From Coursework to Innovation: A Student-Driven *In Silico* Framework for Personalized Cancer Vaccine Design

B. J. Hlavinka[1,2], S. Richardson[4], R. Nurdel[1], H. Cao[1], J. Mclemore[1,3], O. Rogers[1], S. A. Curran[2,5]

[1] Department of Biology & Biochemistry, University of Houston, Science & Research Building 2, 3455 Cullen Blvd., Houston, TX 77004 USA

[2] Curran's Nanophysics Research Group, University of Houston, Science & Research Building 1, 3507 Cullen Blvd., Houston, TX 77204 USA

[3] Department of Mathematics, University of Houston, Philip Guthrie Hoffman Hall, 3551 Cullen Blvd, Houston, TX 77004 USA

[4] Department of Biomedical Engineering, University of Houston, Science and Engineering Research Center, 3517 Cullen Blvd, Room 2027, Houston, TX 77204

[5] Department of Physics, University of Houston, Science & Research Building 1, 3507 Cullen Blvd., Houston, TX 77204 USA

Abstract

Modern oncology is currently undergoing a transformative shift from broad-spectrum treatments to personalized molecular interventions, yet a significant gap remains between rapid biotechnological advancement and public understanding. This research project, emerging from a hybridized undergraduate and postgraduate Biomedical and Industrial Genomics course at the University of Houston, explores a reproducible *in silico* workflow[27,28] developed and put into practice by the University of Houston Sequencing Core under the guidance of Dr. Preethi Gunaratne for the development of personalized neoantigen vaccines. A case study utilizing real-world de-identified patient data consisting of an MMP14/C1QBP gene fusion associated with aggressive Non-Hodgkin Lymphoma demonstrates how raw genomic split-reads can be transformed into a tailored therapeutic roadmap. At its core, this project was a catalyst for student-led innovation, as researchers were granted full-autonomy to integrate their unique academic

backgrounds into the modeling of therapeutic interventions. This collaborative environment fostered creative discovery, yielding diverse intellectual contributions ranging from specialized radiological interventions and microfluidic monitoring to the development of sophisticated algorithmic optimization programs. By navigating the transition from genomic education to hypothetical clinical synthesis, the class successfully bridged the gap between theoretical knowledge and practical application. The findings indicate that while industrial scaling for generic off-the-shelf vaccines remains a long-term challenge due to high genomic diversity, the boutique patient-specific model is a viable next-generation feasibility. This work underscores the *Translational Ideal*: the inherent necessity of transparent scientific communication to demystify complex genomics, replace intellectual curiosity with data-driven understanding, and ensure that the next generation of life-saving medicine is both mathematically optimized and biologically sound.

Introduction

Clinical Context

The landscape of modern oncology is shifting from broad-spectrum treatments to personalized molecular interventions, yet a significant gap remains between rapid biotechnological advancement and public understanding. While surgery, radiation, and traditional chemotherapy remain foundational, current guidelines emphasize combining these with immunotherapy and targeted treatments to overcome drug resistance and reduce systemic toxicity. It is the goal of the authors to demystify the process of creating targeted therapy proposals, demonstrating that these interventions are scientifically optimized, biologically grounded, and rigorously tested against real-world benchmarks. This paper serves as a technical and educational case study in biomedical and industrial genomics, utilizing real-world de-identified oncological patient data to design a

hypothetical multi-tiered therapeutic strategy based on previous work envisioned and implemented by The University of Houston Sequencing Core (UHSeq) combined with student researcher innovations.

Collaborative Framework

This research emerged from a unique split-level Biomedical and Industrial Genomics course taught by Dr. Preethi Gunaratne that hybridized the learning experiences and output goals of student researchers at both the undergraduate and postgraduate level. The academic structure was designed to mirror a professional dry-laboratory environment. Each researcher was tasked with an *in silico* workflow to take a raw oncologic gene fusion extracted from RNA-sequencing data and, in a stepwise process, offer a hypothetical peptide pool for the de-identified patient to whom the gene fusion belonged. Researchers were then grouped according to cancer types and collaborative work culminated in a final presentation of individualized therapeutic models where each student researcher offered intellectual contributions to the modeling of therapeutic intervention for the overall cancer type. These groups were given carte blanche to use their own expertise and knowledge from other courses and internship experiences when creating the hypothetical therapeutic approach, resulting in discovery and innovation far beyond the scope of the class. Outside the technical results, the course aimed to familiarize student researchers with cutting edge clinical genomics techniques and necessary background information regarding the methods used during the workflow while fostering the environment for creativity.

