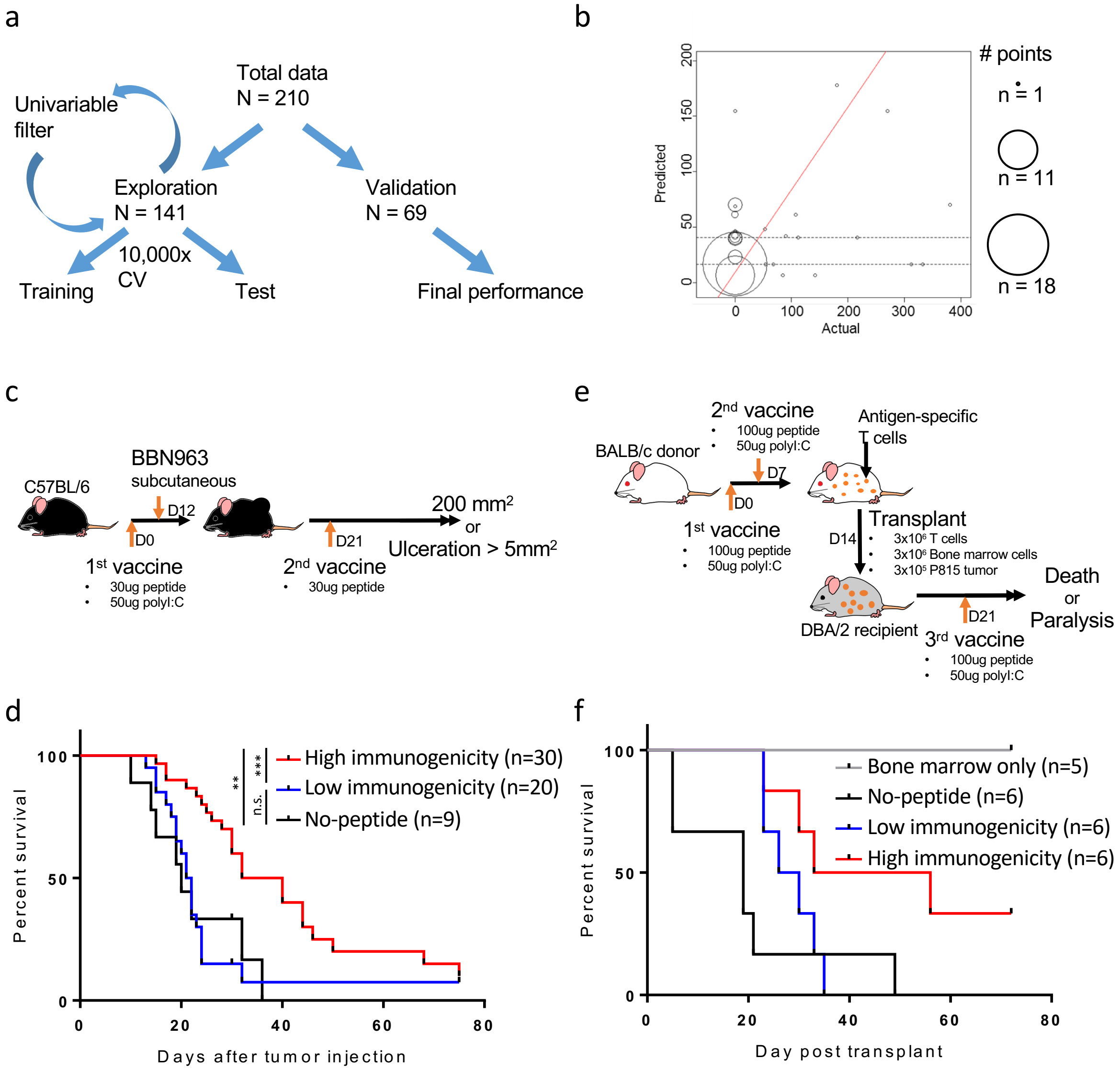


Figure 3.3 Performance and validation of the gradient boosting model (GBM) approach for predicting neoantigen/mHA immunogenicity. (A) Schema of the cross-validation approach used for GBM model building. (B) Performance of the final GBM model in validation set, showing actual (x-axis) versus predicted (y-axis) immunogenicity scores. Size of each point represents number of antigens at each coordinate. Red line represents the line of best fit, with p-value of fit shown above the graph. (C, E) Schema for in vivo validation experiments, with tumor vaccine studies performed in (C) BBN963 basal-like bladder cancer and (E) the P815 mastocytoma syngeneic transplant model. (D, F) Kaplan-Meier survival curves for animals bearing (D) BBN963 basal-like bladder cancer and (F) P815 mastocytoma syngeneic transplants. Animals treated with predicted high (red) or low (blue) immunogenicity antigens, no-peptide control (black), or bone marrow only control (grey). Data in (D) represents two independent experiments.

Data in (F) represents one independent experiment. Statistics performed with log-ranked testing (**p<0.01; ***p<0.001).

Correlates of predicted immunogenicity in human class I epitopes.

Although the immunogenicity prediction algorithm was designed and validated in mice, we hypothesized that similar rules of immunogenicity may exist among human neoantigens. To study this, we ran previously predicted MHC I neoantigens from TCGA through our machine-learning algorithm, generating immunogenicity scores for each epitope[65]. As expected, we observed a correlation between the number of HINs identified by our model (>95th percentile) with number of total neoantigens (Pearson correlation $p<0.0001$; Supplementary Fig. S11). Therefore, we performed regression studies between HIN count and immune features without controlling for total neoantigen burden. We observed significant association between HIN count and IFNγ, cytotoxicity, CD8+ T-cell and total T cell, and B-cell immune gene signatures (IGS) among the dataset (not controlling for cancer type; Figure 3.4A). Assessing these associations individually by tumor type, we observed that the most significant associations were encompassed by the colon (COAD) and lung (LUAD) adenocarcinoma cancer types (Figure 3.4B). Within COAD, ta positive association between HIN count and many T-cell and cytotoxicity signatures. To identify potential drivers of this pattern, we looked for the association between HIN count and MSI status, observing that MSI-high COAD tumors had significantly higher HIN counts (Figure 3.4C; Supplementary Fig. S12).

In contrast, LUAD demonstrated a negative association with signatures of anti–PD-1 responsiveness and several immune cell signatures. Regression analysis between these negative IGS features and LUAD oncogene/tumor suppressor copy numbers demonstrated preferential association with MYC copy number (q-value <0.05; Figure 3.4D; Supplementary Fig. S13). To demonstrate that MYC amplification provided a pro-tumorigenic signal in LUAD, we observed significantly greater expression of genes corresponding with cell cycle gene patterns (Supplementary Fig. S14A, S14C, and S14D), as well as enrichment of downstream genes involved in the MYC pathway (Supplementary Fig. S14B) among MYC-amplified tumors. This increased proliferation pattern was additionally associated with decreased sharing of T-cell receptor sequences from tumor-infiltrating T cells in MYC-amplified tumors, suggesting a potentially altered antitumor immune response (Supplementary Fig. S14E). We did not find HIN count to correlate with MYC copy number (Pearson p>0.5), suggesting tumor immunogenicity burden and MYC target expression may be independent predictors for immune exclusion and checkpoint inhibition resistance.

Figure 4

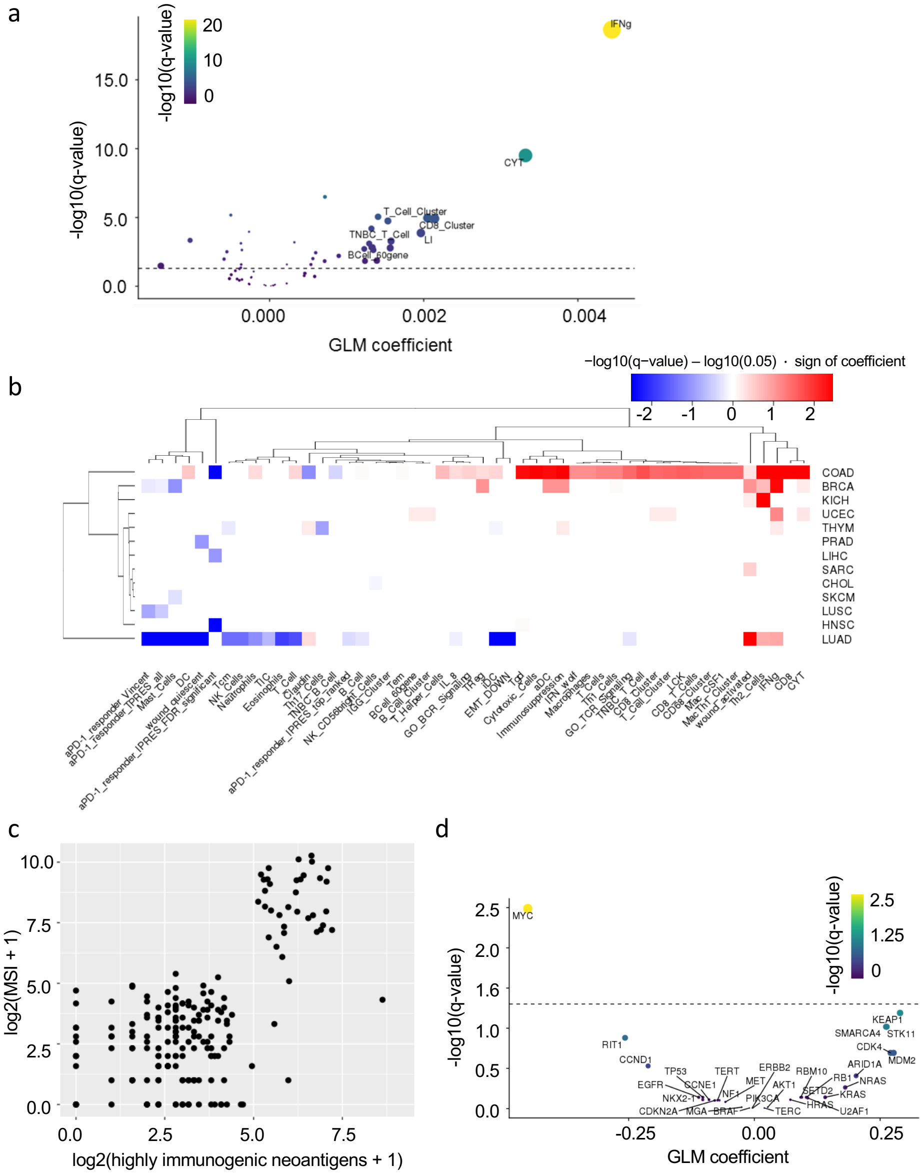


Figure 3.4 Correlative analysis of predicted neoantigen immunogenicity in TCGA human datasets. (A) Volcano plot representing generalized linear method generalized linear method (GLM) coefficient (x-axis) and –log10(q-value)(y-axis) between numbers of highly immunogenic neoantigens (HINs) and immune gene signatures (IGSs) in TCGA pan-cancer datasets. (B) Heatmap representing GLM regression between numbers of HINs and IGS for each TCGA cancer subset. Color represents direction of coefficient (red: positive; blue: negative), and shade represents –log10(q-value) magnitude. (C) Number of HINs (x-axis) versus microsatellite instability (MSI) score (y-axis) for a TCGA colorectal carcinoma (COAD) dataset. (D) Volcano plot representing GLM coefficient (x-axis) and –log10(q-value)(y-axis) between average

expression of IGS in (B) with significantly negative association with HIN burden and oncogene/tumor suppressor copy numbers in a TCGA lung adenocarcinoma (LUAD) dataset. (A, D) Dashed line represents q-value=0.05.

Out-of-frame neoantigen epitopes promote anti-tumor immunity.

Through the design of our neoantigen prediction algorithm, we considered predicted tumor epitopes across all open reading frames. As such, subsets of our predicted neoantigens were frameshifted epitopes that contained a mutation. We hypothesized that through mechanisms, such as novel splice variants and ribosomal dysfunction, a subset of these out-of-frame predicted antigens could arise in the tumor, allowing for a viable target with greater heterogeneity from self-antigen. Indeed, two of the predicted high immunogenicity neoantigens used in our BBN963 vaccine studies were predicted to be out-of-frame (B2, C2), although still providing therapeutic efficacy over predicted low immunogenicity and no-peptide controls. One of these antigens (B2) demonstrated computational evidence of translation in the out-of-frame context using two de novo transcriptome assemblers Trinity[78] and StringTie[61], whereby the presence of a 5’ start codon was identified with no intervening stop codon up to the B2 antigen site.

With therapeutic and computational evidence in favor of B2 antigen-mediated antitumor immune responses against BBN963 tumors, we next confirmed the presence of B2/MHC tetramer–specific CD8+ T cells infiltrating within the tumor of BBN963-bearing animals, suggesting an antigen-driven T-cell response (Figure 3.5A; Supplementary Fig. S15). Tetramer sorting and peptide-pulsed dendritic cell cocultures of B2-specific T cells demonstrated approximately 40-fold expansion of T cells within 10 days (<5x105 to >2x106), with maintenance of a B2-enriched population Supplementary Fig. S16). Using a flow cytometric cytotoxicity assay, coculture of B2-specific T cells with the BBN963 cell line demonstrated >1.7x increase in killing of target cells (15.25%) compared to the OTI T-cell irrelevant control (8.85%). Neither B2-specific nor OTI T cells demonstrated killing of irrelevant splenocytes control cells (1.1% and 0.85%, respectively; Figure 3.5B; Supplementary Fig. S17). B2-specific T-cell killing of BBN963 over that of OTI T-cell controls was additionally confirmed using a 51Cr-release cytotoxicity assay (Figure 3.5C). Altogether, these results suggested the presence of a cytotoxic CD8+ T-cell response against the out-of-frame B2 neoantigen in BBN963.

Figure 5

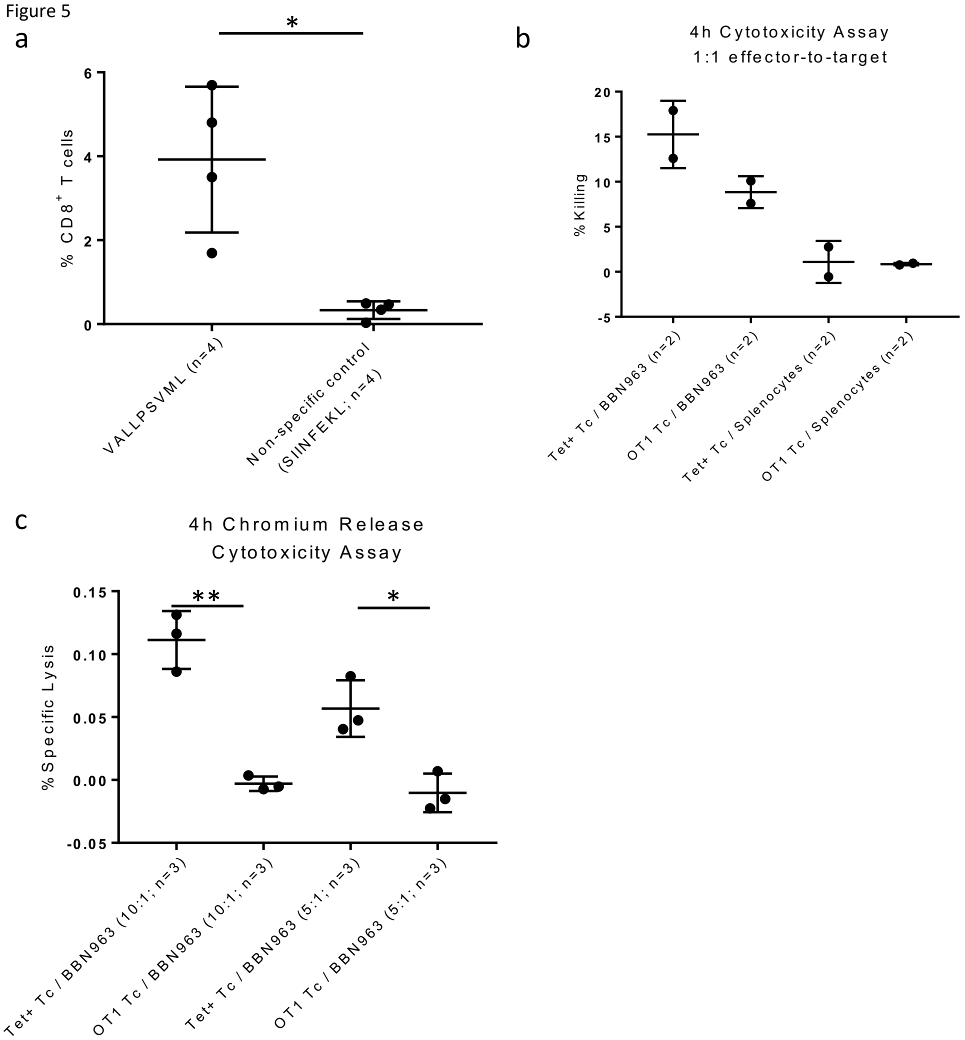


Figure 3.5 Analysis of out-of-frame epitope B2-specific T cells. (A) Percent B2 tetramer–positive (left) versus irrelevant SIINFEKL-tetramer control (right) BBN963 tumor-infiltrating CD8$^+$ T cells. Statistics performed with Mann-Whitney u-test (*p<0.05). (B) 4-hour flow cytometric cytotoxicity assay, comparing percent specific killing in 1:1 cocultures of B2 tetramer–specific T cells or OTI irrelevant T-cell controls versus BBN963 target or irrelevant splenocyte target control. Percent killing represents spontaneous target death background subtracted values. Statistics not performed for (B) due to n=2 sample size across all groups. (C) 4-hour chromium-51 release assay, comparing percent specific killing in cocultures of B2 tetramer–specific T cells or OTI irrelevant T-cell controls versus BBN963 targets at 10:1 and 5:1 effector-to-target ratios. (A-C) Error bars represent mean±standard deviation. Data from each graph represents one independent experiment, respectively. Statistics performed with Welch's t-test (**p<0.01; *p<0.05).

3.3 Discussion

The study presented here addressed two unanswered questions regarding tumor antigens: 1) what features of a tumor antigen sequence are associated with immunogenicity, and 2) can inclusion of non-canonical, out-of-frame epitopes provide viable targets for anti-tumor therapeutic vaccination? We demonstrated that peptide-intrinsic features of predicted tumor antigens could discriminate epitopes with therapeutic efficacy, and that inclusion of out-of-frame epitopes among this pool could provide antitumor immunity against these alternative antigens. We showed that reading frame was not a significant determinant for immunogenicity (at least among peptides with predicted binding affinity <500 nM), and that exclusion of frame-filtering could identify out-of-frame epitopes with therapeutic antitumor, cytotoxic activity. Although the optimal rules for immunogenicity may differ between in-frame and out-of-frame tumor antigens, our relatively limited training set was underpowered to discriminate between these two classes. As such, future studies should be performed to address the biological differences between in- and out-of-frame tumor antigens, and what methods can most optimally identify clinically relevant epitopes in each class.

Our analysis of class I epitopes demonstrated similar trends in expression of features significantly associated with immunogenicity, as well as potential generalizability of our final gradient boosting algorithm for human MHC. We were unable to demonstrate here whether MHC haplotype may influence immunogenicity prediction, given our murine models were limited to two haplotypes. That said, it may be the case that certain features may significantly impact immunogenicity in a way that is conserved across haplotypes. A potential bias in our analysis is the variation in distribution of ELISpot scores among various models. This variation is likely a product of both our selection process (i.e. peptide selection based on a threshold predicted MHC binding affinity) as well as the number of predicted epitopes available for screening in each model. As methodology for antigen prediction and validation was conserved across all models, as well as biological validation performed across two independent tumor models, we do not believe there to be significant underlying biological differences among epitopes identified between different tumor models.

Despite the increased interest in neoantigen-based therapeutic tumor vaccine therapy, prediction algorithms capable of directly predicting for neoantigen immunogenicity are lacking[6], and no neoantigen immunogenicity predictor trained specifically on tumor antigen data exists. Current neoantigen immunogenicity predictors are instead trained on databases containing immunogenicity scores from all potential MHC-binding epitopes, of which the biology may not closely match that of mutation-derived tumor antigens. An example of this biological disparity is observed in the vastly different immune response rates between neoantigens and tumor-specific endogenous retroviral epitopes[79], whereby the concept of a "self" and "non-self" antigen is not considered as a feature for immunogenicity prediction among current algorithms. Training our model specifically on "self" tumor antigenic sources instead allows for greater specificity of selection for peptide-intrinsic features which correlated with ex vivo validated IFNγ release scores. Our final predictive algorithm demonstrated capacity to select for therapeutically relevant epitopes, as observed in our treatment studies where predicted high immunogenicity peptides controlled tumor burden significantly better than predicted low immunogenicity peptides and no-peptide control groups. This model demonstrated a high true-negative rate, which is ideal in the setting of filtering out many weakly immunogenic epitopes to select for a small pool of targets for therapy.

Validation experiments in BBN963 and P815 models were performed as a combination of prophylactic and therapeutic vaccines, rather than strictly treating animals after tumor injection. This method was selected due to the intrinsically low efficacy of free-peptide vaccines, whereby differences in therapeutic efficacy may not be observed between predicted high and low immunogenicity antigens[80]. As such, although these experiments provided evidence for the biological relevance of our computational model, development of more robust therapeutic vaccine platforms are still necessary for improving response rates to peptide-based, tumor-specific antigen vaccines. From these studies, we observed that predicted high immunogenicity peptides had greater benefit in the therapeutic vaccine setting than predicted low immunogenicity peptides. As such, we reasoned that although HIN count and total neoantigen burden were correlated, the most HINs were the key drivers of immunity. Thus, we performed regression studies between HIN count and immune features without controlling for total neoantigen burden. Analysis of human neoantigen data from TCGA demonstrated association between presence of HINs with features of immune response in colon and lung adenocarcinoma. Although the association between HIN count and immune gene signature expression, as well as MSI-status, in COAD agreed with the classical view of a tumor-antigen driven immune response, the negative association with immune features (including signatures of anti–PD-1 responsiveness) in LUAD is less clear. A report from Jerby-Arnon and colleagues demonstrates an association between resistance to immune checkpoint inhibition and MYC target expression[81]. As such, we initially hypothesized that MYC target expression may be the common driver for immune exclusion, anti–PD-1 non-responsiveness, as well as high HIN burden. However, MYC copy number did not correlate with HIN count, suggesting MYC expression and HIN burden are independent pathways of immune exclusion and checkpoint inhibitor resistance in LUAD. Further studies are necessary to more closely examine the relationship between tumor immunogenic neoantigen burden and immune features, elucidating why higher immunogenicity burden is unexpectedly negatively associated with IGS patterns.

Currently, it is not well understood what frequency of tumor antigens arise from conventional in-frame antigens versus non-conventional antigenic sources, such as retroviral/retrotransponson expression, intron expression, and out-of-frame translation. A study from Laumont et al. suggests that non-coding regions are the main source of tumor-specific antigens in acute lymphoblastic leukemia patient samples[69], providing evidence that current methods for neoantigen prediction may be limited by filtering for in-frame exon regions. Laumont and colleagues used an RNA-based screening approach whereby k-mers derived from tumor RNA-seq reads were directly screened against matched-normal RNA k-mers, keeping only tumor specific regions. Compared to conventional exome-based TSA calling, this RNA-based screening approach allowed for identification of a broader repertoire of epitopes, consistent with the increased frequency of non-canonical TSAs identified by Laumont et al. compared to this current study[69]. Although Laumont and colleagues used a mass spectrometry approach to confirm expression of out-of-frame epitopes, no computational methods have been used to identify such non-canonical epitopes. As such, our study relies upon a naïve, non-biased approach for screening out-of-frame antigens, whereby we combined conventional exome-based SNP antigen calling with identification of potential epitopes across all open reading frames. We demonstrated that the frame of an epitope did not associate with immunogenicity, but inclusion of out-of-frame epitopes could provide therapeutic benefit. This analysis highlighted how some proportion of SNV neoantigens predicted to be out-of-frame may still maintain expression and capacity to trigger a cytotoxic T-cell response against the tumor. If such antigens are further confirmed in human cancers, there are important implications that will need to be addressed: 1) whether the biology and immunogenicity of these out-of-frame "SNV" antigens more closely mirrors that of classical SNV-neoantigens or whether they are instead more similar to INDEL-derived neoantigens or alternative neoantigens, such as tumor-specific endogenous retroviral antigens; and 2) if reading frame filters should be applied to current neoantigen calling algorithms in order to most effectively capture the targetable antigen landscape of a tumor.

3.4 Materials and Methods

Cell lines.

The B16F10 cell line was purchased from ATCC (CLR-6475) and cultured according to the ATCC protocol. The P815 cell line was purchased from ATCC (TIB-64), transduced with luciferase as previously described[82], and cultured according to the ATCC protocol. The BBN963, UPPL1541, and MB49 cell lines were obtained and passaged as previously described [67]. The T11 model was obtained and passaged as previously described[83]. All cells used in this study were derived from viably frozen stocks of the above cell lines, with aliquots derived within ≤5 passages of the original stock. No mycoplasma testing was performed. No further authentication was performed on cell lines directly purchased from ATCC (B16F10, P815) or those received directly from the deriving lab (T11: Charles M. Perou, UNC Lineberger; BBN963, UPPL1541: William Y. Kim, UNC Lineberger). MB49 cell line was authenticated through transcriptomic analysis, as previously described[67].

Animal studies.

All experiments described in this study were approved by the UNC Institutional Animal Care and Use Committee (IACUC). Animals used in this study, their vendor source, and respective tumor cell lines included: C57BL/6J (Jackson Laboratories; B16F10), C57BL/6 (Charles River Laboratories; BBN963, MB49, UPPL1541), DBA/2J (Jackson Laboratories; P815), and BALC/c (Jackson Laboratories; T11). Tumor injection routes and cell numbers for all models and experiments included: B16F10: Flank subcutaneous (s.c.),105 cells; BBN963: Flank s.c., 107 cells; MB49: Flank s.c., 105, UPPL1541: Flank s.c., 106, T11: Mammary fat pad intradermal, 104, P815: Tail vein intravenous, 3x105. Tissue collection and DNA/RNA isolation is described in the "Neoantigen and mHA prediction" section below. Graft-versus-Host disease (GvHD) scoring was performed as previously described[84], with score defined as the sum of five components of posture, fur, activity, skin, and weight loss on a 0-2 scale.

Tissue Dissociation.

All single-cell suspensions mentioned in the below methods sections were derived using the below listed protocol. Tissues were homogenized in cold PBS using the GentleMACs Dissociator and the samples were passed through a 70 µM cell strainer using a 5 mL syringe plunger. The samples were centrifuged for seven minutes at 290 RCF, 4°C, decanting the supernatant. The remaining pellet was resuspended into 1 mL of ACK lysis buffer (150 mM NH4Cl, 10 mM, KHCO3, 0.1 nM Na2EDTA in DPBS, pH 7.3) for 2 minutes at room temperature before quenching with 10 mL of cold media. The samples were centrifuged for seven minutes at 290 RCF, 4°C, resuspended in 10 mL of cold media, and passed through a 40 µM cell strainer.

Neoantigen and mHA prediction.

Neoantigen prediction was performed as previously described[67]. Briefly, mice were injection with tumors (Figure 3.1A) in the route and counts listed above, and monitored until tumor size reached 100mm3 by caliper measurement ((l x w^2)/2, where w is the smaller of two perpendicular tumor axes), at which point mice were humanely sacrificed with CO2 asphyxiation followed by cervical dislocation. P815 tumor samples were collected directly from cell line culture (105 cells per sample). RNA was extracted from single-cell suspensions of tumors using Qiagen RNeasy Mini kit (cat. # 74104), and DNA was extracted from single-cell suspensions of tumors and matched-normal tail clippings or livers using Qiagen DNeasy kit (cat. # 69504), all according to manufacturer's protocol. Whole exome and transcriptome library preparation was performed using Agilent SureSelect XT All Exon and Illumina TruSeq Stranded mRNA library preparation kits, respectively. Libraries were sequenced via 2x100 runs on an Illumina HiSeq 2500 at the UNC High Throughput Sequencing Facility (HTSF). Tumor mutations were called using UNCeqR (https://lbc.unc.edu/~mwilkers/unceqr_dist/)[85], filtering for SNV mutations with at least 5x coverage by RNA-seq. Translated 8-11mer (class I) or 15mer (class II) peptides were derived across all open reading frames, and then predicted for MHC binding affinity using NetMHCPan3.0 (http://www.cbs.dtu.dk/services/NetMHCpan-3.0/)[17]. Class I minor mismatch antigens were predicted similarly in the P815 model against the BALB/c histocompatible donor. Predicted binders were filtered by binding affinity <500 nM, generally accepted cutoff for immunogenicity as previously noted[8,20,28], with top epitopes screened for immunogenicity using a vaccine/ELISpot approach, as described below.

Vaccine/ELISpot screening.

Predicted neoantigen peptides (MHC I: n = 210; MHC II: n = 68) were synthesized by New England Peptide (Gardner, MA), using custom peptide array technology (Supplemental Data File 1). Non-tumor bearing wildtype animals were vaccinated with predicted neoantigen peptides of their respective predicted haplotype, given as a subcutaneous injection of a pool of 8 equimolar peptides (5 nmol total peptide) and 50 µg poly(I:C) (Sigma, cat. # P1530) in PBS. A second identical injection was repeated 6 to 7 days after primary injection. Mice were humanely sacrificed with CO2 asphyxiation followed by cervical dislocation 5 to 6 days after the second injection. Spleens were harvested and prepared into single cell suspension, as described above. Splenocytes were plated in triplicate at 5 × 105 cells per 100 µL media (RPMI 1640 (Gibco cat. # 11875-093) with 10% FBS (Gemini cat. # 900-108) onto an IFNγ capture antibody-coated ELISPOT plate (BD Biosciences, cat. # 551083) according to manufacturer protocol for 48-72 hours, along with 1 nmol of a single peptide against which the respective mouse was vaccinated. Immunogenicity was defined as the average number of spot-forming cells (SFC) identified using an ELISpot plate reader (AID Classic ERL07), with no-peptide background subtracted from each epitope.

Computational analysis.

Variables used in neoantigen immunogenicity regression and modelling were derived using the “aaComp” command of the R package “Peptides” (v2.4; https://cran.r-project.org/web/packages/Peptides/index.html). Using features derived from this command (Tiny, Small, Aliphatic, Aromatic, Non-polar, Polar, Charged, Basic, Acidic), variables were derived by the presence (1) or absence (0) of each feature at each absolute and relative position along each antigen, at the site of SNV mutation along each antigen, at the first or last 3 amino acid residues (beginning/end) or middle residues (middle) of each antigen, or difference (loss: -1, gain: 1, or no change: 0) of each feature in the mutated versus reference antigen along SNV mutation site.

For all GLM and predictive modeling analyses, low variance variables (defined by the "nearZeroVar" function of the "caret" package) were removed prior to further analysis. Generalized linear models (GLM) using the R "glm" function were used for all non-modeling univariable and multivariable linear regression analyses, with significance reported as false discovery rate (FDR)-adjusted p-values (q-value) using the R "p.adjust" command. Backward stepwise regression for multivariable modelling was performed using the R "stepAIC" command of the "MASS" package (https://cran.r-project.org/web/packages/MASS/index.html), optimized on Akaike Information Criterion (AIC). Backward stepwise regression was performed by starting with all variable candidates and testing the deterioration of the model with removal of each variable.

For immunogenicity prediction modeling, analyses were performed using the R package "caret" as a wrapper for running each multivariable approach: GLM, elastic net, random forest, gradient boosting, and linear and radial support vector machine methods. For cross validation, data were split into exploration (n = 141) and validation (n = 69) sets using the caret "createDataPartition" function, confirming statistically non-significant differences in measured immunogenicity between exploration and validation sets (Mann-Whitney p > 0.2). Model performance was derived from Pearson correlation coefficients between ELISpot immunogenicity and predicted immunogenicity scores, using a 10,000-fold cross-validation (2/3rd random resampling) approach within the exploration set, with the input predictor variables limited to those that demonstrated significant univariable correlation in >50% of 1000-fold bootstrapping iterations (2/3rd resampling) within the exploration set. The final gradient boosting machine-learning algorithm immunogenicity prediction of MHC I epitopes can be accessed at https://github.com/vincentlaboratories/neoag.

To explore for computation evidence of out-of-frame transcripts, StringTie (https://ccb.jhu.edu/software/stringtie/)[61] and Trinity (https://github.com/trinityrnaseq/trinityrnaseq/wiki)[78] were used for de novo assembly of transcripts from BBN963 RNA-seq data, according to standard workflow provided in the above links.

Peptide treatment studies.

BBN963 basal-like bladder cancer model treatment began with pre-tumor vaccination with 30 µg of a single peptide (or no-peptide control) and 50 µg poly(I:C) adjuvant injected in 100 µL PBS intradermally in the flank of 8-10 week old female C57BL/6 mice (Charles River). Twelve days after vaccination, 1x107 BBN963 cells were injected in 100 µL PBS subcutaneously in the flank, ipsilateral to the vaccine site. On day 21 post primary vaccination, animals were given a vaccine booster with 30 µg of the initial respective peptide with no poly(I:C) adjuvant. This booster was delivered in 100 µL PBS intradermally in the skin directly adjacent to the tumor. Animals were monitored for tumor growth via caliper measurement and survival every 2-3 days for the remainder of the study, with UNC IACUC defined endpoints of area >200 mm2 or ulceration >5 mm in the longest diameter.

For P815 treatment studies, 8-12 week old male BALB/c donors (Jackson Laboratory) were vaccinated on days 0 and 7 with 100 µg total peptide (3-4 pooled equimolar peptides, or no-peptide control) and 50 µg poly(I:C) adjuvant in 100 µL PBS intradermally in the flank. DBA/2 recipients were treated with 800 rad total body irradiation on day 13. On day 14, splenic-derived T cells and bone marrow cells were isolated from donor BALB/c animals. T cells were isolated from single-cell splenocyte suspensions uisng the Miltenyi Pan T Cell Isolation Kit II (cat. #130-095-130), according to manufacturer's protocol. Bone marrow cells were isolated as previously described[86]. Recipient DBA/2 animals were given tail-vein IV injections of 3x106 T cells, 3x106 bone marrow cells, and 3x105 P815-luciferase tumor cells (or bone-marrow only control). DBA/2 recipients were given a booster vaccine on day 21 after primary vaccine (100 µg total peptide, 50 µg poly(I:C)), with animals monitored every 2-3 days for survival, with UNC IACUC defined endpoints of bilateral hind-limb paralysis. Luciferase imaging studies were performed on days 8, 13, 22, 26, and 35 after transplant, using an IVIS imaging system on animals given 3 mg intraperitoneal D-luciferin (Perkin Elmer, cat. # 122799) 10 minutes prior to imaging.

Tetramer studies.

Peptide/MHC tetramer and cell surface protein staining were performed as described previously[87]. Briefly, viable, single-cell suspensions derived from tumors (approximately 107 total cells) were treated with 50 nM dasatinib (Sigma-Aldrich, cat. # CDS023389) for 30 minutes at 37°C, and then stained using approximately 10 µg/mL tetramer on ice for 30 minutes. Tetramers were generated using the MBL Quickswitch Quant H-2 Kb Tetramer Kit-PE (cat. # TB-7400-K1), using peptides VALLPSVMNL or SIINFEKL irrelevant control, according to manufacturer protocol. Cells were then washed and incubated on ice with biotin-conjugated anti-PE antibody (5 µg/ml; BioLegend; PE001) for 20 minutes, followed by 2 washes, then further incubation with streptavidin, R-PE conjugate (SAPE, 5 µg/mL) for 10 minutes on ice. Cells were then washed and stained for viability using BD fixable viability dye FVS620 according to the manufacturer's directions. Last, cells were Fc blocked (anti-mouse CD16/CD32; 2.4G2, BD Biosciences) for 10 minutes on ice, followed by surface staining for 20 minutes on ice with the following markers: CD45 (BV510; 30-F11), CD4 (FITC; RM4-5), CD8 (APC/Fire-750; 53-6.7) (All antibodies purchased from BD Biosciences). Aquisition was performed using a BD LSRFortessa flow cytometer. FlowJo flow cytometry software version 10 was used for analyses of all flow cytometric data. Cells were selected using gates defined by single color controls and FMO or irrelevant tetramer controls, with tetramer-positive/negative CD8+ T cells defined within live, singlet (by FSC-A versus FSC-H), CD45+, CD4–/CD8+ gates.

Ex vivo T-cell expansion and cytotoxicity assays

For tetramer-sorted T-cell isolation, CD8+ T cells were isolated from BBN963 tumor single-cell suspensions (as described above) using the Miltenyi Dead Cell Removal Kit (130-090-101) followed by the Miltenyi CD8a+ T Cell Isolation Kit (130-104-075), both according to manufacturer protocol. CD8+ T cells stained with tetramer as described above and sorted on the BD FACSJazz. Tetramer sorted T cells or column-sorted (Miltenyi CD8a+ T Cell Isolation Kit) CD8+ T cells from OT1 (C57BL/6-Tg(TcraTcrb)1100Mjb/J, Jackson Laboratories cat. # 003831) splenocytes were cultured in complete RPMI media (RPMI 1640 (Gibco cat. # 11875-093) with 10% FBS (Gemini cat. # 900-108), 1% sodium pyruvate (100nM; Gibco cat. # 11360-070), 1% non-essential amino acids (10mM; Gibco cat. # 11140-050), 1% l-glutamine (Gemini cat. # 400-106), 1% HEPES buffer (1Ml Corning cat. # 25-060-Cl), and .1% 2-mercaptoethanol (55nM; Gibco cat. # 2198502)) in the presence of recombinant murine IL7 (10 ng/mL, Peprotech cat. #

217-17), IL15 (10 ng/mL, Peprotech cat. # 210-15), and IL2 (100 IU/mL, Peprotech cat. # 212-12) for 72 hours. Antigen-specific T-cell expansion was performed using a previously described protocol[88]. Briefly, all sorted T cells were recovered at 106 cells/mL for 72 hr in RPMI complete media in the presence of IL7 (10 ng/mL), IL15 (10 ng/mL), and IL2 (10 ng/mL) at 37°C and 5% CO2. T cells were then cocultured in RPMI complete media and IL7, IL15, and IL2, alongside peptide-pulsed DCs (2.5 µg/mL peptide) which had been pulsed approximately 18 hr prior to coculture. Media and cytokines were changed every 2-3 days, letting cells expand for 7-10 days after coculture with peptide-pulsed dendritic cells before downstream assays.

For flow cytometric-based cytotoxicity assays, target cells (BBN963 or an irrelevant splenocyte control) were pre-labelled in 5 µM CFSE for 15 minutes prior to co-culture. Tetramer-sorted and antigen-expanded T cells (per above section) were cocultured alongside targets at a 1:1 ratio, with 1x105 of each target and effector population. Cells were plated on a v-bottom 96-well polypropylene plate, centrifuged at 300 x g for 1 minute, and incubated at 37°C, 5% CO2 for 4 hours. After incubation, cells were stained using FVS700 viability dye (BD Biosciences, cat. # 564997) according to manufacturer's directions. Aquisition was performed on a BD LSRFortessa flow cytometer. FlowJo flow cytometry software version 10 was used for analyses of all flow cytometric data. Cells were identified as targets (CFSE+) or effectors (CFSE–), looking for percent viability among targets. Percent killing was reported as frequency of dead targets, background subtracted from no-effector control wells.

The cytotoxic activity of T cells was evaluated using a standard 4-hour 51Cr release assay[89]. In brief, 5x103 51Cr-labeled (Perkin Elmer cat. # NEZ030001MC) BBN963 target cells per well were plated in triplicate in a 96 well v-bottom plate with different ratios (10:1 and 5:1

effector:target) of effector cells and incubated for 4 hours at 37°C. The supernatant was collected and analyzed with a gamma-counter (Perkin Elmer). Before labeling, target cells were incubated for 2 hours at 37°C with the specific peptides (100 nM) and washed twice with complete medium. Target cells were incubated with medium alone or in 1% Triton X-100 (Sigma-Aldrich) to determine the spontaneous and maximum 51Cr release, respectively. The mean percentage of specific lysis of triplicate wells was calculated as follows: [(test counts - spontaneous counts) / (maximum counts - spontaneous counts)] x 100%.

TCGA data analyses

MapSplice-aligned, RSEM-quantified RNA-Seq expression matrices and survival data were downloaded from FireBrowse (http://firebrowse.org/). Expression matrices were merged between all cancer types, upper quartile normalized within each sample, and log2 transformed. Immune gene signatures (IGS) were derived from previously described signatures[12,90–93], with expression calculated as the mean expression of each gene within the signature. TCGA LAML samples were omitted from analysis in order to prevent skewing of IGS patterns. Inclusion criteria were those defined by TCGA pan-immune working group, according to previous studies[65]. MHC I neoantigen expression used for machine-learning algorithm immunogenicity prediction were obtained from publicly available data derived in previous studies[65]. TCGA pan-cancer dataset (n = 11,092; LUAD n = 515, COAD n = 283) analyses were performed according to the above “Computational analysis” methods section.

Differential gene expression analysis was performed using DESeq2 (https://bioconductor.org/packages/release/bioc/html/DESeq2.html)[94]. Gene set enrichment analysis (http://software.broadinstitute.org/gsea/index.jsp)[95], Ingenuity pathway analysis (https://www.qiagenbioinformatics.com/products/ingenuity-pathway-analysis/)[96], and DAVID gene ontology analysis (https://david.ncifcrf.gov/)[97] were performed from respective web portals. T-cell receptor diversity analysis was performed from previously published MiXCR-derived TCR reads[65] – MiXCR is an analytic tool for TCR inference from whole transcriptome RNA-seq data[98].

Statistical analyses

Statistical analyses for survival (displayed as Kaplan-Meier plots) were performed using log rank test, with no statistical correction. Differences in cytotoxicity in tetramer-sorted cytotoxicity assays were determined via Welch's t-test, with no statistical correction. Differences in tetramer-positive T cell populations were determined via Mann-Whitney u-test, with no statistical correction. All above analyses were performed in Graphpad Prism 8. All other analyses and corresponding statistical tests are described in the above "Computational analysis" methods section.

Supplemental material

All supplemental figures and tables cited in Chapter 3 are listed according to the original published manuscript, which can be found at https://cancerimmunolres.aacrjournals.org/content/7/10/1591.figures-only.