The Workflow: Project Overview

The first part of this project follows a reproducible pipeline designed by UHSeq to turn a genetic anomaly into a molecular target. This workflow is broadly divided into four critical phases:

1. Genomic Identification: Detecting chromosomal translocations (split-reads) that create non-self, chimeric RNA transcripts found exclusively in tumor cells.

2. Molecular Intervention: Utilizing CRISPR-Cas9 to target the 3' UTR of the fusion transcript, ensuring the oncogenic protein is expressed at levels sufficient for immune recognition.

3. Peptide Optimization: Processing the fusion protein to identify a core neopeptide pool based on Human Leukocyte Antigen[6] (HLA) coverage and thermodynamic stability.

4. Clinical Integration: Benchmarking results against Patient-Derived Xenograft (PDX) mouse models to predict treatment efficacy and determine the best drug-vaccine combinations for the de-identified patient.

This workflow was done on an individual basis with each student researcher assigned a split-read based on de-identified patient-specific oncogenic tissue that was lysed, sequenced, and analyzed by UHSeq. From this single bit of data, all other processes were carried out on open access platforms, meaning if *anyone* had access to a split-read, they could perform this workflow (provided they had the understanding of genomics required to carry it forward).

The Translational Ideal

The ultimate aim of this communication article is two-fold: to demystify and destigmatize the idea that cutting-edge techniques and analytics can only be performed following extensive postgraduate experience and training and to emphasize that collaborative, multidisciplinary scientific approaches to some of the greatest contemporary scientific problems can yield novel outcomes—even when proposed at the undergraduate level. By providing a transparent, data-driven

explanation of mRNA-based therapies, we demonstrate that these interventions are not experimental gambles, but mathematically-optimized and biologically grounded tools.

> *Author's Note/Academic Disclaimer: while the workflow imagined and implemented by UHSeq uses de-identified patient data and industrial-grade software at the forefront of cancer research, the results reported herein are* in silico *models designed to test the feasibility of rapid, personalized vaccine design and not meant to represent true clinical proposals. Student inventions are proprietary and patient data has been depersonalized. For ease of communication and unless otherwise noted, this paper will exemplify the MMP14/C1QBP gene fusion and its related cancer—Activated B-Cell-like Diffuse Large B-Cell Lymphoma (BCL-ABC), an aggressive and metastatic non-Hodgkin lymphoma—as this was the split read assigned to the primary author.*

Methods

Split Read

AAGGATGGCAAATTCGTCTTCTTCAAAG___GAGACAAGCATTGGGTGTTTGATGAGGC
GTCCCTGGAACCTGGCTACCCCAAGCACATTAAGGAGCTGGGCCGAGGGCTGCCTACC
GACAAGATTGATGCTGCTCTCTTCTGGATGCCCAATGGAAAGACCTACTTCTTCCGTGG
AAACAA___GTACTACCGTTTCAACGAAGAGCTCAGGGCAGTGGATAGCGAGTACCCCA
AGAACATCAAAGTCTGGGAAGGGATCCCTGAGTCTCCCAGAGGGTCATTCATGGGCAG
CGAT**GAAG|GAGA**CAAAGCTTTTGTTGATTTCCTGAGTGATGAAATTAAGGAGGAAAG
AAAAATTCAGAAGCAT

Figure 2.1—*The initial split read data given. Emboldened region shows the fusion point of two otherwise unassociated genes.*

Split-reads: Identification of structural variations where a single read of patient-specific DNA maps to different locations, marking breakpoints.

The primary project began with the identification of a unique chromosomal translocation wherein a portion of one gene fused with another to create a chimeric RNA transcript—a non-self target found exclusively in tumor cells. These gene fusions still transcribe mRNA, provided they encode an open reading frame (ORF). This anomalous junction can be exploited through genomic intervention to function as a molecular beacon for the immune system, signaling that the tumor cells to which they (or their resultant protein translations) belong should be sought out as invasive and biologically harmful. From here, the original gene fusion partners must be identified.

Blat

The entire split read was entered into BLAT[9] and the two sequences with the highest identities that were aligned at "Start" and "End" were determined to be the best models for gene fusion partners. The 5' partner was found to be **MMP14 (exons 1-9)** of Chromosome 14**.** The model **NM_004995.4** was the best fit for the purpose as it contained the most exons to be used for building the fusion model. The 3' partner was found to be **C1QBP (exons 2-6)** on Chromosome 17. The **NM_001212.4** model was used. After the gene partners were determined, the combined fusion gene MMP14/C1QBP was stitched together, confirming the breakpoint given in the split-read.