CHAPTER 4 BIOINFORMATICS ANALYSIS OF BLADDER TUMORS AND BREAST CANCER BRAIN METASTASES[2]

4.1 Immunogenomic characterization of Bladder Tumors

4.1.1 Introduction

In the United States, bladder cancer is the fourth most common malignancy in men, with approximately 74,000 new cases and 16,000 deaths expected in 2015. Bladder cancer is histologically divided into low-grade or high-grade tumors that are associated with distinct genomic alterations and differences in prognosis[99]. Low-grade tumors are almost uniformly noninvasive (Ta) and have a 5-year survival rate of 96%. In contrast, high-grade tumors can become muscle-invasive and metastatic and are associated with a 5-year survival rate ranging from 70% (muscle-invasive) to 5% (metastatic).

Multiple studies have now identified distinct RNA expression subtypes within both low- and high-grade bladder cancer[100–106]. Building upon the work of our colleagues, we and others have recently described distinct subtypes of high-grade muscle-invasive urothelial carcinoma (UC), luminal and basal, that reflect attributes of their corresponding breast cancer subtypes. These studies highlight the similarities in the underlying biology between breast and bladder cancer[101,106]. In addition to the originally reported molecular subtypes of breast cancer (luminal A, luminal B, her2-enriched, and basal-like), a claudin-low subtype of breast cancer has been

[2] Chapter 4.1 has been was originally published as Kardos J, Chai S, Mose LE, Selitsky SR, Krishnan B, Saito R, et al. Claudin-Low Bladder Tumors are Immune Infiltrated and Actively Immune Suppressed. JCI Insight 1, (2016)

more recently identified and is characterized by a stromal phenotype, lack of luminal differentiation marker expression, enrichment for epithelial-to-mesenchymal transition (EMT) markers, cancer stem cell–like features, and immune response genes[92].

Clinical trials using immune checkpoint Abs targeting the PD1/PD-L1 axis have recently shown promise in a portion of patients with advanced UC, with the premise that activation of immune checkpoint pathways, including PD-L1, results in active immunosuppression[107]. Despite the excitement surrounding PD1/PD-L1 axis inhibition in treating advanced UC, only approximately 30% of patients respond. Therefore, the majority of patients display intrinsic resistance to PD1/PD-L1 inhibition, and a priori identification of these patients would clearly be beneficial.

We report here the discovery of a claudin-low subtype of high-grade, muscle-invasive UC defined by biologic characteristics of the claudin-low subtype of breast cancer. Claudin-low tumors were uniformly enriched for immune gene signatures but simultaneously expressed immune checkpoint molecules, demonstrating that, despite being immune infiltrated, claudin-low tumors are also actively immunosuppressed. Interestingly, the predicted neoantigen burden was not significantly increased in claudin-low tumors. Instead, they highly expressed cytokines and chemokines associated with leukocyte chemotaxis into the tumor immune microenvironment as a result of an imbalance between PPARγ and NF-κB signaling. These results highlight the association between molecular subtype and the degree of immune infiltration and immune suppression and suggest that mechanisms other than neoantigen burden can drive the development of immune infiltrated tumors and also that claudin-low tumors are poised to respond to immune checkpoint inhibition.

4.1.2 Results

Identification of a claudin-low subtype in bladder cancer.

Previous studies have identified a claudin-low molecular subtype of breast cancer[108]. Given the previously documented similarities in gene expression patterns between breast and bladder cancer[101,106], we asked whether a claudin-low subtype also exists in bladder cancer. To this end, we performed unsupervised hierarchical clustering on 408 high-grade, muscle-invasive bladder tumors from the The Cancer Genome Atlas (TCGA) urothelial bladder carcinoma (BLCA) data set using gene signatures representative of biologic characteristics that are known to define breast cancer claudin-low tumors such as an enrichment for tumor-initiating cells (TICs) and an EMT (Figure 4.1A)[92,108]. Specifically, we used gene lists of the tight-junction claudins (*CLDN3*, *CLDN4*, and *CLDN7*) and a previously published bladder cancer–derived TIC signature[90]. In addition, we derived a bidirectional (EMT_UP and EMT_DOWN), bladder cancer–specific, notch-dependent EMT gene signature from the publicly available Gene Expression Omnibus (GEO) gene expression data set (GEO GSE60564) (Supplemental Table 1). Unsupervised hierarchical clustering with these gene signatures revealed a distinct cluster that had characteristics of a claudin-low subtype (Figure 4.1A, highlighted in green).

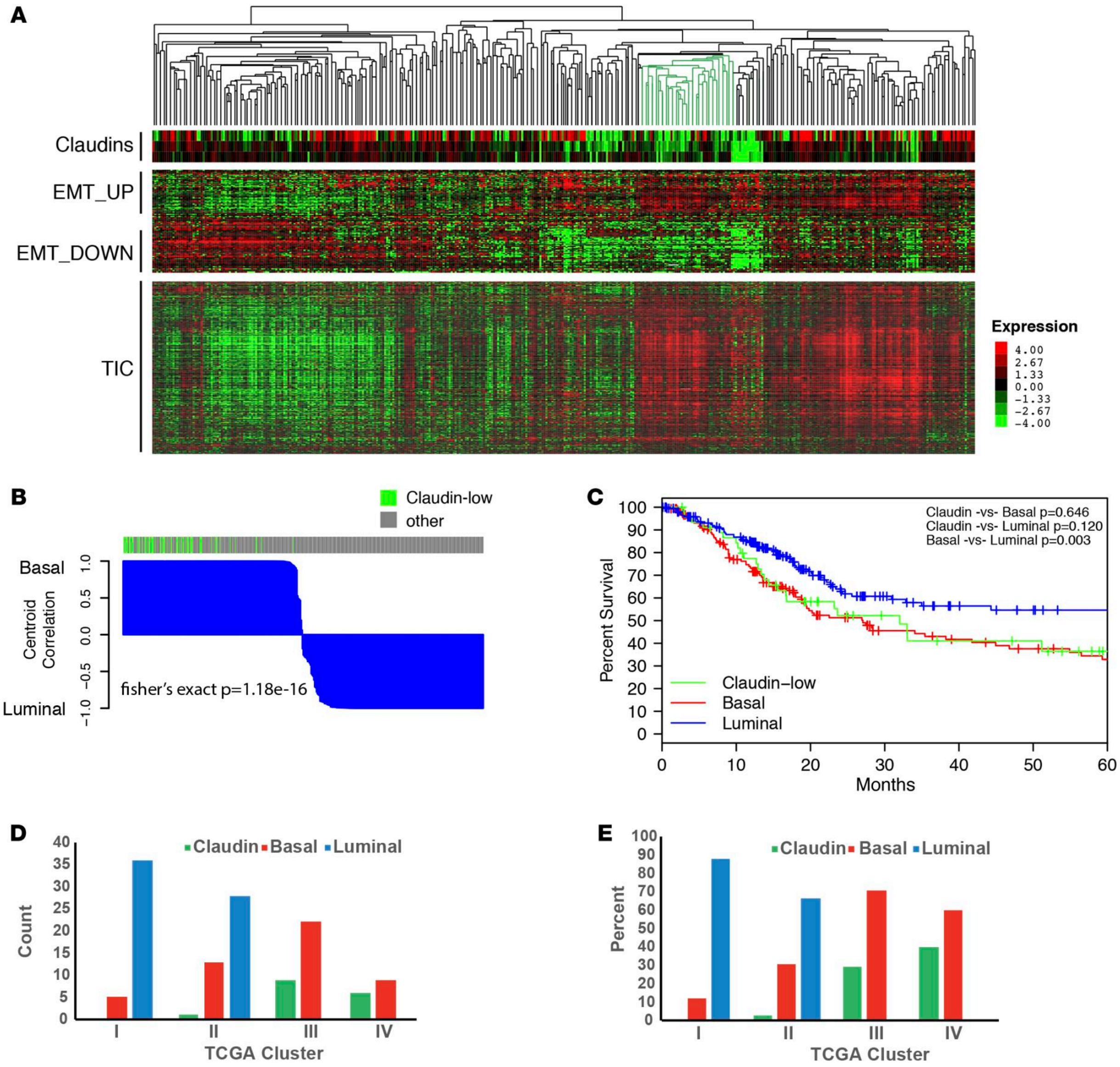


Figure 4.1 Identification of a claudin-low subtype in bladder cancer. (A) Unsupervised clustering of TCGA muscle-invasive UC samples. Samples were clustered on the basis of expression of tight-junction claudins, a bidirectional EMT signature, and a TIC signature. The tumors identified as claudin-low are highlighted in green on the dendogram. *n* = 408. (B) Waterfall plot showing correlation with the basal and luminal centroids as defined by BASE47 classification; claudin-low tumors are highlighted in green. Claudin-low tumors were significantly enriched in the BASE47 basal subtype (Fisher's exact test $P = 1.18 \times 10^{-16}$) and were highly correlated with the basal centroid (Pearson's correlation $P = 9.33 \times 10^{-15}$). *n* = 408. (C) Kaplan-Meier plot showing overall survival of bladder cancer by molecular subtype. Significance was determined by log-rank testing with a Bonferroni correction. *n* = 408. (D and E) Bar graphs showing the classification of TCGA UC tumors by TCGA mRNA cluster subtype (*x* axis) and our subtype classifications (*y* axis) by count and percentage. *n* = 129. EMT, epithelial-to-mesenchymal transition; TCGA, The Cancer Genome Atlas; TIC, tumor-initiating cell; UC, urothelial carcinoma.

To ensure that the set of tumors within the presumed claudin-low cluster were homogeneous and distinct from adjacent clusters of tumors, we performed a Gaussian distribution analysis, starting with the smallest cluster and iteratively repeated the analysis with the addition of adjacent clusters using SigClust R software (Supplemental Figure 1A)[109]. This method identified a conserved node of 48 tumors that had consensus enrichment for claudin-low features, and these tumors were therefore defined as claudin- low. All 48 claudin-low tumors were classified as basal by our BASE47 subtype classifier (Fisher's exact *P* = 1.18 × 10–16)[101], and when examined for their correlation to the BASE47 basal or luminal centroid, they were found to be highly basal (Figure 4.1B). Further supporting the notion that these tumors exhibit features of claudin-low breast cancer, we applied the previously defined breast cancer–specific claudin-low classifier to the TCGA BLCA tumors and found a significant enrichment (Fisher's exact *P* = 1.10 × 10–18) of the breast cancer–defined claudin-low tumors within the bladder claudin-low cluster (Supplemental Table 2). Given these findings, we propose a 3-subtype classification of bladder cancer consisting of basal (~40%), luminal (~50%), and claudin-low (~10%) tumors. While basal-like bladder cancer consistently has a worse clinical outcome [100,101,105,106], consistent with previous work on breast cancer[92], we did not find an observable significant difference in overall survival rates between patients with claudin-low tumors and those with basal tumors (Figure 4.1C).

A 40-gene classifier, bladder claudin-low 40, accurately predicts claudin-low tumors.

To define a minimal set of genes that could accurately classify claudin-low bladder tumors, we applied prediction analysis of microar- rays (PAMs) to the TCGA BLCA tumors and derived a 40-gene signature, bladder claudin-low 40 (BCL40) (Supplemental Table 3), which accurately classifies bladder tumors into claudin-low and non–claudin-low subtypes, with a training error rate of 0.23 and 0.13, respectively. When combined with the previously validated bladder cancer analysis of subtypes by gene expression (BASE47) predictor[101], this provides a 3-class predictor that can accurately classify bladder tumors as claudin-low, basal, or luminal.

In order to validate the predictor, we compiled a 130-tumor metadata set from 2 previously compiled published data sets (GEO GSE48277)[106]. The BASE47 and BCL40 predictors identified 36 claudin-low tumors (~30%), 27 basal tumors (~20%), and 67 luminal tumors (~50%). We found that these subtypes were phenotypically similar to the initially derived subtypes in our discovery set of TCGA bladder tumors as measured by expression of the EMT, TIC, and claudin gene signatures (Supplemental Figure 1, B–E). Furthermore, we ran a transcriptome-wide correlation analysis between the basal, luminal, and claudin-low tumors identified in the discovery (TCGA BLCA) and validation data sets (GEO GSE48277) and found a strong correlation in gene expression between the subtypes identified in the discovery and validation data sets (basal [Pearson's R = 0.459, P < 2.2 × 10–16], claudin-low [Pearson's R = 0.805, P < 2.2 × 10–16], and luminal [Pearson's R = 0.809, P < 2.2 × 10–16]) (data not shown). This further confirmed that the subtypes identified across separate data sets had consistent genome-wide RNA expression profiles.

Comparison with MD Anderson and TCGA UC intrinsic molecular subtype classifications.

We next examined whether our claudin-low subtype merely recapitulated any of the existing molecular subtypes of UC published by MD Anderson or TCGA. A comparison of our claudin-low, basal, and luminal predictions on the 408 provisional TCGA BLCA tumors with the MD Anderson oneNN classification system (p53-like, basal, and luminal)[106] re-demonstrated the high concordance of luminal subtype designations[110] as well as the notion that claudin-low tumors arise primarily from basal tumors (Supplemental Figure 2, A and B). We further compared our claudin-low, basal, and luminal predictions on the 129 published TCGA BLCA tumors with TCGA 4-subtype classification (clusters I, II, III, and IV)[102]. Our claudin-low tumors were primarily found in TCGA clusters III and IV (Figure 4.1, D and E). These comparisons further strengthen the notion that claudin-low tumors do not merely recapitulate a previously described molecular subtype of bladder cancer.

The claudin-low subtype displays unique, intrinsic genomic alterations and gene expression patterns.

We next examined the association between molecular subtype and genomic events within significantly mutated or copy number–altered genes identified as being altered at a greater than 5% frequency within TCGA BLCA data set[102]. A comparison of claudin-low and basal subtypes revealed that claudin-low tumors had significantly increased rates of RB1, EP300, and NCOR1 mutations, an increased percentage of tumors with EGFR amplification, as well as decreased rates of mutations in FGFR3 and ELF3 (Figure 4.2, A and B). Relative to the luminal subtype, claudin-low tumors revealed a significantly higher rate of mutation of TP53, RB1, and EP300 and an increased percentage of tumors with EGFR amplification. Conversely, luminal tumors (compared with claudin-low tumors) had a significantly higher rate of PPARG amplification and mutation of KDM6A, ELF3, and FGFR3. These results are in keeping with the notion that genomic alterations and their subsequent effects on signal transduction and transcription may be partially responsible for differences in gene expression subtypes.

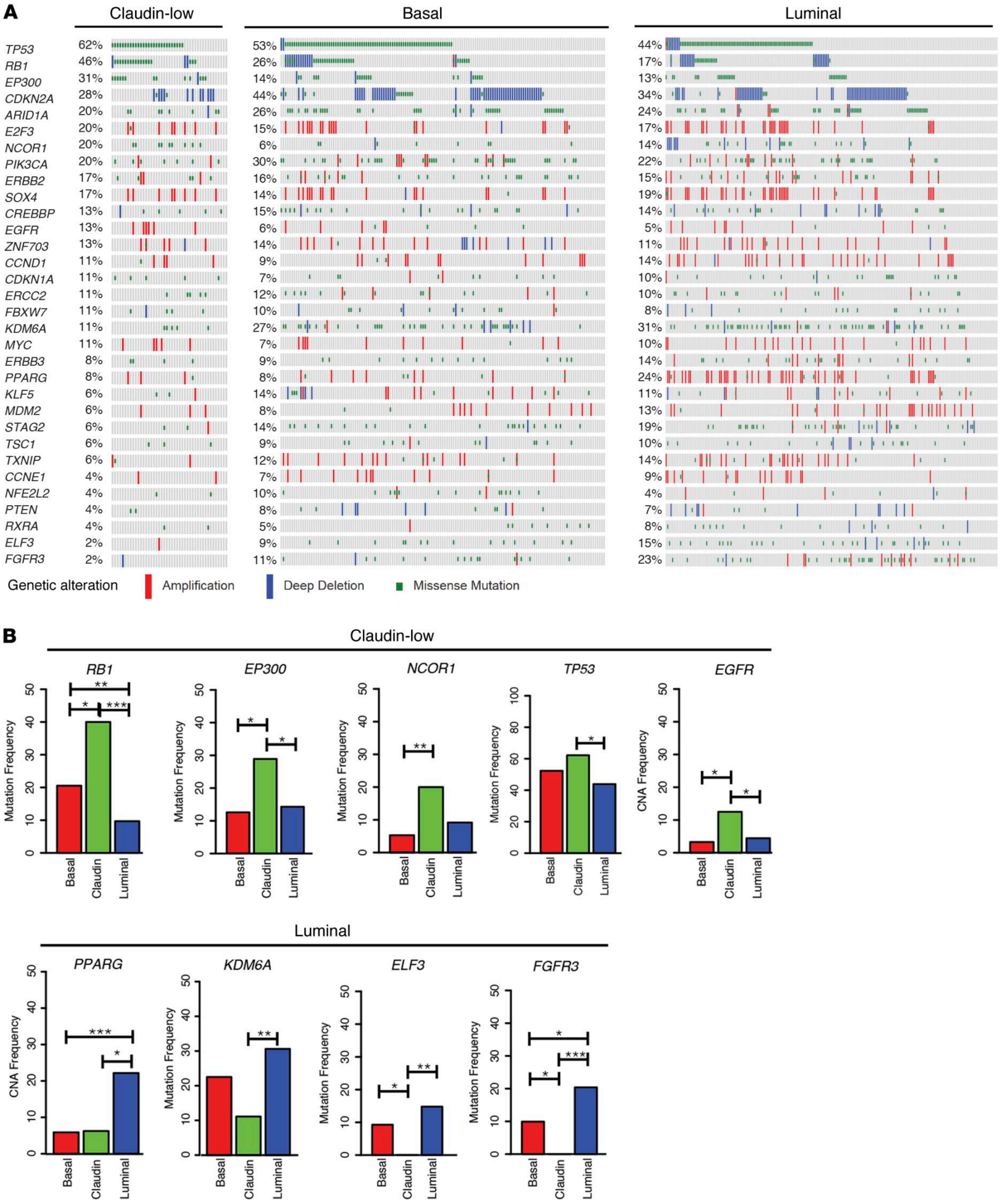


Figure 4.2 Genomic characterization of bladder cancer subtypes. (A) Oncoprint of genomic copy number alterations and mutations by bladder cancer subtype for genes previously identified as significantly mutated or copy number altered in more than 5% of bladder tumors. n = 408. (B) Bar plots of genes that were identified to have a significant ($P < 0.05$) difference in either gene mutation or copy number alteration (CNA) between the claudin-low and basal and/or luminal subtypes. *$P < 0.05$, **$P < 0.01$, and ***$P < 0.001$, by Fisher’s exact test.

To further understand the gene expression patterns that differentiate claudin-low tumors, we performed 2-class significance analysis of microarrays (SAMs), comparing each subtype against all of the other tumors (e.g., claudin-low vs. basal plus luminal). We detected a significant number of differentially expressed genes (FDR = 0.05) (Supplemental Figure 3A and Supplemental Table 4) by this comparison as well as by a direct comparison of each subtype with another (e.g., claudin-low vs. basal) (Figure 4.3A and Supplemental Table 5). Ingenuity Pathway Analysis (IPA) revealed that, compared with both basal and luminal tumors, claudin-low tumors had significant enrichment in the upstream regulators IFNG, TNF, and TGFB1, which are well-known proinflammatory cytokines (IFN-γ and TNF-α) and pro-EMT (TGF-β) growth factors (Supplemental Table 6). Additionally, claudin-low tumors had higher levels of IL4 and IL13 signaling relative to signaling levels in basal and luminal tumors, respectively. Further IPA analysis demonstrated enrichment of other immune-associated pathways in claudin-low tumors (Supplemental Figure 3B). These observations are in keeping with the EMT phenotype, which is a defining characteristic of claudin-low tumors, but are also strongly suggest that these tumors are heavily immune infiltrated.

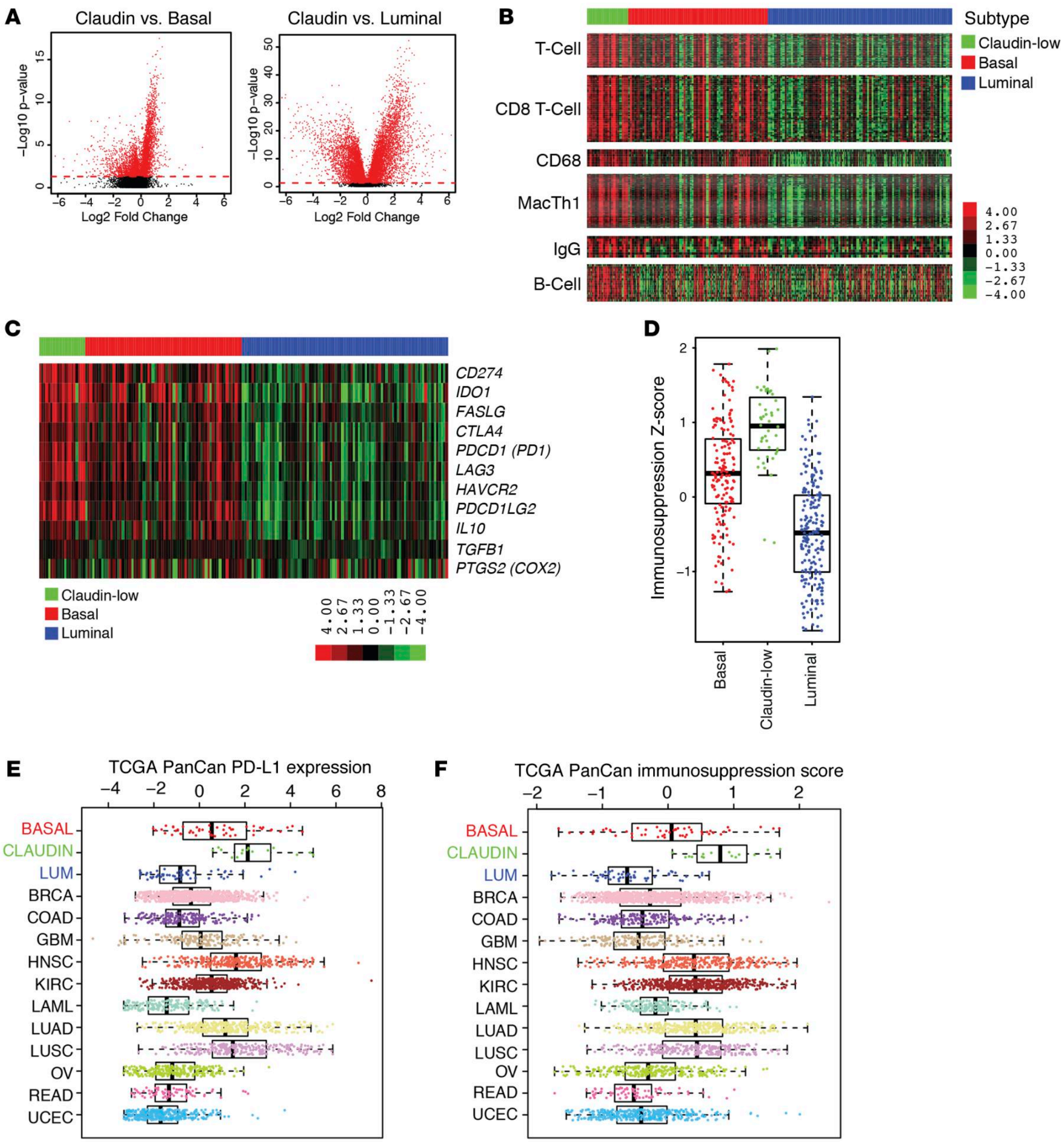


Figure 4.3 Immune characterization of bladder cancer subtypes. (A) Volcano plot of $\log_2$ fold change of median gene expression and $-\log_{10} P$ value of gene expression across bladder tumor subtypes. Dashed line across the plots corresponds to a significance threshold of $P = 0.05$. $n = 408$. Significance was calculated using Student's t test with a Bonferroni correction. (B) Heatmaps of supervised clustering of bladder tumor subtypes across previously identified immune signatures. $n = 408$. (C) Heatmap of supervised clustering of bladder tumor subtypes across an immune suppression gene signature. $n = 408$. (D) Box plot of immune suppression gene signature z score across bladder tumor subtypes. $n = 408$. (E) Box plot of PD-L1 gene expression across the Pan-Cancer tumor types. $n = 3,602$. (F) Box plot of immune suppression gene signature z scores across the Pan-Cancer tumor types. $n = 3,602$. The box plots denote the interquartile range (IQR), with the box representing Q1 to Q3, the line denoting Q2, and the whiskers extending an additional 1.5 times the IQR beyond Q1 and Q3. The dots represent data points. BLCA, bladder urothelial carcinoma; BRCA, breast cancer; COAD, colon adenocarcinoma; GBM, glioblastoma multiforme; HNSC, head and neck squamous cell

carcinoma; KIRC, kidney renal clear cell carcinoma; LAML, acute myeloid leukemia; LUAD, lung adenocarcinoma; LUSC, lung squamous cell carcinoma; OV, ovarian serous cystadenocarcinoma; READ, rectum adenocarcinoma; UCEC, uterine corpus endometrial carcinoma; LUM, luminal; TCGA, The Cancer Genome Atlas; PanCan, Pan-Cancer.

Claudin-low tumors are enriched in immune gene signature expression.

To better characterize the immune cell populations present within claudin-low tumors, we used previously defined gene signatures indicative of specific cellular immune populations[91] and examined their expression by molecular subtype. All examined signatures appeared to be and were statistically enriched in the claudin-low subtype when each signature was collapsed into a single value per tumor (z score) (Figure 4.3B and Supplemental Figure 4). To assess the level of immunosuppression, we examined the expression of a panel of immune checkpoint molecules (immunosuppression score) derived from the literature and found that they were uniformly highly expressed in claudin-low tumors compared with expression levels in both basal and luminal tumors, respectively (Figure 3.3, C and D).

Bladder cancer as a whole expressed moderate levels of PD-L1 and our immunosuppression score relative to the spectrum of 12 tumors in TCGA Pan-Cancer analysis (Supplemental Figure 5, A and B)[111]. When broken down by subtype, however, claudin-low tumors in particular had very high levels of PD-L1 expression (Figure 4.3E) and high expression of the immunosuppression score (Figure 4.3F). In aggregate, these findings indicate that claudin-low tumors consistently harbor a high level of immune infiltration that is matched by a high level of active immune suppression. Basal tumors, in contrast, have a more heterogeneous phenotype, while luminal tumors appear to have a paucity of immune cells or immune checkpoint expression. In keeping with this notion, there was a strong correlation between the immune signatures and the immunosuppression signature (Supplemental Figure 5C) across all tumors.

The presence of an immune infiltrate has been shown to be prognostic in other cancers[112]. In muscle-invasive bladder cancer, specifically, the presence of CD8+ tumor-infiltrating lymphocytes (TILs)[113] and a low ratio of FOXP3 to CD4 or CD8 expression on TILs[114] have been associated with improved disease-free and overall survival rates. In keeping with the work by Sharma et al.[113], Cox proportional hazards (Cox PH) modeling for each immune gene signature across all tumors in TCGA BLCA data set showed that only the CD8_T_Cell signature was prognostic (Cox PH = 0.846, P = 0.047) (Figure 4.4A), further supporting the unique importance of CD8+ TILs. When Cox PH modeling was performed within each subtype, none of the signatures were prognostic within the claudin-low and luminal subtypes. However, within the basal subtype, numerous signatures were prognostic, including the Ig signature, macrophage signature, T cell signature, CD8+ T cell signature, and immunosuppression signature (Figure 4.4B). We believe these findings are consistent with immune gene signatures being consistently upregulated in the claudin-low subtype and downregulated in the luminal subtype, respectively, while the basal subtype has a more heterogeneous range of gene signature expression, allowing for a more dynamic range across which these subtypes can be prognostic. Supporting this, the basal subtype had the largest SD of immune signature expression across all signatures (basal vs. claudin: P = 0.007; basal vs. luminal: P = 0.097, Bonferroni-corrected Student’s t test).

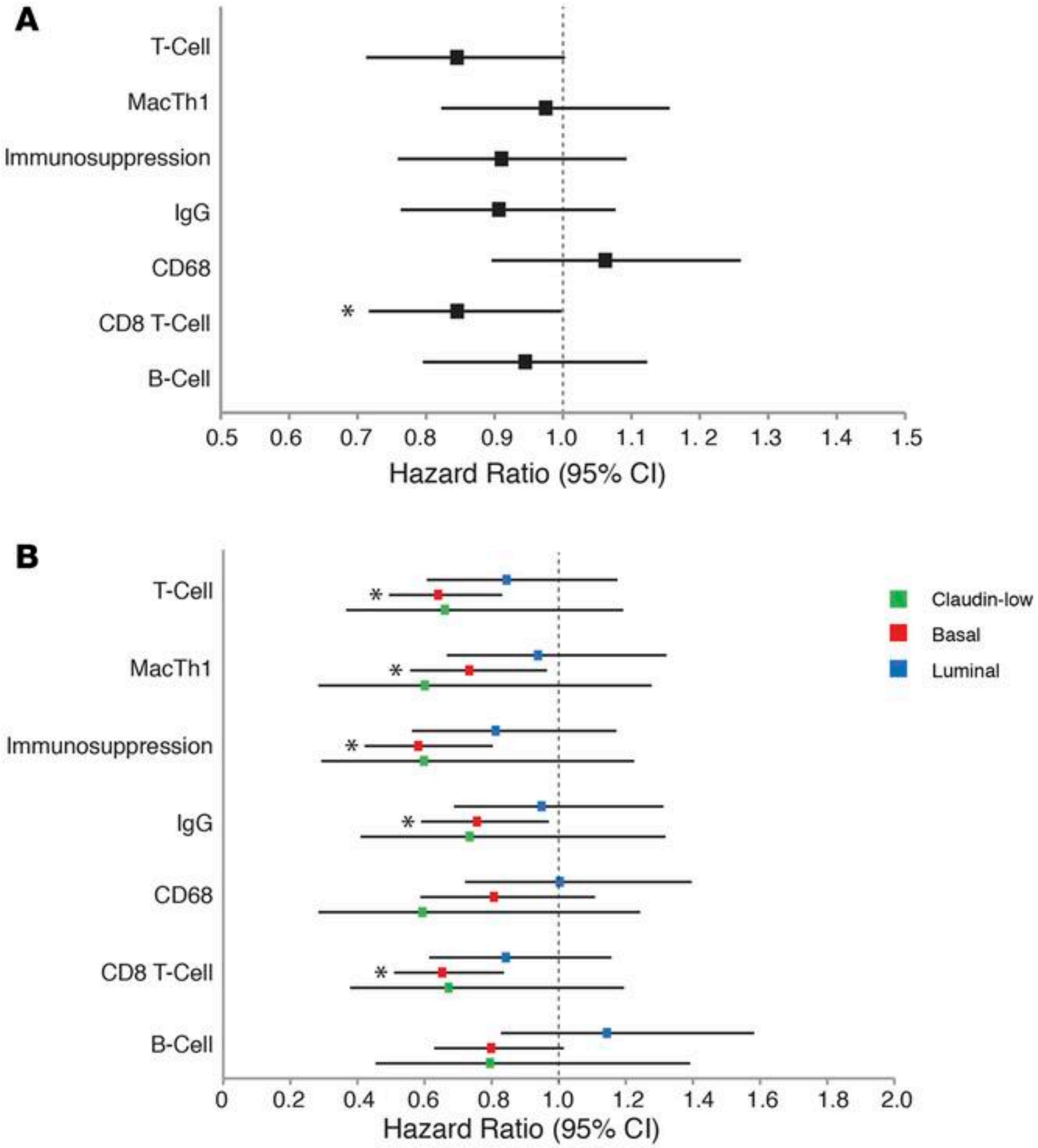


Figure 4.4 Immune gene signatures have prognostic value across bladder cancer subtypes. (A) Forest plot of Cox PH ratios of the immune gene signatures across all tumors, with a 95% CI indicated around the values. *n* = 408. (B) Forest plot of Cox PH ratios of the immune gene signatures within defined tumor subtypes, with a 95% CI indicated around the values. *n* = 408. **P* < 0.05, prognostically significant signatures by Cox PH modeling. Cox PH, Cox proportional hazard.

Specific T cell receptor and B cell receptor gene segment expression levels are prognostic in bladder cancer subtypes.

An antigen-driven T cell and/or B cell response would be expected to drive clonal expansion of T cells and/or B cells, resulting in decreased diversity of T cell receptor (TCR) and/or B cell receptor (BCR) repertoires. In addition, if a clonally expanded immune response was active intratumorally, this should be reflected in associations of specific TCR and/or BCR gene segment expression with improved survival. For example, decreased TCR diversity has been associated with response to immune checkpoint inhibition in melanoma[115] and has been shown to be prognostic in bladder cancer[116]. To evaluate this concept in TCGA bladder samples, we fit Cox PH models to test the association of expression of each TCR or BCR gene segment with survival and calculated the number of prognostic gene segments by subtype. To establish null distributions for the number of gene segments expected in each subtype, we used the bootstrap resampling method previously published by our group[91]. For both TCR gene segments (Figure 4.5A) and BCR gene segments (Figure 4.5B), a significantly higher number of gene segments than expected by chance were prognostic in the basal subtype, but not in the claudin-low or luminal subtype. Figure 4.5, C and D show the specific gene segments that were prognostic in each subtype. Prognostic segments were found in multiple TCR and BCR families, with a small number of gene segments discovered in multiple subtypes (i.e., TRBV11-2). This suggests that adaptive immune responses important in endogenous antitumor immunity are not uniform in TCR and BCR usage between the subtypes.

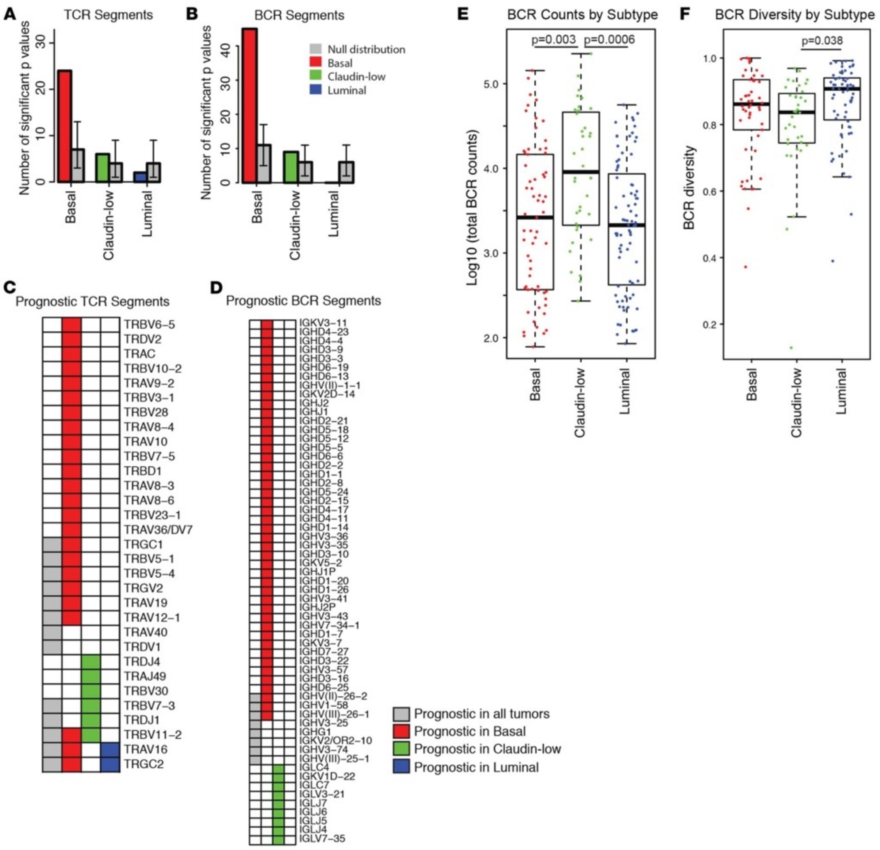


Figure 4.5 BCR and TCR segment expression is prognostic. (A) Number of TCR gene segments by subtype in which increased expression was significantly associated with improved survival by Cox PH model fit. Null distributions (gray bars) with 95% CIs were generated for each by bootstrap resampling of non-TCR genes and calculation of the number of significant *P* values that were similarly associated with prolonged survival. *n*= 292. (B) Number of BCR gene segments by subtype in which increased expression was significantly associated with improved survival by Cox PH model fit. Null distributions (gray bars) with 95% CIs were generated for each by bootstrap resampling of non-TCR genes and calculation of the number of significant *P* values that were similarly associated with prolonged survival. *n* = 292. (C) Specific TCR gene segments in which increased expression was significantly associated with improved survival by Cox PH model fit for all tumors (gray boxes), basal tumors (red boxes), claudin-low tumors (green boxes), and luminal tumors (blue boxes). (D) Specific BCR gene segments in which increased expression was significantly associated with improved survival by Cox PH model fit for all tumors (gray boxes), basal tumors (red boxes), claudin-low tumors (green boxes), and luminal tumors (blue boxes). (E) Log base 10 number of reads supporting any BCR V(D)J rearrangement are shown by subtype. *n* = 181. Mann-Whitney *U*–Wilcoxon test with an FDR multiple testing correction was used to determine significance. (F) Repertoire diversity by subtype. The box plots in E and F denote the interquartile range (IQR), with the box representing Q1 to Q3, the line denoting Q2, and the whiskers extending an additional 1.5 times the IQR beyond Q1 and Q3. The dots represent data points. *n* = 150. Mann-Whitney *U*–Wilcoxon test with an FDR multiple testing correction was used to determine significance. BCR, B cell receptor; Cox PH, Cox proportional hazard; TCR, T cell receptor.

Despite the presumed importance of assessing T cell and B cell clonality in tumor immunology, at present, this can only be done by direct TCR or BCR sequencing. Our group has developed a bioinformatics method (VDJician) to accurately and efficiently reconstruct rearranged BCR V(D)J sequence repertoires from short-read RNA-sequencing data. We applied this to TCGA bladder data to evaluate whether overall BCR expression (Figure 4.5E) and/or repertoire diversity (Figure 4.5F) varied by subtype. BCR expression was higher and repertoire diversity lower (indicative of clonal expansion) in the claudin-low subtype relative to that observed in the luminal subtype, which is consistent with the presence of a selective antigen-directed response in claudin-low tumors. These results, in conjunction with our previous findings, indicate that claudin-low tumors are immune infiltrated and have an active immune response within the tumor microenvironment.

Predicted neoantigen burden does not vary significantly by bladder cancer subtype but is selectively associated with survival in basal tumors.

Neoantigens are altered peptides derived from tumor-intrinsic mutant proteins that are presented by MHC molecules and can drive robust antitumor T cell responses[4]. This is in contrast to self-antigens that may be overexpressed in tumors but have been subjected to central immune tolerance [117]. Neoantigens derived from tumor-specific genomic aberrations can be predicted using whole-exome sequencing of paired tumor and matched normal samples, and expression is confirmed by incorporation of RNA expression data. The predicted neoantigen number has been positively associated with favorable clinical outcomes in multiple tumor types[118] as well as with response to immune checkpoint inhibition in melanoma[10,119] and non–small-cell lung cancer[120]. These results suggest an important protective role for the endogenous repertoire of T cells able to target tumor cells. In order to determine whether neoantigen burden varied by bladder tumor subtype, we implemented an informatics pipeline based on the approach published by Rajasagi et al.[28] and applied this to TCGA bladder data (Figure 4.6). There was a noisy but clear correlation between predicted neoantigen burden and the number of somatic mutations (Figure 4.6A: left y axis and right y axis, respectively) (Spearman’s R = 0.79, P < 2 × 10–16, Figure 4.6B). Interestingly, claudin-low tumors, despite having a high level of immune infiltration and active immunosuppression, did not have a significantly different level of predicted neoantigens compared with that of basal or luminal subtypes (Figure 4.6C).

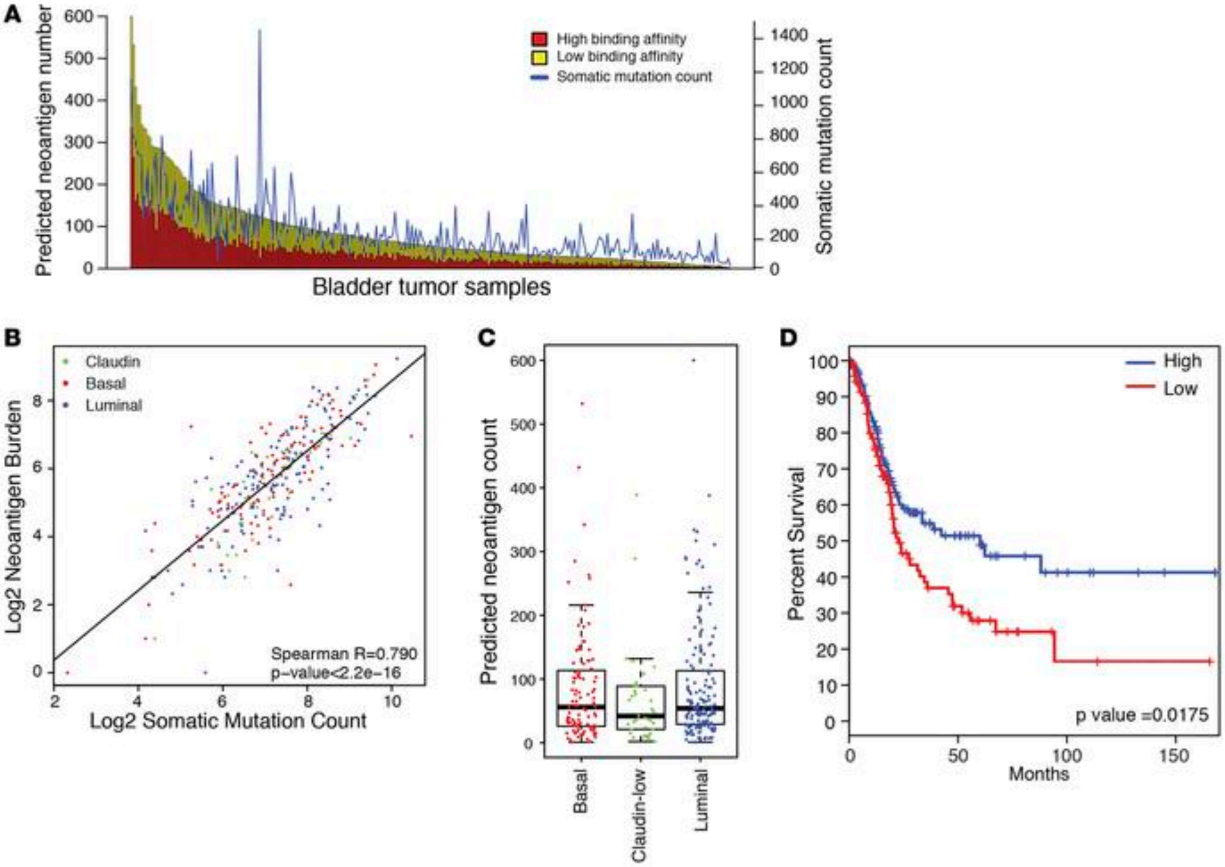


Figure 4.6 Predicted neoantigen burden by bladder cancer subtype. (A) Stacked bar plot showing the number of predicted neoantigens in each bladder tumor with a predicted IC_{50} of less than 50 nm (red bars) and less than 150 nm (yellow bars). Numbers of predicted neoantigens are shown

in the left *y* axis. Blue line and right *y* axis show the number of missense mutations per tumor. *n* = 289. (B) Scatter plot of somatic missense mutations ($\log_2$) versus predicted neoantigen burden ($\log_2$) across TCGA data set. Significance and correlation were determined using Spearman's rank test. *n* = 289. (C) Box plot showing the number of predicted neoantigens with an IC_{50} of less than 50 nm by tumor molecular subtype. Subtypes were not significantly different (*P* > 0.05). Significance was determined by 1-way ANOVA. *n* = 289. The box plots denote the interquartile range (IQR), with the box representing Q1 to Q3, the line denoting Q2, and the whiskers extending an additional 1.5 times the IQR beyond Q1 and Q3. The dots represent data points. (D) Kaplan-Meier plot showing survival of bladder cancer patients with high (greater than median value, blue line) versus low (less than median value, red line) predicted numbers of neoantigens. Vertical hash marks indicate censored data. Significance was determined by log-rank test. *n* = 289. TCGA, The Cancer Genome Atlas.