CRISPR

Student researchers were tasked with designing CRISPR-Guide RNAs (gRNAs) to block microRNA(miRNA)-mediated downregulation of the fusion transcript. The 3' partner of the fusion sequence was entered into TargetScan[20] as it contained the 3' untranslated region (UTR) (the site at which miRNA can bind and turn off a translated mRNA sequence from being transcribed). TargetScan returned an 8-base pair target seed region in position 78-85 of the 3' UTR.

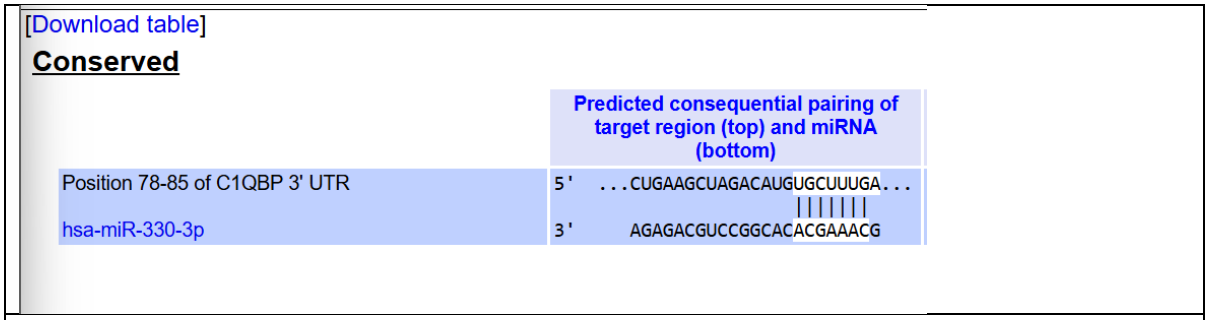

Figure 2.3—*Seed region (white highlight) for miRNA interaction found in the 3' Untranslated Region of the second fusion partner. Position, specific miRNA interaction and antisense pairing are given.*

3'-UTR of C1QBP where seed is found:

5'AGCAGACAGATGCTGAAAGCCATAGTTTCATGGCAGGCTTTGGCCAGTGAACA

AATCCTACTCTGAA**GCTAGACATGTGCTTTG*AAA**TGATTATCATCCTAATATCAT

GGGGGAAAAAATACCAAATTTAAATTATATGTTTTGTGTTCTCATTTATTATCATT

TTTTTCTGTACAAATCTATTATTTCTAGATTTTTGTATAACATGATAGACATAAAAT

TGGTTTATCTCCTCCAA**-3'**

To inhibit the ability of a miRNA to downregulate this transcription, this binding seed region must be cleaved in the DNA. This seed can be a cleavage attack site for a CRISPR-Cas complex but requires a guide RNA and the appropriate CRISPR-associated (Cas) protein to function as described. To design this functional complex, the seed region was found in the 3' UTR of the C1QBP DNA by searching for the DNA equivalent region (substituting Ts for Us: UGCUUUGAàTGCTTTGA). The protospacer adjacent motif (PAM) was determined by interpretation of Cas rules. Cas9 will cut between the 3rd and 4th nucleotides 5' to the PAM site if the motif sequence reads "NGG" where N is any nucleotide and G is Guanine. While Cas9 typically utilizes NGG, non-canonical PAMs are increasingly recognized in specialized industrial genomics contexts. Therefore, this PAM site was proposed to be TGA (underlined) for this *in silico* design and the Cas used was a Cas9. The CRISPR cleavage site is given by the asteriskàTGCTTTG*A. The guide RNA (gRNA) attachment site is highlighted on the DNA. This sequence needed to be retranscribed into RNA to build the gRNA à GCUAGACAUGUGCUUUGAAA.

This strategy established a "killswitch-killswitch" against normal non-coding RNA function using CRISPR-Cas9 to target the 3' UTR seed region of hsa-miR-330-3p, effectively

disrupting the oncogenic transcript's ability to be prematurely downregulated by microRNA binding. By keeping the chimeric mRNA active for translation, the CRISPR-Cas cleavage ensures that the oncogenic cell produces an abundance of the fusion protein to be marked for immunological intervention. By isolating and identifying the fusion region, researchers continued to the ultimate clinical goal of this project—development of a personalized neoantigen vaccine using small neopeptides mimicking the protein fusion region.