To assess the association between predicted neoantigen burden and subtype, we performed Cox PH analysis with the predicted neoantigen count as the potential explanatory variable. In the basal but not claudin-low or luminal subtypes, an increased number of predicted neoantigens was associated with prolonged survival (P = 0.025). For all bladder tumors taken together, the association was significant as well (P = 0.005). Figure 4.6D shows survival curves for all bladder tumors divided by the median predicted neoantigen count into high versus low neoantigen burden. Analyzed in this way as well, high neoantigen burden was associated with prolonged overall survival. Therefore, while there is a high correlation between bladder cancer molecular subtype and immune signature expression, this does not appear to be explained by the predicted neoantigen number.

Claudin-low tumors express high levels of cytokines and chemokines normally repressed by PPARG.

Given that predicted neoantigen burden was relatively similar across molecular subtypes, we explored the possibility that claudin-low tumors harbor an immune infiltrate because of increased production of proinflammatory cytokines and chemokines. To this end, we examined the relative expression of a panel of cytokines and chemokines (Supplemental Table 7) and their receptors among bladder subtypes and found that the majority of them were significantly upregulated in claudin-low tumors relative to expression levels in both basal and luminal tumors (Figure 4.7, A and B). We noted that NF-κB target genes in particular were significantly upregulated in the claudin-low subtype compared with expression in both the basal and luminal subtypes (Fisher's exact P value = 1.885 × 10–8, data not shown).

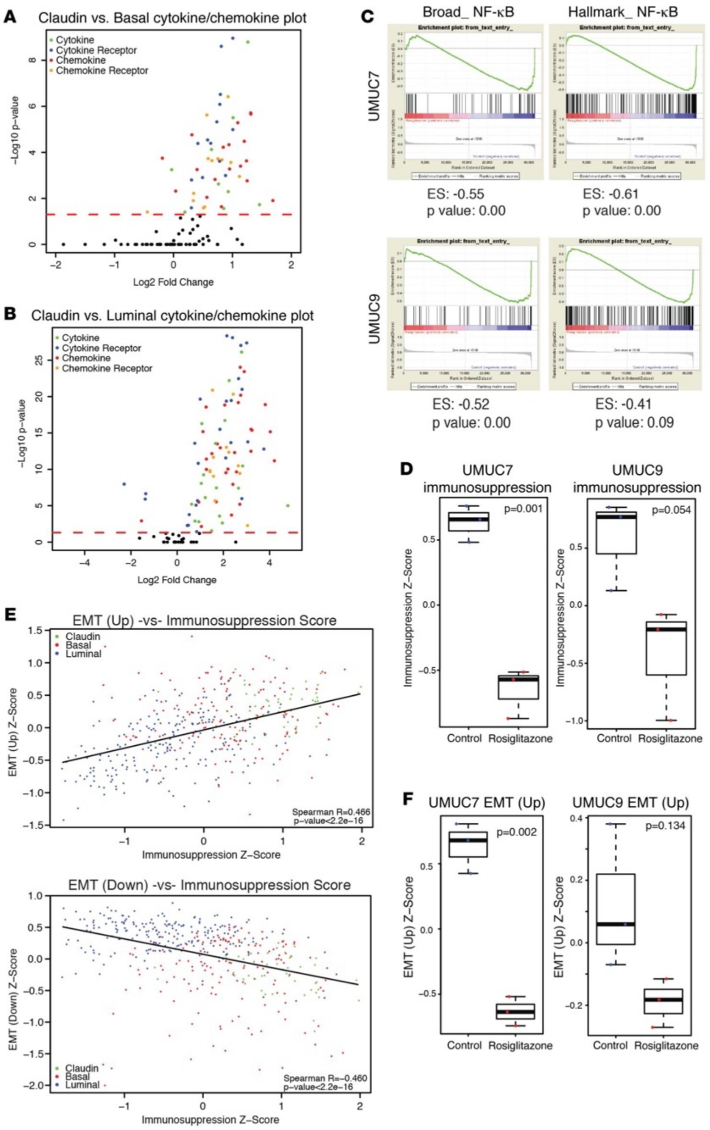


Figure 4.7 Predicted neoantigen burden by bladder cancer subtype. (A) Stacked bar plot showing the number of predicted neoantigens in each bladder tumor with a predicted IC_{50} of less than 50 nm (red bars) and less than 150 nm (yellow bars). Numbers of predicted neoantigens are shown in the left *y* axis. Blue line and right *y* axis show the number of missense mutations per tumor. *n* = 289. (B) Scatter plot of somatic missense mutations ($\log_2$) versus predicted neoantigen burden ($\log_2$) across TCGA data set. Significance and correlation were determined using Spearman's rank test. *n* = 289. (C) Box plot showing the number of predicted neoantigens with an IC_{50} of less than 50 nm by tumor molecular subtype. Subtypes were not significantly different ($P > 0.05$). Significance was determined by 1-way ANOVA. *n* = 289. The box plots denote the interquartile range (IQR), with the box representing Q1 to Q3, the line denoting Q2, and the whiskers extending an additional 1.5 times the IQR beyond Q1 and Q3. The dots represent data points. (D) Kaplan-Meier plot showing survival of bladder cancer patients with high (greater than median value, blue line) versus low (less than median value, red line) predicted numbers of neoantigens. Vertical hash marks indicate censored data. Significance was determined by log-rank test. *n* = 289. TCGA, The Cancer Genome Atlas.

A defining transcriptional program of urothelial differentiation and of luminal bladder tumors is activation of peroxisome proliferator-activated receptor γ (PPARG) signaling[121]. Consistent with this, we noted that PPARG was significantly amplified in luminal relative to claudin-low tumors (Figure 4.2B). Because PPARG is known to directly inhibit NF-κB signaling[122], we hypothesized that heightened PPARG activity might play a role in restraining the proinflammatory effects of NF-κB. Using a publicly available gene expression data set (GEO GSE48124), we noted that the expression changes induced by treatment with rosiglitazone, a PPARγ agonist, in UMUC7 and UMUC9 bladder cancer cells predicted suppression of the upstream regulator NFKB1 as well as a number of genes known to be activated by NF-κB (STAT5A, IL6, TNF, CCL5) (Supplemental Table 8). Furthermore, rosiglitazone-treated UMUC7 and UMUC9 cells had downregulation in gene signatures of NF-κB activation as assessed by gene set enrichment analysis (GSEA) (Figure 4.7C). Interestingly, we saw that rosiglitazone treatment resulted in significant downregulation of immune checkpoint molecules (such as PDL1, PDL2, IL12, and PGSL2) found in our immunosuppression signature (Figure 4.7D). In aggregate, these data support the notion that downregulation of PPARγ activity results in unopposed NF-κB signaling, which contributes to the proinflammatory milieu of claudin-low tumors as well as to their high level of active immune suppression.

Finally, in keeping with recent work demonstrating that EMT is associated with immune checkpoint molecule expression[123,124], we observed a strong correlation between our bladder cancer–derived EMT signatures and multiple immune signatures including our immunosuppression score: R = 0.462 [EMT (Up)] and R = –0.471 [EMT (Down)]; P < 2.2 × 10–16 (both "Up" and "Down") (Figure 4.7E). Furthermore, given the important role of PPARγ in terminal urothelial differentiation[122], we hypothesized that it may be a critical regulator of epithelial-mesenchymal balance in urothelial cancers. Indeed, we found that PPARγ activation (by rosiglitazone) in UMUC7 and UMUC9 cells decreased levels of our EMT (Up) signature (Figure 4.7F).

4.1.3 Discussion

Herein, we characterize the claudin-low, molecular subtype of high-grade UC. Claudin-low bladder tumors are defined by high levels of EMT, enrichment for TIC signatures, and low expression levels of tight-junction claudins. In addition, claudin-low tumors are enriched in specific genomic alterations (e.g., mutations in EP300 and NCOR1 as well as amplification in EGFR) and have a distinct transcriptional profile. Furthermore, we found that claudin-low tumors are highly enriched in all immune gene signatures examined, but also express high levels of immune checkpoint molecules. In contrast to melanoma and non–small-cell lung cancer, the predicted neoantigen burden did not appear to correlate with immune infiltration in bladder cancer. Instead, claudin-low tumors appeared to downregulate PPARγ signaling, resulting in unopposed NF-κB activity and contributing to a proinflammatory milieu (Figure 4.8).

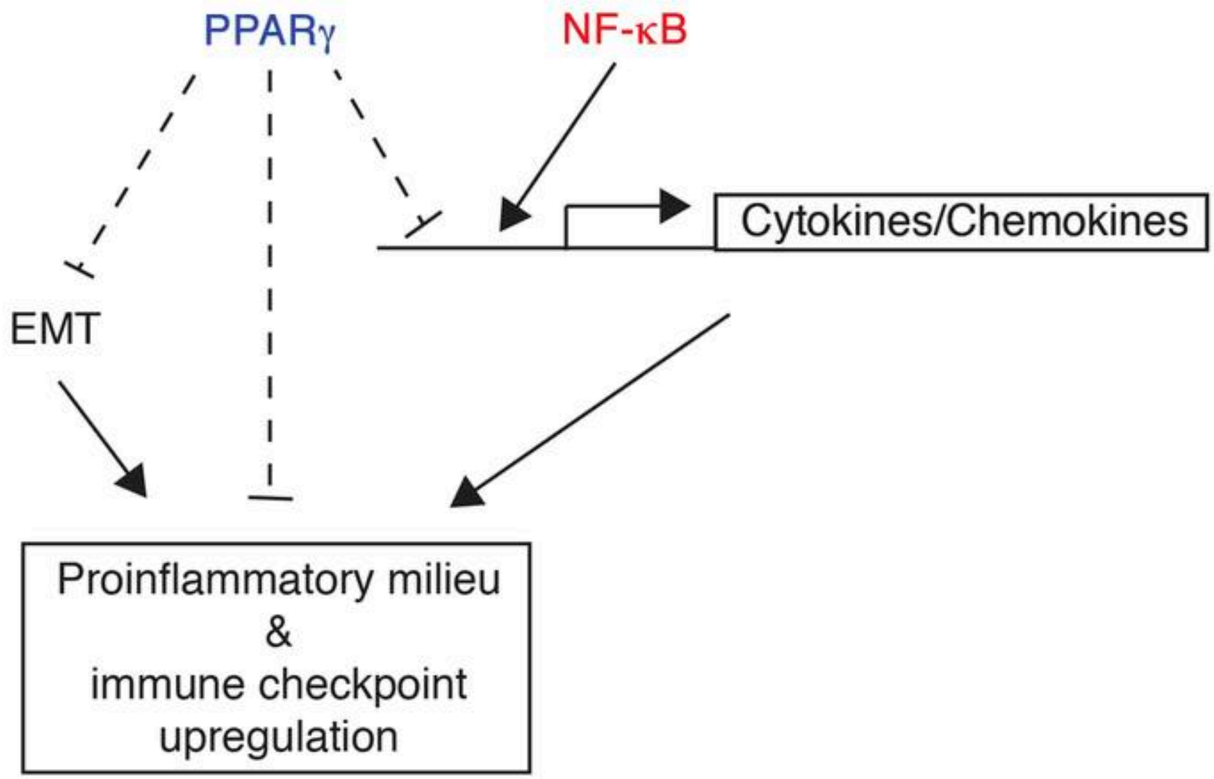


Figure 4.8 Model of immune infiltration across bladder cancer subtype**.** Proposed model of immune response regulation through PPARγ and NF-κB signaling.

In our study, as in previous studies, expression of the various immune gene signatures was highly correlated, including high correlations between gene signatures associated with specific cellular subpopulations (CD8+ T cells, B cell lineage, Th1-polarizing macrophages) and the immunosuppression gene signature. This supports the claim that tumors growing in the presence of immune cell influx must adaptively suppress the antitumor response in order to survive. Immune gene signature expression levels, the prognostic value of immune gene signatures, and TCR and BCR gene segment expression divide the bladder cancer subtypes into 3 groups: (a) low infiltrate with nonsignificant prognostic value (luminal); (b) heterogeneous infiltrate with significant prognostic value (basal); and (c) high infiltrate with nonsignificant prognostic value (claudin-low). We hypothesize that the lack of prognostic benefit in claudin-low and luminal tumors is driven by different mechanisms. Luminal tumors were sparsely infiltrated and showed low expression levels of molecules associated with immunosuppression (Supplemental Table 9). In contrast, claudin-low tumors showed a substantial but ineffective infiltrate in the context of high expression levels of immunosuppression markers. Immune features may fail to be prognostic in luminal tumors because no infiltrate is present, whereas they fail in claudin-low tumors because, despite a dense infiltrate, the level of immunosuppression overwhelms active antitumor immunity. Basal tumors have the highest degree of variability in immune gene signature expression, and in this model, some basal tumors will have generated an immune response that is competing more effectively (though ultimately insufficiently to clear tumor) with tumor-driven immune suppression. While additional studies are required to test this hypothesis, our data suggest that claudin-low tumors as a whole, as well as a subset of basal tumors, are poised for response to immune checkpoint blockade.

The different molecular aberrations that characterize the bladder cancer subtypes may yield differential exposure of antigens to the immune system, resulting in skewing of the tumor-infiltrating TCR and/or BCR repertoires in predictable ways should the antigens be public (i.e., shared between multiple patients). Though our study was not designed to formally test this, we report here a high degree of variability, in which adaptive immune gene segments were prognostic among the bladder cancer subtypes, an effect that would be expected if TCR/BCR repertoire features associated with tumor targeting were to vary by tumor subtype. Interestingly, in the basal subtype, multiple TCR gene segments associated with γδ T cells were found to be significantly prognostic ($P < 0.05$ by Cox PH). As this specific subset of T cells is involved in adaptive immunity at mucosal surfaces and able to respond to mycobacteria, γδ T cells may be involved in antitumor immunity and an attractive target for the development of biomarkers of response to bladder cancer immunotherapy, including bacille Calmette-Guérin (BCG), which is commonly given for nonmuscle invasive disease.

We report here the VDJician algorithm that performs de novo assembly of repertoires of fully rearranged BCR VDJ sequences. When analyzed, the claudin-low subtype showed the highest expression levels but the lowest repertoire diversity compared with basal and luminal subtypes. This is consistent with the presence of an antigen-driven response in the claudin-low tumors, leading to clonal expansion of antigen-reactive B cell–lineage cells. Plasma cells are known to express high levels of BCR mRNA, and these results would also be consistent with a restricted plasma cell infiltrate. In addition, as plasma cells represent a terminal differentiation in the B cell lineage in response to antigenic stimulation, their presence would also be expected in an antigen-driven response. Future experiments will be necessary to confirm these findings and attempt to map immunogenic epitopes in claudin-low tumors.

In melanoma and a subset of solid tumors, neoantigen burden correlates with expression of perforin and granzyme A (a measure of cytolytic activity)[10,125] and tumors with these attributes appear to be more responsive to CTLA4 checkpoint blockade. In bladder cancers examined in that study, there was a trend toward increased cytolytic activity, with increased predicted neoantigen burden ($P = 0.096$, data not shown)[125]. In contrast, we did not see significant correlations between neoantigen burden and predicted features such as T cell or CD8+ T cell gene signatures, immunosuppression score, or molecular subtype, suggesting that alternate etiologies exist to explain the proinflammatory state of claudin-low and basal tumors relative to that of luminal tumors. In this regard, we observed significant upregulation of cytokines and chemokines in claudin-low tumors and hypothesize that this cytokine milieu is favorable to a proinflammatory state and immune cell influx. We propose that PPARγ activity, through its ability to repress NF-κB, is inversely correlated with this proinflammatory milieu and, therefore, that luminal tumors, which are enriched in PPARG amplification and activation of PPARG gene signatures, have very little inflammation. Conversely, we found that claudin-low tumors, which have relatively low levels of PPARG pathway activation, have high levels of immune infiltration. Therefore, in contrast to the inflamed tumors found in melanoma and non–small-cell lung cancer, which appear driven by neoantigen expression, inflamed bladder cancers have a proinflammatory state induced by an enhanced cytokine/chemokine milieu. It will be important to determine whether altering the balance between PPARγ and NF-κB activity can be used to alter the immune milieu toward a more favorable response to immune therapy and whether other transcriptional programs can be harnessed as well.

Finally, while immune checkpoint inhibition holds great promise, the response rates of various solid tumors remain approximately 20% to 30%, suggesting that many patients will not derive benefit. Our BASE47 and BCL40 gene classifiers, which can accurately subtype high-grade bladder tumors, may serve to identify useful predictive biomarkers of response (i.e., claudin-low) or lack of response (i.e., luminal) to PD1 axis inhibition. Moreover, our studies further validate the notion of subtype-specific therapy in bladder cancer (i.e., basal = chemotherapy; claudin-low = immune checkpoint blockade) and advance the possibility that claudin-low breast tumors may have similar immune features.

4.1.4 Methods

TCGA data set manipulation.

TCGA Bladder Urothelial Carcinoma RNA Expression data set was downloaded from the Broad Institute Firehose Pipeline (http://gdac.broadinstitute.org) on August, 27, 2015. RNA expression was downloaded in a normalized RSEM file. Expression values were log2 transformed, and genes with less than 80% expression across all samples were filtered out. Missing values were imputed using the K-nearest neighbor imputation method. Tumor-adjacent normal samples were removed, and gene expression values were median centered across each gene. TCGA Pan-Cancer data set was downloaded from the Synapse website (https://www.synapse.org) from data set syn 2468297[111]. Genes with less than 80% expression across all samples were filtered out. Missing values were imputed using K-nearest neighbor imputation.

Gene signatures.

Bladder TIC, EMT, and tight-junction claudin gene signatures were used in the classification of a claudin-low subtype. The TIC signature was derived by Chan et al.[90]. The set of claudins used was identified by Prat et al.[92]. The EMT signature is a bidirectional signature derived on the GEO (GEO GSE60564) data set of Notch2 overexpression in a urinary bladder RT4V6 cell line. The data set was mean collapsed onto genes. Genes were filtered for a significant difference (Student's t test, $P < 0.05$) between the control and Notch2-overexpressed (EMT-induced) cell lines and also for their presence in TCGA bladder UC data set. Genes were then ranked on the basis of median difference between the 2 groups. The top 50 genes with the most increased expression in the EMT-induced cells and the top 50 genes with the most decreased expression in the EMT-induced cells were used to create the bladder cancer–specific EMT_UP and EMT_DOWN signatures, respectively. Immune gene signatures used to describe immune cell processes were derived by Iglesia et al.[91]. Z scores were calculated for each claudin, basal, and luminal subtype and box plots made of the distributions. Gene signature z scores were obtained by calculating the z score of each gene within a signature across all samples and taking the median of all gene z scores within a gene signature as the z score of the gene signature.

Identification of a claudin-low class.

Bladder basal and luminal predictions and centroid distances were made using the BASE47 PAM Classifier derived by Damrauer et al.[101]. Breast cancer claudin predictions were made using the Distance-Weighted Discrimination (DWD) Claudin Classifier provided by Prat et al.[92].

Data were clustered on the TIC/EMT (Up and Down)/claudin gene sets using average linkage clustering with a centered correlation similarity metric on the Cluster 3.0 platform. Each gene set was individually clustered across genes using average linkage clustering. Gene sets were collapsed down to z scores, and a conservative node with high TIC/high EMT UP/low EMT DOWN/low claudin gene set was selected. SigClust was run on the node, expanding out to the entire gene set for each increasing node. Differences in gene expression subtypes were determined using SAMs run on R, with an FDR of 0.05. A PAM predictor (BCL40) was derived on the 408 tumor TCGA data set for a claudin/other subtype classifier. A threshold of 6.4 was selected, giving a 40-gene predictor with an overall error rate of 0.14

A validation data set of 130 muscle-invasive UC samples was compiled from 73-sample and 57-sample data sets from GEO (GEO GSE48277][106]. Each data set was mean collapsed onto genes. The data set was combined and batch effect adjusted using parametric empirical Bayesian adjustments through the ComBat function in the sva R package and was then median centered. Genome-wide correlations and significance were calculated using a Pearson's correlation test.

Clinical, mutation, and copy number alteration analysis.

Mutation, copy number, and clinical data were downloaded as mutation packager calls through the Broad Institute Firehose Pipeline (http://gdac.broadinstitute.org) on September 3, 2015. Survival status and overall survival were determined on the basis of the data provided. Oncoprint figures were produced using the downloaded TCGA mutation and copy number alteration (CNA) data. Genes were selected on the basis of previously being identified as having significant mutations or CNAs within the gene[102]. Significance in CNA and mutation across subtypes was determined using Fisher's exact test. Cox PH ratios and CIs were derived using the survival package on the R platform.

Pathway analysis.

Cellular pathway analysis across subtypes was performed using QIAGEN's IPA (www.qiagen.com/ingenuity). Comparison across subtypes was done using the gene list with an FDR of 0.00 as determined by SAM analysis across subtypes.

Gene signature expression analysis.

Supervised clustering of samples was performed across all tumor samples by claudin, basal, and luminal subtypes. Genes within each signature were clustered using average linkage on Cluster 3.0. Significance across gene signature z scores was calculated using Student's t test. Cytokines and chemokines were identified using a RegEx search to capture all members of the molecular families (Supplemental Table 7). Volcano plots were produced using Bonferroni-adjusted Student's t test P values, and fold change was calculated using normalized RSEM expression values. NF-κB gene signatures were accessed through Molecular Signatures Database (MSigDB) or compiled by the Broad Institute. GSEA software was used to produce enrichment plots (http://www.broad.mit.edu/gsea/)[95]. UMUC7 and UMUC9 cell line data were accessed through GEO data sets GSE48124 and GSE47993, respectively. Expression values were mean collapsed onto genes. Gene signatures were compiled on the basis of existing gene lists. Significance was calculated by collapsing gene signatures into z scores as described above, and 2-tailed Student's t tests were performed across gene signatures.

TCR and BCR gene segment expression analysis. Expression levels of 353 BCR gene segments and 240 TCR gene segments were determined for TCGA bladder tumor samples with available TCGA mRNA-sequencing data and survival data using bedtools (version 2.17.0). Gene expression values were normalized to the upper quartile of total reads within a sample as previously described[126]. Survival analyses were performed using a Cox PH model to derive P values and coefficients for each gene segment using the Cox PH function in the survival package in R. The number of gene segments that were significantly associated with improved survival ($P < 0.05$ and coefficient <0) was calculated for each bladder tumor subtype. Null distributions describing the expected number of prognostic gene segments for each subtype were estimated with 95% CIs according to the bootstrap method previously published by our group[91].

Fisher's exact test was used to compare the number of BCR segments and TCR segments significantly associated with improved survival among all subtypes.

Analysis of rearranged BCR repertoires using VDJician.

The VDJician software accepts mRNA-sequencing data mapped to the genome as input and builds a deBruijn graph of read pairs that map to IgH loci or have similarity with germline IgH alleles as well as all unmapped reads. The graph is traversed exhaustively, resulting in a set of putative contigs. Anchor sequences near the 3′ end of V segments and the 5′ end of J segments are identified in an up-front indexing step. If a contig contains a sequence within a configurable distance of a V anchor and a J anchor, the anchors are a reasonable distance apart, and conserved amino acids that typically bind a CDR3 segment are present (cysteine and tryptophan for IgH), the contig is considered a candidate. The original set of reads is mapped to candidate contigs, which are then further filtered on the basis of coverage and read pair information. VDJician outputs a final set of contigs along with alignments of the original reads mapped to these contigs. This output was passed to RSEM for transcript quantification. The total BCR count was calculated by summing the read count values for all predicted BCR sequences for each sample. Evenness was calculated by dividing the Shannon-Wiener diversity index by the number of BCR sequences for each sample (example expression in R): -sum((read count/sum(read count)) * log(read count/sum(read count))))/log(number of BCR sequences). P values were determined using a Mann-Whitney U–Wilcoxon test.

Neoantigen prediction.

The bladder cancer data set used for neoantigen prediction consisted of 289 samples with available TCGA mRNA-sequencing data, exome-sequencing data, and tumor-specific mutation annotation data[102]. Neoantigens were predicted using a bioinformatics pipeline similar to that developed by Rajasagi et al.[28]. Tumor-specific single nucleotide variant annotation data were downloaded from the Broad Institute Firehose Pipeline (http://gdac.broadinstitute.org). Pysam was used to determine RNA-sequencing read coverage of missense mutations, and bedtools (version 2.17.0) was used to determine the exome-sequencing read coverage of missense mutations. Nine- and ten-mer peptides derived from 3 ORFs with all possible combinations of missense mutations that overlap the genomic location of peptide in the ENCODE reference transcript set were considered in the peptide generation pipeline. DNA sequences corresponding to peptides were retrieved and translated in silico into protein sequences. The expression levels of each peptide generated were determined by the lowest missense mutation RNA-sequencing read coverage. PHLAT was used to identify the HLA class I (HLA-A, HLA-B, HLA-C) type of each tumor sample[127]. Binding affinity to MHC molecules expressed by the tumor for all possible 9- and 10-mer peptides generated from missense mutations was predicted using NetMHCpan (version 2.8). Binding affinity of peptides to null alleles, alternatively expressed alleles, and alleles not supported by NetMHCpan were not predicted. Peptides were then filtered by their binding affinities (IC50 nM) to each class I allele in the tumor sample's HLA type and RNA expression level of the predicted source transcript(s). Peptides with an IC50 value of less than 150 nM for at least 1 class I allele and RNA read support of at least 2 reads were considered predicted neoantigens.

Statistics.

A P value of less than 0.05 was considered significant across all analyses performed. SigClust statistical analysis software was used to determine significance in Figure 4.1A and Supplemental Figure 1A. A Fisher's exact test was used in Figure 4.1B; Figure 4.2, A and B; and Supplemental Table 2. A Pearson's correlation was performed in Figure 4.1B. A log-rank test of survival difference was performed in Figure 4.1C (Bonferroni-corrected) and Figure 4.6B. A

Bonferroni-corrected 2-tailed Student's t test was performed in Figure 4.3A; Figure 4.7, A, B, D, and F; Supplemental Figure 1, B–E; and Supplemental Figure 4. Cox PH modeling was performed in Figure 4.4, A and B, and Figure 4.5, A–D. A Mann-Whitney U–Wilcoxon test with an FDR multiple testing correction was performed in Figure 4.5, E and F. A Spearman's rank correlation was used in Figure 4.6B, Figure 4.7E, and Supplemental Figure 5C. A 1-way ANOVA was used in Figure 4.6C. GSEA significance testing was used in Figure 4.7C. SAM significance testing was performed in Supplemental Figure 3A and Supplemental Tables 4 and 5. IPA significance testing was used in Supplemental Figure 3B and Supplemental Tables 6 and 8.

Study approval.

No experiments included in the manuscript used animal or human subjects and, as such, did not require IRB approval.

Supplemental material

All supplemental figures and tables cited in Chapter 4.1 are listed according to the original published manuscript, which can be found at https://insight.jci.org/articles/view/85902.

4.2 Immunogenomic characterization of Triple Negative Breast Cancer Brain Metastases

4.2.1 Introduction

Triple negative breast cancer (TNBC) is an aggressive subtype of breast cancer with high metastatic potential. African-American women are more likely to present with TNBC than Caucasian women. Once metastatic, half of patients with TNBC will develop brain metastases (BM)[128]. Regardless of treatment strategy, African-American women have shorter survival times, highlighting disparities in TNBC incidence and clinical outcomes[129,130]. While TNBC BM are routinely treated with radiotherapy, early responses are not durable and expected survival remains less than one year[131]. There are no approved systemic therapies to treat TNBC BM. Against this background, monoclonal antibodies that boost adaptive immune responses have yielded durable responses in incurable solid tumors, including metastatic TNBC[132,133]. To enhance this activity, novel strategies combine co-stimulatory agonists with co-inhibition repressors, including targeting 4-1BB to increase cytolytic T-lymphocyte activity[134]. To target immune suppressive phenotypes of tumor-associated macrophages, small molecule inhibitors against CSF1R are in early phase clinical trials. Some CSF1R inhibitors, such as BLZ945, cross the blood brain barrier (BBB) and are active against preclinical brain tumor models[135]. Radiotherapy not only transiently disrupts the BBB, but also synergizes with immunotherapy in murine models of several cancers, including breast cancer[136].

Despite the early successes of immunotherapy and synergy with radiotherapy, patients with brain metastases have largely been excluded from immunotherapy trials. Reasons for exclusion include several assumptions: (1) monoclonal antibodies cannot cross the BBB, (2) immune responses in the immune-privileged CNS may be prohibitively toxic, and (3) evidence of productive immune responses against TNBC BM is lacking.

Both the challenge of the blood brain barrier and the paucity of data on the biologic underpinnings and immune response of BCBM contribute to inadequate therapies for this disease. We sought to characterize the genomic and immune landscape of TNBC BM to foster the development of effective brain permeable anti-cancer agents, including immunotherapy. Our result challenge these assumptions by showing: (1) immune infiltration of TNBC BM is correlated with improved patient survival, (2) immune signatures including a T cell signature correlate positively with improved survival in brain metastases, and (3) gene signatures associated with response to immunotherapy are upregulated in TNBC BM. Collectively, our data support exploration of immunotherapy in patients with TNBC BM.

4.2.2 Results

Mutational Analysis of TNBC brain metastasis and primary breast tumors.

We evaluated the specific mutations present in TNBC brain metastases and primary tumors that metastasize to the brain via whole exome sequencing (Figure 4.9). Among the brain metastases and primary tumors, *TP53* was found to be one of the most commonly mutated genes, a hallmark of TNBC. Notably, many other genes associated with DNA damage repair processes were also mutated in brain metastases, including *ATM* and *ATR*. This is of interest as currently ATM and ATR inhibitors are in clinical development and several inhibitors of these pathways are brain permeable. In addition, several genes associated with PI3K signaling were mutated in brain metastases tissues, including *PIK3R1*; in primary tumors, *PIK3R1 and PIK3CA*. This finding is also clinically-relevant as inhibitors of the PI3K and AKT pathway are in development and brain permeable. Finally, *FAT1,* which functions as an adhesion molecule and/or signaling receptor likely important in developmental processes and cell communication, and *GNAS* were altered in both brain and primary tumor tissues. We believe that further exploration inhibiting these pathways in preclinical models of TNBC would be prudent as a next step in investigation.

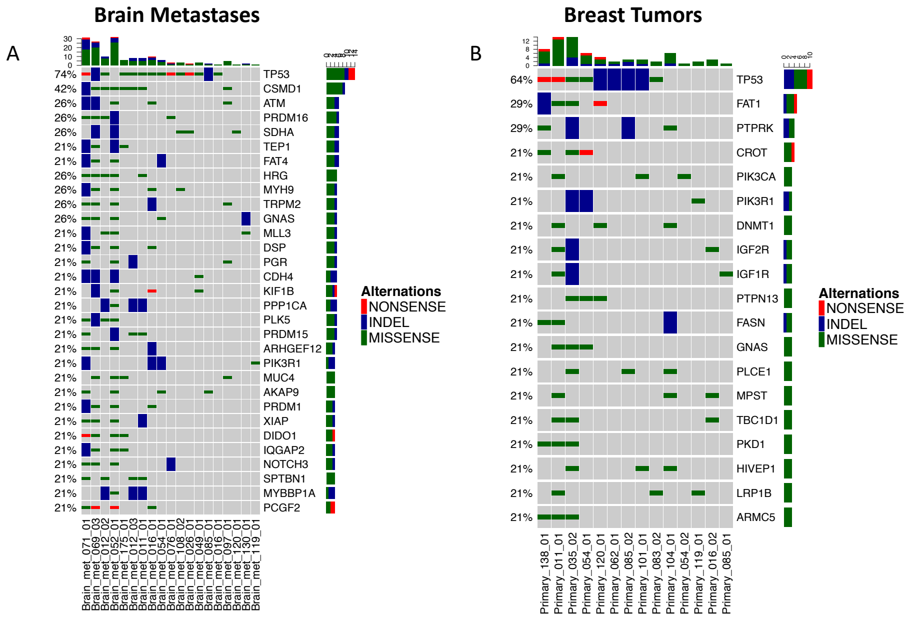


Figure 4.9 Genomic characterization of primary breast and brain metastasis tumors. Brain metastases (A) and primary tumor (B) samples with matched normal tissues were analyzed for single nucleotide mutations, small insertions/deletions, and copy number variations. Oncoprints of genes that were identified to be mutated in at least 21% of brain metastases and primary breast tumors, respectively. The copy number variation landscape analysis was also performed, but no significant differences were found between primary breast tumors and brain metastases.

Lower immune gene signature expression in brain metastasis compared with primary breast tumors.

To characterize the immune microenvironment of primary breast tumor and breast cancer brain metastasis, the expression levels for 56 immune signatures were calculated (Figure 4.10). Globally, immune gene signatures were lower in brain metastases (*P < 0.05, **P < 0.01, and *** P < 0.001). Notably, gene signatures involved in B cell signaling, dendritic cells, mast cells, T central memory cell, T effector memory cells and regulatory T cells were statistically lower in brain metastases tissues compared to primary TNBC that eventually metastasized to the brain. On the contrary, an IPRES-derived responder signature was higher in brain metastases tissues, while the non-responder signature was higher in primary TNBC’s. This observation supports our hypothesis that breast cancer brain metastases will respond to immune checkpoint blockade, perhaps more robustly than response of primary tumors, supporting continued evaluation of immunotherapy in TNBC brain metastases in the preclinical and clinical settings.

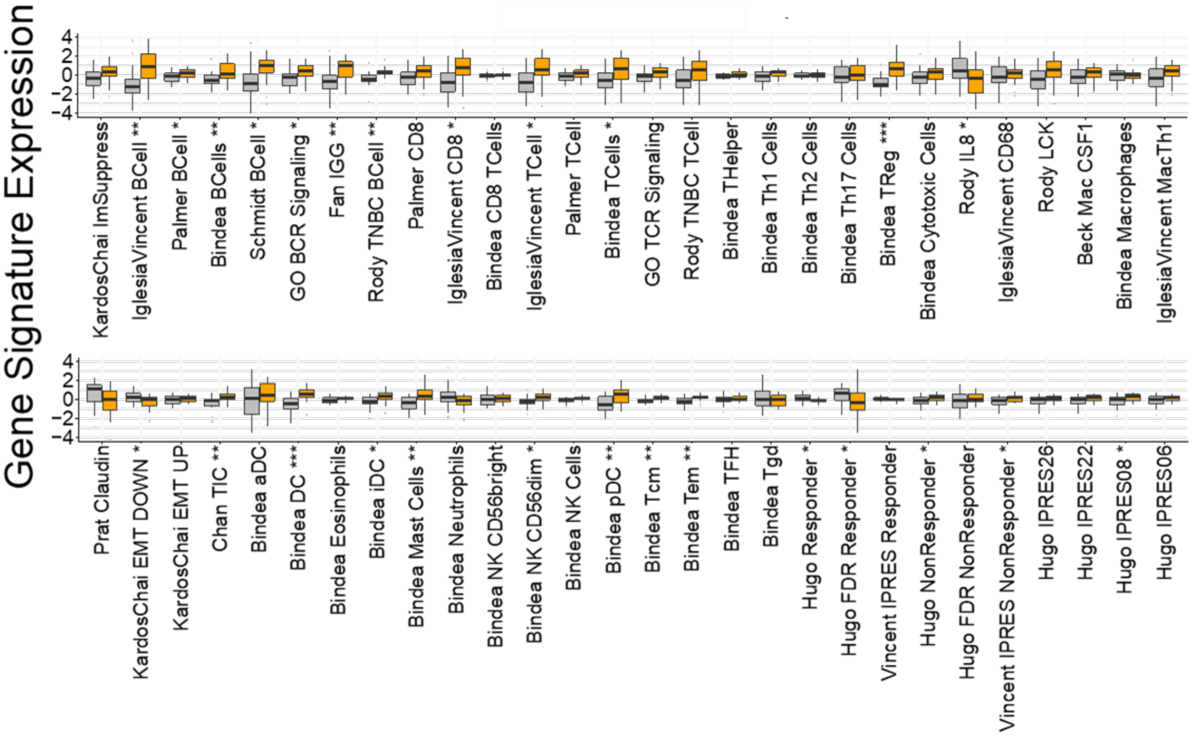


Figure 4.10 Immune gene signatures are differentially expressed between brain metastasis and primary breast tumors. 56 immune gene signatures were used for this analysis. Boxplots show most significantly different signatures. The mean expression levels of all genes in the signature defines the signature score for a given tumor. Significance was determined by 1-way ANOVA (*P < 0.05, **P < 0.01, and *** P < *0.001*). Many signatures (IGG Cluster, B cell, effector memory T cell, dendritic cells) were lower in brain metastases compared to primary tumors, indicating the brain metastases are less immune infiltrated. A signature of responsiveness to PD-1 inhibition in melanoma was higher in brain metastases compared with primary tumors. DC, dendritic cell; aDC, activated dendritic cell; iDC, intestinal dendritic cell; pDC, plasmacytoid dendritic cell; Tem, effector memory T cell; Tcm, Central memory T cells; Tgd, Gamma delta T cells; TReg, Regulatory T cell; TIC, tumor initiating cell; IPRES derived responder and non-responder to PD1 inhibition; Th1, T helper type 1 cells; Th2, T helper type 2 cells; Th17, T helper type 17 cells; TFH, T follicular helper cells; EMT, epithelial-to-mesenchymal transition.

T cell receptor is less diverse with higher reads counts in primary tumors compared to brain metastases.

An antigen-driven T cell response would be expected to lead to clonal expansion of T cells, causing decreased diversity of TCR repertoires. To evaluate this concept in our samples, we used MiXCR to infer and quantify expression levels of rearranged TCR V(D)J sequences from RNA-seq data. The result (Figure 4.11) clearly shows a decreased diversity of T cell receptor with increased read count in in primary tumors compared to brain metastases (*P < 0.05 and **P < 0.01), consistent with the presence of less antigen-driven T cell responses in brain metastases than in primary tumors.

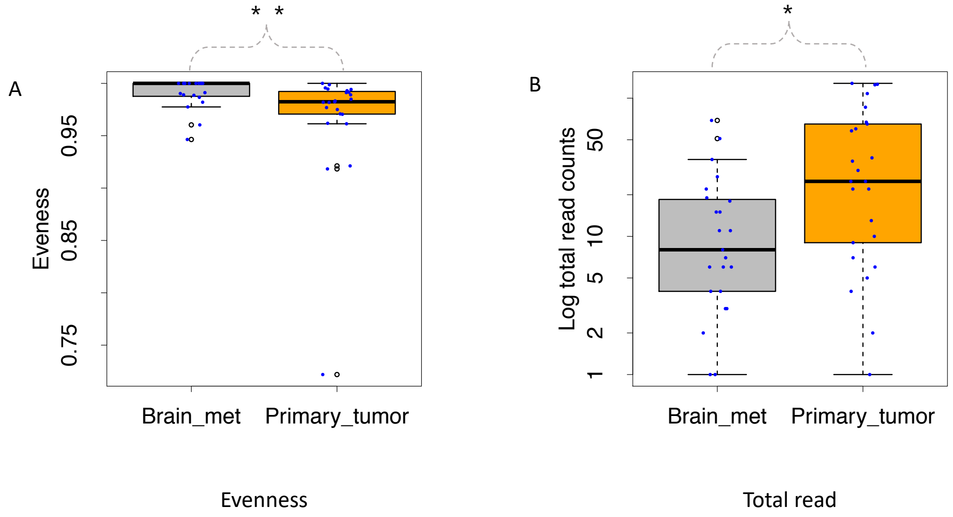


Figure 4.11 TCR segment expression is significantly different between tumors and brain metastasis. Number of reads supporting rearranged TCR V(D)J sequence repertoires and repertoires diversity were analyzed by tissue. In tumors T-cell repertoires were less diverse with higher counts compared to brain metastases. The difference was statistically significant using Wilcoxon rank-sum test for both diversity and counts in T-cells, shown in A and B.

Differential gene and pathway expression between primary breast and brain metastasis tumors

Next, we evaluated differential gene expression between TNBC brain metastases tissues and primary TNBC that eventually metastasize to the brain (Figure 4.12A). From this analysis, we found that many genes involved in neuronal processes were higher in the brain metastases compared to primaries (i.e. GFAP – astrocyte marker, MOBP – involved in myelin sheath, GAP43 – involved in neuronal plasticity). We also used single sample GSEA (ssGSEA) to evaluate differentially expressed signaling pathways between TNBC brain metastases and primary tumors (Figure 4.12B). Our results also illustrated many neurologic signaling pathways were differentially-expressed in brain metastases tumors; oncogenic signatures such as BRCA1, AKT, mTOR, and KRAS were differentially expressed in primary tumors while RB, cyclic AMP, mTOR and EIF4E were differentially expressed in brain metastases.