ORF Finder

Student researchers were tasked with extracting the open reading frame and building a general neopeptide pool. For this fusion to be a viable target for peptide vaccine therapy, it must be able to be translated—there must be an open reading frame created by the fusion without any premature stop codons created by the merging of the gene transcript. The nucleotide sequence of the fusion gene was inserted into the NCBI Open Reading Frame Finder[15] to determine its open reading frame, then compared against the wild type-peptides given in RefSeq[1,14] from both fusion partners to confirm the fusion sequence. The ORF of the fusion gene was compared against the coding sequence (CDS) of the two wild type contributors to determine peptide sequence and if there were any other mutations contributing to the hybrid translation by running through BLASTp and comparing alignments. After comparison of alignments, the 677 amino acid neopeptide was stitched together at the 5' partner translation to 3' partner translation fusion point.

MSPAPRPPRCLLLPLLTLGTALASLGSAQSSSFSPEAWLQQYGYLPPGDLRTHTQRSPQSLSAA
IAAMQKFYGLQVTGKADADTMKAMRRPRCGVPDKFGAEIKANVRRKRYAIQGLKWQHN
EITFCIQNYTPKVGEYATYEAIRKAFRVWESATPLRFREVPYAYIREGHEKQADIMIFFAEGF
HGDSTPFDGEGGFLAHAYFPGPNIGGDTHFDSAEPWTVRNEDLNGNDIFLVAVHELGHA
LGLEHSSDPSAIMAPFYQWMDTENFVLPDDDRRGIQQLYGGESGFPTKMPPQPRTTSRPSV
PDKPKNPTYGPNICDGNFDTVAMLRGEMFVFKERWFWRVRNNQVMDGYPMPIGQFWRG
LPASINTAYERKDGKFVFFKGDKHWVFDEASLEPGYPKHIKELGRGLPTDKIDAALFWMP
NGKTYFFRGNKYYRFNEELRAVDSEYPKNIKVWEGIPES**PRGSFMGSDE**
--
GDKAFVDFLSDEIKEERKIQKHKTLPKMSGGWELELNGTEAKLVRKVAGEKITVTFNINN

SIPPTFDGEEEPSQGQKVEEQEPELTSTPNFVVEVIKNDDGKKALVLDCHYPEDEVGQED EAESDIFSIREVSFQSTGESEWKDTNYTLNTDSLDWALYDHLMDFLADRGVDNTFADELV ELSTALEHQEYITFLEDLKSFVKSQ
Figure 2.4—677aa neoepitope fusion peptide. For ease of determination, translation from 5' partner is listed above, 3' partner is listed below. In reality, the gene fusion partners translate seamlessly creating a viable protein that exists nowhere in the body except the tumor environment. Emboldened amino acids show the 20-peptide sequence used to build the neoantigen pool.

9mers, 10mers, 11mers

The 20-peptide sequence used to build the peptide pool was determined to be **PRGSFMGSDE|GDKAFVDFLS.** Eleven-amino-acid-long polymers (11mers) were built by ensuring that at least one peptide on either side of the fusion contributed to a neopeptide—ensuring that no wild type peptides would be perturbed by the autoimmune response—after which 9mers and 10mers were built from the 11mer pool, an optimal size for MHC-I binding.

9-11mers:		
9mers	**10mers**	**11mers**
GSFMGSDE|G **SFMGSDE|GD** **FMGSDE|GDK** **MGSDE|GDKA** **GSDE|GDKAF** **SDE|GDKAFV** **DE|GDKAFVD** **E|GDKAFVDF**	**RGSFMGSDE|G** **GSFMGSDE|GD** **SFMGSDE|GDK** **FMGSDE|GDKA** **MGSDE|GDKAF** **GSDE|GDKAFV** **SDE|GDKAFVD** **DE|GDKAFVDF** **E|GDKAFVDFL**	**PRGSFMGSDE|G** **RGSFMGSDE|GD** **GSFMGSDE|GDK** **SFMGSDE|GDKA** **FMGSDE|GDKAF** **MGSDE|GDKAFV** **GSDE|GDKAFVD** **SDE|GDKAFVDF** **DE|GDKAFVDFL** **E|GDKAFVDFLS**
Figure 2.5—*neopeptide pool created around the junction (|) of the peptides, shown here with at least one amino acid on either side of the fusion to maintain non-wild type determination.*		

Viability of the Peptide Pool

UHSeq then provided student researchers with the patient's Optitype[19] (patient-specific HLA typing) file and a NetMHCPan[23] (bioinformatics tool used to predict binding between peptides and Major Histocompatibility Complex molecules) file and tasked with determining the best peptide pool for their patient. The Optitype for the MMP14/C1QBP patient:

A*02:01, A*68:01 B*44:02, B*35:01 C*05:01, C*04:01

Human Leukocyte Antigen Nomenclature: Gene Locus (A) Antigen Group (02:) specific protein type (01)*

After finding the associated Human Leukocyte Antigens (HLAs)—proteins found on the surface of most nucleated cells that act as identification tags, allowing the immune system to distinguish between healthy self-cells and foreign invaders—in the NetMHCPan file, any peptides that bound fewer than 3 HLAs or had an eluted ligand percentile (EL) rank higher than 32.7, were considered low priority peptides and were initially discarded from the proposed neopeptide vaccine pool. However, some anomalies needed to be considered. The question of how best to weigh the coverage (number of HLAs bound by a single peptide) against the ability to bind at the surface of the cell (EL rank) arose. An early assumption to average the binding ability across all peptides while giving the highest priority to the most coverage proved not to be fruitful.

Fusion genes	MMP14-C1QBP					
PDX model	ST5380					
Peptides	A*02:01	A*68:01	B*44:02	B*35:01	C*05:01	C*04:01
FMGSDEGDK	29.96	20.5				
MGSDEGDKA				16.065	17.417	
GSDEGDKAF	12.428		4.893	2.468	0.09	0.649
SDEGDKAFV			13.223		19.595	15.099
SFMGSDEGDK		9.526				
FMGSDEGDKA	5.955				20.909	
MGSDEGDKAF	27.613		9	2.005	0.327	1.817
GSDEGDKAFV	10.866				1.327	12.391
GSFMGSDEGDK		8.681				
SFMGSDEGDKA	28.37					25.923
FMGSDEGDKAF	2.973		12.05	2.672	0.208	0.61
MGSDEGDKAFV	30.182			25.1	6.113	20.808
DEGDKAFVD			23.75			
EGDKAFVDF			13.202	3.367	1.143	1.185
SDEGDKAFVD			27.222			
DEGDKAFVDF			0.984	12.5	14.419	18.985
EGDKAFVDFL			22.667	27.312	6.089	13.689
GSDEGDKAFVD					15.432	
SDEGDKAFVDF			7.186	22.371	8.79	19.365
DEGDKAFVDFL			5.357	28.857		

Figure 2.6—*The twenty best peptides for the patient, determined solely by EL Rank. Only 4 of the 24 peptides had EL rank above 32.7 across the board, showing no binding. These are blacked out and correspond with data given as "No Binding" by the Peptide Pipeline. Note A*68:01 only has coverage by 3 neopeptides, none of which cover ≥3 HLAs*

Optimizing the Neopeptide Pool

Student researchers were tasked with creating a vaccine for their patient using the 20 "best" neopeptides. "Best" was to be determined by further research including de-identified patient specific data linked to Patient Derived Xenograft (PDX) mouse models which included data about previous treatment and therapeutic exposures, the tumor extraction site, and how the mouse models reacted to specific chemo interventions as well as any immunotherapy that had been assayed. Further Human Protein Atlas[10] research into the fusion gene partners in relation to their biological function and clinical relevance in relation to assigned cancer type and other cancers enabled students to build hypotheses on the potential impact of the resultant fusion protein. Student researchers were further grouped by cancer types and instructed to contribute to a presentation detailing these cancer specifics, the individual projects, and the final, most lofty goal of developing a therapeutic regimen for these patients including, if at all possible, the development of a multivalent generic cancer immune-booster. Thus ended the established workflow provided by UHSeq, but what followed revealed the ingenuity of the students and the true novelty of this class.

Student Researcher Innovations

Student groups were granted autonomy to utilize their unique expertise and knowledge acquired through other classes and internships leading to discoveries and innovations that extended far beyond the standard curriculum. This collaborative environment encouraged creativity within the class, where researchers brought their diverse academic backgrounds to solve the problem of

personalized therapy. These contributions generally trended toward the answering of four central questions regarding optimization, safety, patient outcome, and communication:

How best to optimize the neopeptide pool?

The core of the therapeutic challenge involved determining the most effective method for prioritizing neopeptide ranking, specifically weighing binding affinity (EL Rank) against the breadth of HLA coverage. A simple average was seen to be ineffective as discussed in section 2.6, so several students used their own intuition to propose different prioritizations; however, two students rose to the challenge by developing two distinct algorithmic approaches.

Hy Cao's Peptide Pipeline (v5.1) is a Python-based engine designed to prioritize the dynamic biological constraints of a single patient. Its primary innovation lies in the integration of the Pepsickle library, which models proteasomal cleavage to ensure that the selected neopeptides are naturally produced in vivo and avoids "self" mimicry that could be seen as wild type and ultimately be ignored by the immune system. The pipeline utilizes differential evolution and stochastic optimization to navigate a complex reward and penalty system. This logic favors peptides with maximal HLA coverage while penalizing those with poor thermodynamic stability or high EL Ranks. Furthermore, the pipeline employs automated "knee detection" to mathematically identify ideal cutoff points for core and support pools, ensuring the final selection is grounded in mathematical optimization rather than arbitrary thresholds.