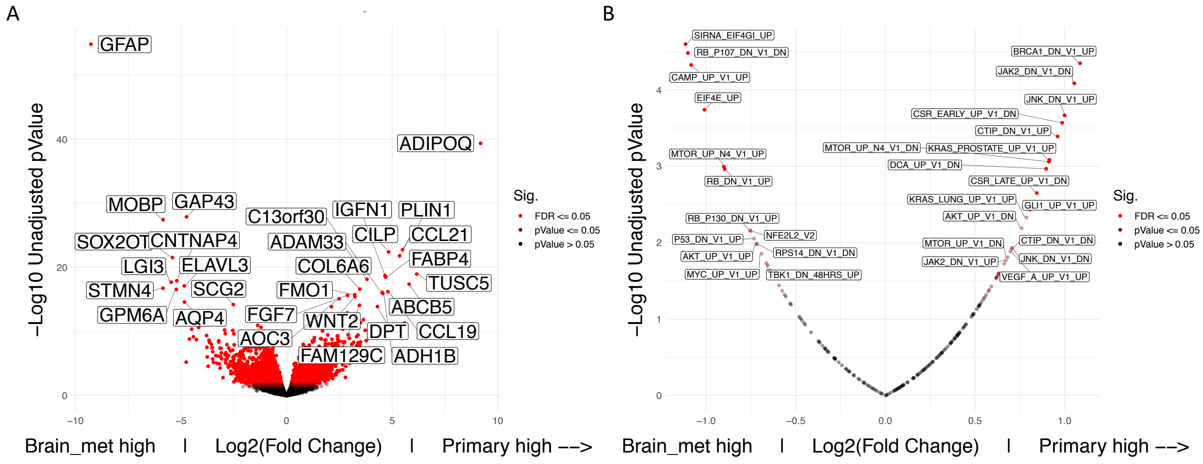


Figure 4.12 Differentially expressed genes and pathways between primary breast and brain metastasis tumors. (A) Volcano plot of log2 fold change of gene expression levels with -log10 unadjusted P value of gene expression shown. The 20 genes with lowest P values were labeled. Many genes involved in neuronal processes were high in brain metastases as compared to primary breast tumor, indicating possible normal brain tissue admixed within brain metastases. (B). Volcano plots of log2 fold change of ssGSEA canonical pathway and oncogenic signature expression values with –log10 unadjusted P value of gene expression were analyzed. The oncogenic signature, with 20 lowest P values were labeled. Oncogenic signatures such as BRCA1, AKT, CtIP, and KRAS were differentially expressed in primary tumors while MTOR, EIF4E, and RB family members were differentially expressed in brain metastases.

Immune signatures correlate with improved survival in primary tumors and to brain metastases.

We examined the prognostic implications of immune gene signatures in TNBC brain metastases and primary tumors using Cox proportional hazards regression with immunogenomics features as predictor variables and overall survival as the response variable. A forest plot of Cox PH ratios of a subset is shown, with a 95% CI indicated around the values (Figure 4.13). Immune signatures such as macrophage, Th1 cell, IPRES responder, and low EMT correlated positively with survival in primary tumors (Figure 4.13A) and immune signatures such as central memory T cell and IPRES responder correlated positively with survival when expressed in brain metastases tissues (Figure 4.13B). The positive correlation of IPRES responder signature with survival suggest brain metastases might respond well to immunotherapies such as anti-PD-1 therapy.

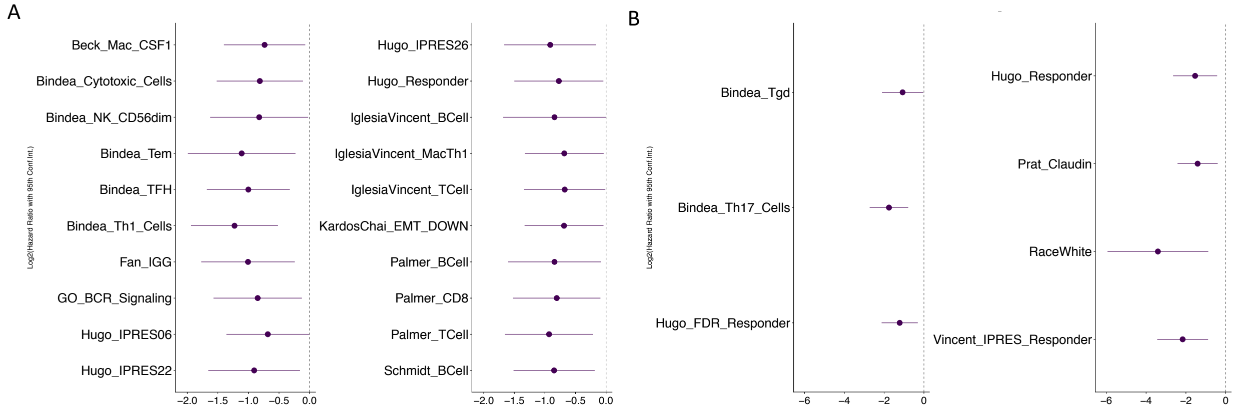


Figure 4.13 Prognostic immune gene signatures in primary breast tumors and brain metastasis. Cox PH ratios were determined for immune gene signatures correlated with survival in primary breast tumors (A.) and brain metastases (B.). Shown are Forest plots of Cox PH ratios of a subset that are significantly correlated, with a 95% CI indicated around the values.

Neoantigen burden in primary tumors and brain metastases.

Neoantigens are peptides derived from tumor specific mutations and are presented by MHC molecules[4]. They can drive robust antitumor T cell responses[4]. Neoantigens are predicted using whole-exome sequencing of paired tumor and matched normal samples, and expression is confirmed by using RNA expression data. The predicted neoantigen number has been positively associated with favorable clinical outcomes in many tumor types[118]. Our neoantigen prediction results suggest that the neoantigen burden is higher in primary tumors compared with brain metastases samples (Figure 4.14). The low level of immune infiltration in brain metastases samples might explain the lower level of antigen-driven T cell response in brain metastases.

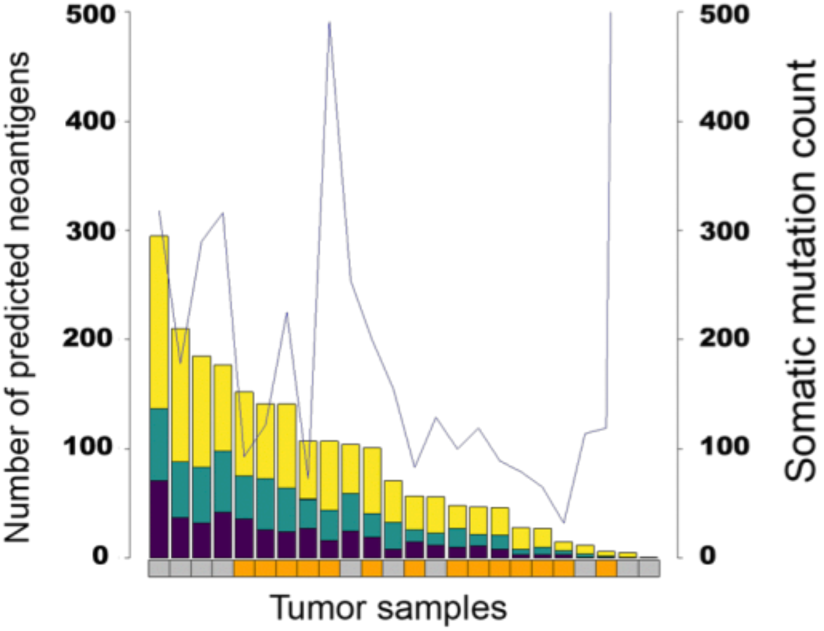


Figure 4.14 Predicted neoantigen burden in breast cancer brain metastases. Stacked bar plot showing the number of predicted neoantigens in each tumor sample with a predicted IC50 of less than 50 nm (dark purple bars), less than 150 nm (dark green bars), and l3e0ss than 500nm (yellow bars). Blue line and right y axis show the number of somatic mutations per tumor.

4.2.3 Discussion

Our results illustrate several key findings and enrich our understanding of the genetic and immunologic landscape of TNBC brain metastases. First, we illustrate a lower immune gene signature globally in TNBC brain metastases when compared to primaries that eventually metastasize to the brain. We also learned that (1) the capacity to respond to immunotherapy (as illustrated by higher expression of the IPRES-derived responder signature in brain metastases tissues), coupled with (2) a higher mutational and neoantigen burden (3) improved prognosis among those whose tissues illustrate higher expression of the IPRES- derived responder signature all lead to the conclusion that immunotherapy to treat TNBC brain metastases is a promising therapeutic strategy worthy of continued investigation. We will continue to investigate immunotherapy in preclinical TNBC intracranial models, and ultimately, in patients with progressive TNBC brain metastases. While we found several neuronal genes and pathways with increased expression in TNBC brain metastases, we are unable to ascertain from bulk tissue RNA sequencing data the contribution of non-malignant neuronal tissue to these results. As such we do not interpret them as necessarily indicative of TNBC biology. In addition, our WES data points to mutations in several DNA damage repair pathways in TNBC brain metastases, thus combination immunotherapy with PARP inhibitors is of interest, and has shown respectable activity in extracranial metastatic TNBC[137]. Thus, we have developed and have approval for a phase II study of the PARP inhibitor, niraparib, in combination with the PD1 antibody (TSR-042) to treat patients with stable or progressive TNBC brain metastases. This analysis is the largest genomic and immune analysis of TNBC brain metastases to date. Our results provide valuable insight into the molecular underpinnings of this aggressive disease and will certainly lead to additional preclinical work and clinical trials for patients with TNBC brain metastases.

4.2.4 Methods

Data processing.

RNA-seq pipeline: The mRNA sequencing data for 49 tumor samples were aligned to hg19 reference genome using MapSplice 2.0.1.9[62]. Gene expression was quantified using RSEM 1.1.13[138]. Gene expression values were upper quartile normalized, $\log_2$ transformed. Missing values were converted to zero.

Whole exome pipeline.

The Whole Exome Sequencing Data for 33 tumor and matched normal samples were aligned to hg19 reference genome using BWA 0.7.9a[139]. The generated bam files were realigned with ABRA 0.96[140]. Somatic variant calling results generated from Strelka 1, UNCeqR 0.1.14, and Cadabra 0.96 were merged together and annotated with SnpSift 1.3.4 and snpEff 3.3[63,85,141–143]. Annotation information was retrieved from COSMIC 20150210, ExAC 0.3, and dbSNP 132[144–146].

Genomic mutation analysis.

Pysam 0.83[53] was used to parse mutation information from VCF files generated by the whole exome pipeline. Circos plot was generated using R package circlize[147]. Oncoprint figures were generated using mutation information and R package ComplexHeatmap[148].

Immune signature expression analysis.

The expression values of 56 immune signatures were calculated by averaging the expression values of genes in the signature. Genes with less than 70% expression across all samples were filtered out. Significance was determined by 1-way ANOVA.

Analysis of rearranged TCR and BCR repertoires.

MiXCR 2.1.2 was used for inference of TCR repertoires[98]. And V'DJer 0.12 was used for estimation of BCR repertoire[149]. V'DJer 0.12 outputs a set of assembled contigs along with a bam file that contains alignment information of the original reads to these contigs. The output files were passed to salmon 0.13.1 for transcript quantification[150]. Evenness was calculated by dividing the Shannon-Wiener diversity index by the number of TCR or BCR sequences for each sample. *P* values were determined using a Wilcoxon Rank Sum test.

Neoantigen prediction.

The data set used for neoantigen prediction consisted of 26 samples with available paired tumor and normal exome-sequencing data along with tumor mRNA-sequencing data. Neoantigens derived from DNA variants were predicted using a bioinformatics pipeline similar to that developed by Rajasagi et al[28]. Pysam 0.83 was used to determine RNA-sequencing read coverage and exome-sequencing read coverage of missense mutations[53]. Eight-, Nine-, ten-, and eleven-mer peptides derived from 3 ORFs with all possible combinations of missense mutations that overlap the genomic location of peptide in the GENCODE reference transcript set were considered in the peptide generation pipeline[52]. DNA sequences corresponding to peptides were retrieved and translated in silico into protein sequences. The expression levels of each peptide generated were determined by the lowest missense mutation RNA-sequencing read coverage. PHLAT was used to identify the HLA class I (HLA-A, HLA-B, HLA-C) type of each tumor sample[127]. Binding affinity to MHC molecules expressed by the tumor for all possible 9- and 10-mer peptides generated from missense mutations was predicted using NetMHCpan (version 4.0)[20]. Binding affinity of peptides to null alleles, alternatively expressed alleles, and alleles not supported by NetMHCpan were not predicted. Peptides were then filtered by their binding affinities (IC50 nM) to each class I allele in the tumor sample's HLA type and RNA expression level of the predicted source transcript(s). Peptides with an IC50 value of less than 150 nM for at least 1 class I allele and RNA read support of at least 2 reads were considered predicted neoantigens.

Differential gene expression and pathway analysis.

The differential gene expression analysis was performed using DESeq2[94]. The differential pathway analysis was performed using Single Sample GSEA[95]. Comparison between brest cancer and brain metastases was done using MSigDB oncogenic gene signatures[151].

Survival analyses.

Survival analyses was performed using Cox Proportional-Hazards model for brain and primary breast tumors and brain metastasis. Hazard ratios were derived from the Cox proportional hazards model for 56 immune signatures, gender, and race. Features with a *P* value less than 0.05 were reported.

CHAPTER 5 CONCLUDING REMARKS

In this dissertation, we presented two algorithms that were developed to facilitate the identification of neoantigens as well as bioinformatics analyses of bladder tumors and breast cancer brain metastases. While different studies were made for each chapter and different bioinformatics algorithms and tools were involved, the central goal remains the same: to identify tumor specific neoantigens using computational analysis of genomics data. For splice variant neoantigens, we developed NeoSplice for specific and comprehensive prediction of splice variant neoantigens using tumor and matched normal RNA-seq data. To access the immunogenicity of tumor specific antigen, we developed a gradient boosting model for predicting immunogenicity using peptide-intrinsic features and demonstrated that out of frame neoepitopes could provide antitumor immunity. We also performed comprehensive genomic and immune characterizations to gain novel insight about immunogenomic features of bladder tumors and triple-negative breast cancer brain metastases.

In general, we have made key progress in the accurate identification of neoantigens that are potentially therapeutic targets. Accurate identification of splice variant neoantigens will expand therapeutic target space for tumors, especially for those tumors like AML where neoantigens derived from SNVs and Indels are few. Understanding of peptide-intrinsic features of predicted tumor antigens that could discriminate epitopes with therapeutic will improve the accuracy of identification of therapeutic effective neoantigens. We expect the work described in this dissertation will become more important as additional neoantigen specific therapeutic platforms enter clinical trials.

ENDNOTES

Author lists (* contributed equally to this work):

[1] Chapter 2: Shengjie Chai, Jan F. Prins, Lisle Mose, Jonathan S. Serody, Paul M. Armistead and Benjamin G. Vincent

[2] Chapter 3: Christof C. Smith*, Shengjie Chai*, Amber R. Washington, Samuel J. Lee, Elisa Landoni, Kevin Field, Jason Garness, Lisa M. Bixby, Sara R. Selitsky, Joel S. Parker, Barbara Savoldo, Jonathan S. Serody and Benjamin G. Vincent

[3] Chapter 4.1: Jordan Kardos, Shengjie Chai, Lisle E. Mose, Sara R. Selitsky, Bhavani Krishnan, Ryoichi Saito, Michael D. Iglesia, Matthew I. Milowsky, Joel S. Parker, William Y. Kim, Benjamin G. Vincent

[4] Chapter 4.2: Benjamin G. Vincent, Maria Sambade, Shengjie Chai, Marni B. Siegel, Luz Cuaboy, Alan Hoyle, Joel Parker, Charles M. Perou and Carey K. Anders

Personal contributions to all sections

[5] Chapter 2: Development and evaluation of algorithm, writing for all sections of the manuscript.

[6] Chapter 3: Algorithmic design, neoantigen prediction of all mouse models.

[7] Chapter 4.1: Neoantigen prediction and immunogenomic analysis represented in Figure 4.5 A-D and Figure 4.6

[8] Chapter 4.2: Immunogenomic analysis represented in all figures.

REFERENCES

1. Alam, R. A brief review of the immune system. *Prim. Care - Clin. Off. Pract.* **25**, 727–738 (1998).
2. Li, A. *et al.* Utilization of Ig heavy chain variable, diversity, and joining gene segments in children with B-lineage acute lymphoblastic leukemia: Implications for the mechanisms of VDJ recombination and for pathogenesis. *Blood* **103**, 4602–4609 (2004).
3. Matsuda, F. *et al.* The complete nucleotide sequence of the human immunoglobulin heavy chain variable region locus. *J. Exp. Med.* **188**, 2151–2162 (1998).
4. Schumacher, T. N. & Schreiber, R. D. Neoantigens in cancer immunotherapy. *Science* **348**, 69–74 (2015).
5. Yarchoan, M., Johnson, B. A., Lutz, E. R., Laheru, D. A. & Jaffee, E. M. Targeting neoantigens to augment antitumour immunity. *Nature Reviews Cancer* **17**, 209–222 (2017).
6. Kim, S. *et al.* Neopepsee: Accurate genome-level prediction of neoantigens by harnessing sequence and amino acid immunogenicity information. *Ann. Oncol.* **29**, 1030–1036 (2018).
7. Sahin, U. *et al.* Personalized RNA mutanome vaccines mobilize poly-specific therapeutic immunity against cancer. *Nature* **547**, 222–226 (2017).
8. Ott, P. A. *et al.* An immunogenic personal neoantigen vaccine for patients with melanoma. *Nature* **547**, 217–221 (2017).
9. Hundal, J. *et al.* pVAC-Seq: A genome-guided in silico approach to identifying tumor neoantigens. *Genome Med.* **8**, (2016).
10. Van Allen, E. M. *et al.* Genomic correlates of response to CTLA-4 blockade in metastatic melanoma. *Science (80-.).* **350**, 207–211 (2015).
11. Turajlic, S. *et al.* Insertion-and-deletion-derived tumour-specific neoantigens and the immunogenic phenotype: a pan-cancer analysis. *Lancet Oncol.* **18**, 1009–1021 (2017).
12. Kardos J, Chai S, Mose LE, Selitsky SR, Krishnan B, Saito R, et al. Claudin-Low Bladder Tumors are Immune Infiltrated and Actively Immune Suppressed. *JCI Insight* **1**, (2016).
13. Zhou, Z. *et al.* TSNAD: An integrated software for cancer somatic mutation and tumour-specific neoantigen detection. *R. Soc. Open Sci.* **4**, (2017).
14. Bjerregaard, A. M., Nielsen, M., Hadrup, S. R., Szallasi, Z. & Eklund, A. C. MuPeXI: prediction of neo-epitopes from tumor sequencing data. *Cancer Immunol. Immunother.* **66**, 1123–1130 (2017).
15. Hundal, J. *et al.* pVACtools: a computational toolkit to identify and visualize cancer neoantigens. *bioRxiv* 501817 (2019). doi:10.1101/501817
16. Richman, L. P., Vonderheide, R. H. & Rech, A. J. Neoantigen Dissimilarity to the Self-Proteome Predicts Immunogenicity and Response to Immune Checkpoint Blockade. *Cell Syst.* **0**, (2019).
17. Nielsen, M. & Andreatta, M. NetMHCpan-3.0; improved prediction of binding to MHC

class I molecules integrating information from multiple receptor and peptide length datasets. *Genome Med.* **8**, (2016).

18. Hoof, I. *et al.* NetMHCpan, a method for MHC class i binding prediction beyond humans. *Immunogenetics* **61**, 1–13 (2009).

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

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

21. Han, Y. & Kim, D. Deep convolutional neural networks for pan-specific peptide-MHC class I binding prediction. *BMC Bioinformatics* **18**, (2017).

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

23. Andreatta, M. *et al.* Accurate pan-specific prediction of peptide-MHC class II binding affinity with improved binding core identification. *Immunogenetics* **67**, 641–650 (2015).

24. Kim, Y. *et al.* Immune epitope database analysis resource. *Nucleic Acids Res.* **40**, (2012).

25. Zhang, Q. *et al.* Immune epitope database analysis resource (IEDB-AR). *Nucleic Acids Res.* **36**, (2008).

26. Sonnhammer, E. L., von Heijne, G. & Krogh, A. A hidden Markov model for predicting transmembrane helices in protein sequences. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **6**, 175–182 (1998).

27. Ley, T. J. *et al.* Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. *N. Engl. J. Med.* **368**, 2059–2074 (2013).

28. Rajasagi, M. *et al.* Systematic identification of personal tumor-specific neoantigens in chronic lymphocytic leukemia. *Blood* **124**, 453–462 (2014).

29. De Necochea-Campion, R., Shouse, G. P., Zhou, Q., Mirshahidi, S. & Chen, C. S. Aberrant splicing and drug resistance in AML. *Journal of Hematology and Oncology* **9**, (2016).

30. Lee, S. C. W. *et al.* Modulation of splicing catalysis for therapeutic targeting of leukemia with mutations in genes encoding spliceosomal proteins. *Nat. Med.* **22**, 672–678 (2016).

31. Zhou, J. & Chng, W. J. Aberrant RNA splicing and mutations in spliceosome complex in acute myeloid leukemia. *Stem Cell Investigation* **2017**, (2017).