In contrast, Jacob McLemore engineered Möbi, a MATLAB-based workflow that approaches the selection problem through Topological Data Analysis (TDA) and population-level clustering. While the Peptide Pipeline focuses on the specific molecular viability of the peptide, Möbi focuses on the high-dimensional patient-level binding behavior within a broader feature space. The program converts patient-specific HLA binding profiles into feature vectors, creating

a matrix that represents the binding landscape of an entire population. By examining the distance and connectivity structure between these vectors—a process known as persistent homology—Möbi divides patients into subgroups based on shared biological traits. Within these topological subgroups, the program applies a linear multi-objective greedy optimization to select candidate peptide pools that provide the most efficient coverage for potential multivalent vaccines.

The "duel" between these programs provided the project with a comprehensive toolkit that balanced biological fidelity with topological efficiency. For our purposes, Cao's proposed pipeline became the most specialized algorithm for "boutique," patient-specific precision by ensuring each peptide is produced and bound efficiently within the host. Conversely, Mclemore's Möbi provides a roadmap for the industrial scaling of neoantigen vaccines by identifying shared immunological landscapes across diverse patient groups.

How to ensure patient safety?

Safety was addressed through a combination of biochemical modeling and genomic surveillance. The Peptide Pipeline integrated Pepsickle[24] modeling to generate cleavage scores, ensuring neopeptides are naturally produced *in vivo* without mimicking "self" sequences that could trigger autoimmunity. Complementing this, one graduate student developed a rigorous validation protocol using the DAVID[25] functional annotation tool to map signaling pathways and the COSMID[26] platform to identify potential CRISPR gRNA off-target effects across the human genome.

How to improve patient outcomes?

Researchers within the primary author's group (Non-Hodgkin Lymphoma) looked toward a holistic "trivalent" strategy to improve clinical success, combining the personalized vaccine with standard-of-care chemotherapy and radiological intervention to target solid tumor metastases.

Sophia Richardson proposed the integration of microfluidic capture systems as a liquid biopsy tool to isolate cancerous B-cells (CD19/CD20), allowing for non-invasive, continuous monitoring of recurrence. Additionally, in other research groups, metabolic interventions such as a ketogenic diet were suggested to exploit the Warburg effect chokepoint, effectively starving metastatic cells that rely on aerobic glycolysis for energy.

How best to communicate the findings?

While the individual presentations highlighted diverse areas of interest, the most significant innovation in communication came from the pedagogical requirement to engage in cross-disciplinary defense. Student researchers were not merely presenting results; they were tasked with answering critical inquiries from peers regarding their specific methodologies. This interrogative approach forced researchers to move beyond their specialized silos to ensure their findings were accessible to the entire cohort.

This academic structure transformed the class from a collection of individuals into a unified think tank. Because not all participants entered the course with the same prerequisite knowledge in NGS or bioinformatics, the high level of research acted as a forced catalyst for democratization. Researchers became science communicators by necessity, drafting explanations and answering peer questions that synthesized the raw data and therapeutic logic. Consequently, this paper represents the project's shift from a laboratory workflow to a collaborative social educational framework, proving that the complexity of industrial genomics can be managed through intentional, competent, peer-led education under the guidance of researchers that paved the way through novel innovations in the field.

Results and Discussion

Challenges Faced

The transition from *in silico* design to a hypothetical clinical roadmap revealed several critical hurdles that required a blend of algorithmic precision and human intuition.

Biological Context and Algorithmic Limitations

In the specific case of the MMP14/C1QBP fusion, researchers encountered significant coverage issues. Specifically, the orphan allele HLA A*68:01 initially lacked representation in the automated peptide pools. This required manual human intervention to override algorithmic thresholds and ensure the patient had "zero-blind-spot" immune coverage for their Optitype HLAs. By widening the scope of the Peptide Pipeline to Common HLAs outside of the patient specific Optitype, the Core Pool of peptides for the vaccination could be adjusted, along with an HLA A*68:01-specific peptide inclusion in each of the support pools.

Cancer Specificity

Because ABC-BCL is a hematologic cancer that is metastatic nearly by definition, the strategy had to account for systemic mobility and the high risk of relapse associated with current standards of care. To address this, the therapeutic strategy moved beyond simple tumor reduction to focus on the biological anomaly created by the MMP14/C1QBP fusion. Since MMP14 acts as a membrane-bound proteolytic enzyme that degrades the extracellular matrix to facilitate cellular escape[3,8,14], and C1QBP simultaneously reprograms energy metabolism via the Warburg effect to sustain growth in hypoxic environments[1,7], the intervention had to be equally multi-faceted. The resulting trivalent approach combined a personalized neoantigen vaccine in combination with monitoring by microfluidic device—designed to provide long-term immunological surveillance against these specific fusion-positive cells—with standard-of-care R-CHOP[12,13] chemotherapy and localized radiological intervention. This integrated defense acknowledges that while chemotherapy may yield an initial -100% T/C ratio in PDX models, the high 40% rate of relapsed or refractory disease

necessitates a highly specific, systemic safety net to target mobile, chemo-resistant cells before they can establish new metastatic masses.

Building a multivalent vaccine for this cancer family remains a significant challenge because the gene fusions analyzed were exceptionally diverse, with each exhibiting a vastly different profile of the Hallmarks of Cancer.[4,5] Because the non-self targets are so heterogeneous, a multivalent pool would require the exact same fusion point across multiple patients, as well as a standardized HLA coverage that accounts for individual genomic contraindications. Consequently, the ideal of an off-the-shelf generic vaccine is currently secondary to the more immediate success of the boutique, patient-specific model. This reality underscores the necessity of the *in silico* workflow to determine if a shared therapeutic target even exists before attempting to scale to an industrial level.

Cross-disciplinary divide: A working example

Perhaps the most pervasive challenge was intellectual opacity—the tendency for high-level research to become inaccessible to those with different foundational backgrounds. This was exemplified by student researcher Ryan Nurdel, who initially struggled with the complexities of bioinformatics and NGS. However, Nurdel's experience ultimately served as the catalyst for one of the project's most significant innovations. As a graduate student with experience in radiation oncology research using Linear Accelerator (Linac) applications, Nurdel's enthusiasm swelledthe moment he realized his specific field of study was essential to the patient's survival. By putting aside academic pride and engaging in peer-to-peer communication—tutored by the primary author—Nurdel was able to bridge his specialized knowledge into the group's final model; the integration of which completely transformed the group's proposal into a robust trivalent approach.

Because one patient presented with a solid metastatic retroperitoneal tumor, he proposed targeted radiation therapy as a primary candidate for eliminating these localized malignancies. He specifically advocated for the use of the Linac system to address the high tumor grade and metastatic capability inherent in the MMP14/C1QBP fusion. Furthermore, Nurdel's plan included involved field radiation for cancerous lymph nodes, providing a critical physical intervention to complement the hypothetical conferred immunity of the neopeptide vaccine. This contribution ultimately proved that the "Translational Ideal" is a systematic necessity. True scientific progress only occurs when researchers demystify their individual expertise to build a holistic, multidisciplinary defense.

Moving Forward

It must be reiterated that all results and proposals reported herein were based on new, biologically unvalidated fusions used for the purpose of familiarizing student researchers with NGS and current methods and techniques in practice at UHSeq and therefore only presented hypothetically. As dictated by a ranking member of UHSeq, this approach would not be appropriate for a methods paper. For instance: while comparing Cao and Mclemore's programs, we have have left the actual computational methods and mathematical mechanics largely unexplained. Stronger validation for the Peptide Pipeline could be obtained by demonstrating its performance using published datasets with experimentally validated neoantigens. However, the primary intention of this paper is to communicate the educational success and novelty of approach for a split-level biomedical and industrial genomics course, which leaves these future methods open to validation.

The impact of this course has extended into professional and institutional roles. Several researchers have since joined UHSeq, and the architects of the Peptide Pipeline and Möbi are currently collaborating on a new utility program for the Core. Olive Rogers has transitioned into

the Head of Extraction for UHSeq, balancing this role with managing all undergraduate researchers in the lab and some business administration. She provided the context for how the split-reads were designed for each of the student researchers.

Looking ahead, there is a strong desire to expand these findings into broader educational outreach, potentially through a specialized glossary or communication curriculum to continue demystifying genomic breakthroughs for a curious public.

Conclusion

Ultimately, this project confirms the feasibility of a "boutique" approach to precision medicine, demonstrating that an actionable, experimental workflow can transform a single split-read into a tailored therapeutic roadmap. By navigating the transition from genomic discovery to algorithmic optimization, we have furthered the idea that personalized vaccine design is not an abstract future goal but a present *in silico* hypothesis as can be seen in the workflow[27,28] established by the researchers in UHSeq. While transitioning this individualized service to an industrial off-the-shelf generic model presents significant hurdles (including the need for exhaustive patient sequencing to avoid "self" contraindications) the methodology established by UHSeq and presented here by the undergraduate authors provides a clear path forward for addressing rare and aggressive cancer-related gene fusions.