32. Kahles, A. *et al.* Comprehensive Analysis of Alternative Splicing Across Tumors from 8,705 Patients. *Cancer Cell* **34**, 211-224.e6 (2018).

33. Jayasinghe, R. G. *et al.* Systematic Analysis of Splice-Site-Creating Mutations in Cancer. *Cell Rep.* **23**, 270-281.e3 (2018).

34. Kranz, L. M. *et al.* Systemic RNA delivery to dendritic cells exploits antiviral defence for cancer immunotherapy. *Nature* **534**, 396–401 (2016).

35. Kreiter, S. *et al.* Mutant MHC class II epitopes drive therapeutic immune responses to

cancer. *Nature* **520**, 692–696 (2015).

36. Castle, J. C. *et al.* Exploiting the mutanome for tumor vaccination. *Cancer Res.* **72**, 1081–1091 (2012).

37. Gubin, M. M., Artyomov, M. N., Mardis, E. R. & Schreiber, R. D. Tumor neoantigens: Building a framework for personalized cancer immunotherapy. *Journal of Clinical Investigation* **125**, 3413–3421 (2015).

38. Patronov, A. & Doytchinova, I. T-cell epitope vaccine design by immunoinformatics. *Open Biol.* **3**, (2013).

39. Dintzis, H. M., Dintzis, R. Z. & Vogelstein, B. Molecular determinants of immunogenicity: The immunon model of immune response. *Proc. Natl. Acad. Sci. U. S. A.* **73**, 3671–3675 (1976).

40. Cole, D. K. *et al.* Modification of MHC Anchor Residues Generates Heteroclitic Peptides That Alter TCR Binding and T Cell Recognition. *J. Immunol.* **185**, 2600–2610 (2010).

41. Singh, N. K. *et al.* Emerging Concepts in TCR Specificity: Rationalizing and (Maybe) Predicting Outcomes. *J. Immunol.* **199**, 2203–2213 (2017).

42. Chowell, D. *et al.* TCR contact residue hydrophobicity is a hallmark of immunogenic CD8$^+$ T cell epitopes. *Proc. Natl. Acad. Sci. U. S. A.* **112**, E1754–E1762 (2015).

43. Huang, L., Kuhls, M. C. & Eisenlohr, L. C. Hydrophobicity as a driver of MHC class i antigen processing. *EMBO J.* **30**, 1634–1644 (2011).

44. Hugo, W. *et al.* Genomic and Transcriptomic Features of Response to Anti-PD-1 Therapy in Metastatic Melanoma. *Cell* **165**, 35–44 (2016).

45. Iglesia, M. D. *et al.* Genomic analysis of immune cell infiltrates across 11 tumor types. *J. Natl. Cancer Inst.* **108**, (2016).

46. Schreiber, R. D., Old, L. J. & Smyth, M. J. Cancer immunoediting: Integrating immunity's roles in cancer suppression and promotion. *Science* **331**, 1565–1570 (2011).

47. Shigehisa Kitano, A. I. Cancer Neoantigens: A Promising Source of Immunogens for Cancer Immunotherapy. *J. Clin. Cell. Immunol.* **06**, (2015).

48. Manber, U. & Myers, G. Suffix arrays: A new method for on-line string searches. in *Proceedings of the Annual ACM-SIAM Symposium on Discrete Algorithms* 319–327 (1990).

49. Holt, J. & McMillan, L. Merging of multi-string BWTs with applications. *Bioinformatics* **30**, 3524–3531 (2014).

50. Ferragina, P. & Manzini, G. Opportunistic data structures with applications. in *Annual Symposium on Foundations of Computer Science - Proceedings* 390–398 (2000).

51. Aho, A. V. & Corasick, M. J. Efficient String Matching: An Aid to Bibliographic Search. *Commun. ACM* **18**, 333–340 (1975).

52. Frankish, A. *et al.* GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res.* **47**, D766–D773 (2019).

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

54. Frazee, A. C., Jaffe, A. E., Langmead, B. & Leek, J. T. Polyester: Simulating RNA-seq datasets with differential transcript expression. *Bioinformatics* **31**, 2778–2784 (2015).

55. Dobin, A. *et al.* STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).

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

57. Trapnell, C. *et al.* Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* **28**, 511–515 (2010).

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

59. Wu, T. D. & Watanabe, C. K. GMAP: A genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* **21**, 1859–1875 (2005).

60. Grabherr, M. G. ., Brian J. Haas, Moran Yassour Joshua Z. Levin, Dawn A. Thompson, Ido Amit, Xian Adiconis, Lin Fan, Raktima Raychowdhury, Qiandong Zeng, Zehua Chen, Evan Mauceli, Nir Hacohen, Andreas Gnirke, Nicholas Rhind, Federica di Palma, Bruce W., N. & Friedman, and A. R. Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data. *Nat. Biotechnol.* **29**, 644–652 (2013).

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

62. Wang, K. *et al.* MapSplice: Accurate mapping of RNA-seq reads for splice junction discovery. *Nucleic Acids Res.* **38**, (2010).

63. Mose, L. E., Perou, C. M. & Parker, J. S. Improved indel detection in DNA and RNA via realignment with ABRA2. *Bioinformatics* (2019). doi:10.1093/bioinformatics/btz033

64. Orellana, C. F. *et al.* EDGAR: Full-length RNA transcript identification by hybrid sequencing and best edit-distance graph alignment of a single molecule read, in High Throughput Sequencing Algorithms and Applications (HitSeq 2016). in *ISMB* (2016).

65. Thorsson, V. *et al.* The Immune Landscape of Cancer. *Immunity* **48**, 812-830.e14 (2018).

66. Maiers, M., Gragert, L. & Klitz, W. High-resolution HLA alleles and haplotypes in the United States population. *Hum. Immunol.* **68**, 779–788 (2007).

67. Saito, R. *et al.* Molecular subtype-specific immunocompetent models of high-grade urothelial carcinoma reveal differential neoantigen expression and response to immunotherapy. *Cancer Res.* **78**, 3954–3968 (2018).

68. Colli, L. M. *et al.* Burden of nonsynonymous mutations among TCGA cancers and candidate immune checkpoint inhibitor responses. *Cancer Res.* **76**, 3767–3772 (2016).

69. Laumont, C. M. *et al.* Noncoding regions are the main source of targetable tumor-specific antigens. *Sci. Transl. Med.* **10**, (2018).

70. Weiss, R. B., Dunn, D. M., Atkins, J. F. & Gesteland, R. F. Slippery runs, shifty stops, backward steps, and forward hops: -2, -1, +1, +2, +5, and +6 ribosomal frameshifting. *Cold Spring Harb. Symp. Quant. Biol.* **52**, 687–693 (1987).

71. Saulquin, X. *et al.* +1 Frameshifting as a novel mechanism to generate a cryptic cytotoxic T lymphocyte epitope derived from human interleukin 10. *J. Exp. Med.* **195**, 353–358 (2002).

72. Macejak, D. G. & Sarnow, P. Internal initiation of translation mediated by the 5′ leader of a cellular mRNA. *Nature* **353**, 90–94 (1991).

73. Bullock, T. N. J., Patterson, A. E., Franlin, L. L., Notidis, E. & Eisenlohr, L. C. Initiation codon scanthrough versus termination codon readthrough demonstrates strong potential for major histocompatibility complex class I- restricted cryptic epitope expression. *J. Exp. Med.* **186**, 1051–1058 (1997).

74. Bullock, T. N. J. & Eisenlohr, L. C. Ribosomal scanning past the primary initiation codon as a mechanism for expression of CTL epitopes encoded in alternative reading frames. *J. Exp. Med.* **184**, 1319–1329 (1996).

75. Malarkannan, S., Horng, T., Shih, P. P., Schwab, S. & Shastri, N. Presentation of out-of-frame peptide/MHC class I complexes by a novel translation initiation mechanism. *Immunity* **10**, 681–690 (1999).

76. Van Den Eynde, B. J. *et al.* A new antigen recognized by cytolytic T lymphocytes on a human kidney tumor results from reverse strand transcription. *J. Exp. Med.* **190**, 1793–1799 (1999).

77. Bruce, A. G., Atkins, J. F. & Gesteland, R. F. tRNA anticodon replacement experiments show that ribosomal frameshifting can be caused by doublet decoding. *Proc. Natl. Acad. Sci. U. S. A.* **83**, 5062–5066 (1986).

78. Haas, B. J. *et al.* De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat. Protoc.* **8**, 1494–1512 (2013).

79. Smith, C. C. *et al.* Endogenous retroviral signatures predict immunotherapy response in clear cell renal cell carcinoma. *J. Clin. Invest.* **128**, 4804–4820 (2018).

80. Irvine, D. J., Hanson, M. C., Rakhra, K. & Tokatlian, T. Synthetic Nanoparticles for Vaccines and Immunotherapy. *Chemical Reviews* **115**, 11109–11146 (2015).

81. Jerby-Arnon, L. *et al.* A Cancer Cell Program Promotes T Cell Exclusion and Resistance to Checkpoint Blockade. *Cell* **175**, 984-997.e24 (2018).

82. Bruce, D. W. *et al.* Type 2 innate lymphoid cells treat and prevent acute gastrointestinal graft-versus-host disease. *J. Clin. Invest.* **127**, 1813–1825 (2017).

83. Roberts, P. J. *et al.* Combined PI3K/mTOR and MEK inhibition provides broad antitumor activity in faithful murine cancer models. *Clin. Cancer Res.* **18**, 5290–5303 (2012).

84. Cooke, K. R. *et al.* An experimental model of idiopathic pneumonia syndrome after bone marrow transplantation: 1. The roles of minor H antigens and endotoxin. *Blood* **88**, 3230–3239 (1996).

85. Wilkerson, M. D. *et al.* Integrated RNA and DNA sequencing improves mutation detection in low purity tumors. *Nucleic Acids Res.* **42**, (2014).

86. Coghill, J. M. *et al.* Separation of graft-versus-host disease from graft-versus-leukemia responses by targeting CC-chemokine receptor 7 on donor T cells. *Blood* **115**, 4914–4922 (2010).

87. Dolton, G. *et al.* More tricks with tetramers: A practical guide to staining T cells with peptide-MHC multimers. *Immunology* **146**, 11–22 (2015).

88. Wölfl, M. & Greenberg, P. D. Antigen-specific activation and cytokine-facilitated expansion of naive, human CD8+ T cells. *Nat. Protoc.* **9**, 950–966 (2014).

89. Quintarelli, C. *et al.* Cytotoxic T lymphocytes directed to the preferentially expressed antigen of melanoma (PRAME) target chronic myeloid leukemia. *Blood* **112**, 1876–1885 (2008).

90. Chan, K. S. *et al.* Identification, molecular characterization, clinical prognosis, and therapeutic targeting of human bladder tumor-initiating cells. *Proc. Natl. Acad. Sci. U. S. A.* **106**, 14016–14021 (2009).

91. Iglesia, M. D. *et al.* Prognostic B-cell signatures using mRNA-seq in patients with subtype-specific breast and ovarian cancer. *Clin. Cancer Res.* **20**, 3818–3829 (2014).

92. Prat, A. *et al.* Phenotypic and molecular characterization of the claudin-low intrinsic subtype of breast cancer. *Breast Cancer Res.* **12**, (2010).

93. Bindea, G. *et al.* Spatiotemporal dynamics of intratumoral immune cells reveal the immune landscape in human cancer. *Immunity* **39**, 782–795 (2013).

94. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, (2014).

95. Subramanian, A. *et al.* Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 15545–15550 (2005).

96. Ingenuity Systems. No Title. *Ingenuity Pathway Analysis* (2013). Available at: https://www.qiagenbioinformatics.com.

97. Huang, D. W. *et al.* The DAVID Gene Functional Classification Tool: A novel biological module-centric algorithm to functionally analyze large gene lists. *Genome Biol.* **8**, (2007).

98. Bolotin, D. A. *et al.* MiXCR: Software for comprehensive adaptive immunity profiling. *Nature Methods* **12**, 380–381 (2015).

99. Goebell, P. J. & Knowles, M. A. Bladder cancer or bladder cancers? Genetically distinct malignant conditions of the urothelium. *Urol. Oncol. Semin. Orig. Investig.* **28**, 409–428 (2010).

100. Volkmer, J. P. *et al.* Three differentiation states risk-stratify bladder cancer into distinct subtypes. *Proc. Natl. Acad. Sci. U. S. A.* **109**, 2078–2083 (2012).

101. Damrauer, J. S. *et al.* Intrinsic subtypes of high-grade bladder cancer reflect the hallmarks of breast cancer biology. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 3110–3115 (2014).

102. Weinstein, J. N. *et al.* Comprehensive molecular characterization of urothelial bladder carcinoma. *Nature* **507**, 315–322 (2014).

103. Dyrskjøt, L. *et al.* Identifying distinct classes of bladder carcinoma using microarrays. *Nat. Genet.* **33**, 90–96 (2003).

104. Sjödahl, G. *et al.* A molecular taxonomy for urothelial carcinoma. *Clin. Cancer Res.* **18**, 3377–3386 (2012).

105. Rebouissou, S. *et al.* EGFR as a potential therapeutic target for a subset of muscle-invasive bladder cancers presenting a basal-like phenotype. *Sci. Transl. Med.* **6**, (2014).

106. Choi, W. *et al.* Identification of Distinct Basal and Luminal Subtypes of Muscle-Invasive Bladder Cancer with Different Sensitivities to Frontline Chemotherapy. *Cancer Cell* **25**, 152–165 (2014).

107. Powles, T. *et al.* MPDL3280A (anti-PD-L1) treatment leads to clinical activity in metastatic bladder cancer. *Nature* **515**, 558–562 (2014).

108. Prat, A. *et al.* Characterization of cell lines derived from breast cancers and normal mammary tissues for the study of the intrinsic molecular subtypes. *Breast Cancer Res. Treat.* **142**, 237–255 (2013).

109. Liu, Y., Hayes, D. N., Nobel, A. & Marron, J. S. Statistical significance of clustering for high-dimension, low-sample size data. *J. Am. Stat. Assoc.* **103**, 1281–1293 (2008).

110. Choi, W. *et al.* Intrinsic basal and luminal subtypes of muscle-invasive bladder cancer. *Nature Reviews Urology* **11**, 400–410 (2014).

111. Hoadley, K. A. *et al.* Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. *Cell* **158**, 929–944 (2014).

112. Fridman, W. H. *et al.* Immune infiltration in human cancer: Prognostic significance and disease control. *Curr. Top. Microbiol. Immunol.* **344**, 1–24 (2010).

113. Sharma, P. *et al.* CD8 tumor-infiltrating lymphocytes are predictive of survival in muscle-invasive urothelial carcinoma. *Proc. Natl. Acad. Sci. U. S. A.* **104**, 3967–3972 (2007).

114. Horn, T. *et al.* The prognostic effect of tumour-infiltrating lymphocytic subpopulations in bladder cancer. *World J. Urol.* **34**, 181–187 (2016).

115. Tumeh, P. C. *et al.* PD-1 blockade induces responses by inhibiting adaptive immune resistance. *Nature* **515**, 568–571 (2014).

116. Choudhury, N. J. *et al.* Low T-cell Receptor Diversity, High Somatic Mutation Burden, and High Neoantigen Load as Predictors of Clinical Outcome in Muscle-invasive Bladder Cancer. *Eur. Urol. Focus* **2**, 445–452 (2016).

117. Snook, A. E., Magee, M. S., Schulz, S. & Waldman, S. A. Selective antigen-specific CD4+ T-cell, but not CD8+ T- or B-cell, tolerance corrupts cancer immunotherapy. *Eur. J. Immunol.* **44**, 1956–1966 (2014).

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

119. Snyder, A. *et al.* Genetic basis for clinical response to CTLA-4 blockade in melanoma. *N. Engl. J. Med.* **371**, 2189–2199 (2014).

120. Rizvi, N. A. *et al.* Mutational landscape determines sensitivity to PD-1 blockade in non-small cell lung cancer. *Science (80-.).* **348**, 124–128 (2015).

121. Choi, N., Zhang, B., Zhang, L., Ittmann, M. & Xin, L. Adult Murine Prostate Basal and Luminal Cells Are Self-Sustained Lineages that Can Both Serve as Targets for Prostate Cancer Initiation. *Cancer Cell* **21**, 253–265 (2012).

122. Peters, J. M., Shah, Y. M. & Gonzalez, F. J. The role of peroxisome proliferator-activated

receptors in carcinogenesis and chemoprevention. *Nature Reviews Cancer* **12**, 181–195 (2012).

123. Mak, M. P. *et al.* A Patient-Derived, Pan-Cancer EMT Signature Identifies Global Molecular Alterations and Immune Target Enrichment Following Epithelial-to-Mesenchymal Transition. *Clin. Cancer Res.* **22**, 609–620 (2016).

124. Chen, L. *et al.* Metastasis is regulated via microRNA-200/ZEB1 axis control of tumour cell PD-L1 expression and intratumoral immunosuppression. *Nat. Commun.* **5**, 5241 (2014).

125. Rooney, M. S., Shukla, S. A., Wu, C. J., Getz, G. & Hacohen, N. Molecular and genetic properties of tumors associated with local immune cytolytic activity. *Cell* **160**, 48–61 (2015).

126. Hammerman, P. S. *et al.* Comprehensive genomic characterization of squamous cell lung cancers. *Nature* **489**, 519–525 (2012).

127. 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* **15**, 409–428 (2014).

128. Lin, N. U. *et al.* Sites of distant recurrence and clinical outcomes in patients with metastatic triple-negative breast cancer: High incidence of central nervous system metastases. *Cancer* **113**, 2638–2645 (2008).

129. Iqbal, J., Ginsburg, O., Rochon, P. A., Sun, P. & Narod, S. A. Differences in breast cancer stage at diagnosis and cancer-specific survival by race and ethnicity in the United States. *JAMA - J. Am. Med. Assoc.* **313**, 165–173 (2015).

130. Plasilova, M. L. *et al.* Features of triple-negative breast cancer Analysis of 38,813 cases from the national cancer database. *Med. (United States)* **95**, (2016).

131. Niwińska, A., Murawska, M. & Pogoda, K. Breast cancer brain metastases: Differences in survival depending on biological subtype, RPA RTOG prognostic class and systemic treatment after whole-brain radiotherapy (WBRT). *Ann. Oncol.* **21**, 942–948 (2009).

132. Emens, L. A. Breast cancer immunotherapy: Facts and hopes. *Clinical Cancer Research* **24**, 511–520 (2018).

133. Nanda, R. *et al.* Pembrolizumab in patients with advanced triple-negative breast cancer: Phase Ib keynote-012 study. *J. Clin. Oncol.* **34**, 2460–2467 (2016).

134. Lynch, D. H. The promise of 4-1BB (CD137)-mediated immunomodulation and the immunotherapy of cancer. *Immunological Reviews* **222**, 277–286 (2008).

135. Quail, D. F. *et al.* The tumor microenvironment underlies acquired resistance to CSF-1R inhibition in gliomas. *Science* **352**, (2016).

136. Sharabi, A. B. *et al.* Stereotactic radiation therapy augments antigen-specific PD-1-mediated antitumor immune responses via cross-presentation of tumor antigen. *Cancer Immunol. Res.* **3**, 345–355 (2015).

137. Vinayak, S. *et al.* TOPACIO/Keynote-162: Niraparib + pembrolizumab in patients (pts) with metastatic triple-negative breast cancer (TNBC), a phase 2 trial. *J. Clin. Oncol.* (2018). doi:10.1200/jco.2018.36.15_suppl.1011

138. Li, B. & Dewey, C. N. RSEM: Accurate transcript quantification from RNA-seq data with

or without a reference genome. in *Bioinformatics: The Impact of Accurate Quantification on Proteomic and Genetic Analysis and Research* (2014). doi:10.1201/b16589

139. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**, 589–595 (2010).

140. Mose, L. E., Wilkerson, M. D., Neil Hayes, D., Perou, C. M. & Parker, J. S. ABRA: Improved coding indel detection via assembly-based realignment. *Bioinformatics* **30**, 2813–2815 (2014).

141. Cingolani, P. *et al.* 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 (Austin).* **6**, 80–92 (2012).

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

143. Cingolani, P. *et al.* Using Drosophila melanogaster as a model for genotoxic chemical mutational studies with a new program, SnpSift. *Front. Genet.* **3**, (2012).

144. Sherry, S. T. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* **29**, 308–311 (2001).

145. Karczewski, K. J. *et al.* The ExAC browser: Displaying reference data information from over 60 000 exomes. *Nucleic Acids Res.* **45**, D840–D845 (2017).

146. Forbes, S. A. *et al.* COSMIC: Exploring the world’s knowledge of somatic mutations in human cancer. *Nucleic Acids Res.* **43**, D805–D811 (2015).

147. Gu, Z., Gu, L., Eils, R., Schlesner, M. & Brors, B. Circlize implements and enhances circular visualization in R. *Bioinformatics* **30**, 2811–2812 (2014).

148. Gu, Z., Eils, R. & Schlesner, M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics* **32**, 2847–2849 (2016).

149. Mose, L. E. *et al.* Assembly-based inference of B-cell receptor repertoires from short read RNA sequencing data with V’DJer. *Bioinformatics* **32**, 3729–3734 (2016).

150. Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. & Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods* **14**, 417–419 (2017).

151. Liberzon, A. *et al.* Molecular signatures database (MSigDB) 3.0. *Bioinformatics* **27**, 1739–1740 (2011).