The success of this research was inseparably linked to the hybrid academic structure of the course, which highlighted a critical need for enhanced scientific communication. The intellectual exchange between researchers, such as the collaborative development of the trivalent approach and the peer-to-peer tutoring that resolved foundational knowledge gaps, proved that the translational ideal is an inherent necessity for progress. As we move beyond the laboratory, it is incumbent upon the scientific community to clarify these processes for the public. By

democratizing our methods, students were empowered to build confidence in seeking to be on the cutting-edge of the latest scientific breakthroughs in genomic research.

These outcomes were only possible due to the unique academic environment fostered by Dr. Gunaratne, and the work done by the researchers at UHSeq. Her trust in the students' capabilities and her patience throughout the complex *in silico* workflow provided the necessary safety net for innovation to flourish. By giving the students complete discretion to integrate their outside lab expertise, she transformed a standard classroom exercise into a high-level research endeavor where even undergraduate contributions to established laboratory protocol could yield novel, publishable outcomes. This course demonstrated that with the acceptance of a unique approach to multi-disciplinary innovation and student engagement supported by mentors who believe in the power of collaborative discovery, the next generation of clinical researchers will be capable of performing established protocols while building upon their own resourcefulness for innovating the future of life-saving medicine.

Acknowledgements

The authors would like to thank Dr. Preethi Gunaratne and UHSeq for their patience, assistance and dedication to student success throughout the whole of the semester, for providing this innovative workflow, prerequisite knowledge, inviting us to engage with genomics research at the cutting edge, and especially for their fact-checking and proofreading of this paper. The authors would also like to thank every other participant in the course for their insightful interrogation of our proposed innovations.

References

[1] C1QBP Complement C1q Binding Protein [Homo sapiens (human)]—Gene—NCBI. **2026**. https://www.ncbi.nlm.nih.gov/gene/708

[2] Diffuse Large B-Cell Lymphoma (DLBCL) Research. *Blood Cancer United*. **2015**. https://bloodcancerunited.org/research/blood-cancer-research-development-progress/lymphoma/diffuse-large-b-cell-lymphoma-dlbcl

[3] Expression of MMP14 in Cancer—Summary. *The Human Protein Atlas*. **2021**. https://www.proteinatlas.org/ENSG00000157227-MMP14/cancer

[4] Hanahan, D. Hallmarks of Cancer: New Dimensions. *Cancer Discovery* **2022**, *12* (1), 31–46. https://doi.org/10.1158/2159-8290.cd-21-1059

[5] Hanahan, D. Hallmarks of Cancer—Then and Now, and Beyond. *Cell* **2026**. https://doi.org/10.1016/j.cell.2025.12.049

[6] HLA Gene May Predict if Cancer Immunotherapy Will Work. National Cancer Institute. **2022**, January 27. https://www.cancer.gov/news-events/cancer-currents-blog/2022/immunotherapy-cancer-biomarker-hla-gene

[7] Homo sapiens Complement C1q Binding Protein (C1QBP), mRNA; Nuclear Gene—Nucleotide—NCBI. **2026**. https://www.ncbi.nlm.nih.gov/nuccore/NM_001212.4?report=GenBank

[8] Homo sapiens Matrix Metallopeptidase 14 (MMP14), mRNA—Nucleotide—NCBI. **2026**. https://www.ncbi.nlm.nih.gov/nuccore/NM_004995.4?report=GenBank

[9] Human BLAT Search. *UCSC Genome Browser*. n.d. https://genome.ucsc.edu/cgi-bin/hgBlat

[10] Human Protein Atlas—Cancer. *The Human Protein Atlas*. **2014**. https://www.proteinatlas.org/humanproteome/cancer

[11] Liu, X.-H.; Liu, H.; Ke, D.; Ke, X.-K.; Wang, Y.-D. Clinical and Prognostic Features of Primary Retroperitoneal Diffuse Large B-Cell Lymphoma: A Single-Center Experience in China. *Chinese Medical Journal* **2021**, *134* (22), 2750–2752. https://doi.org/10.1097/cm9.0000000000001473

[12] Liu, Y.-Y.; Leboeuf, C.; Shi, J.-Y.; Li, J.-M.; Wang, L.; Shen, Y.; Garcia, J.-F.; Shen, Z.-X.; Chen, Z.; Janin, A.; Chen, S.-J.; Zhao, W.-L. Rituximab plus CHOP (R-CHOP) Overcomes PRDM1-Associated Resistance to Chemotherapy in Patients with Diffuse Large B-Cell Lymphoma. *Blood* **2007**, *110* (1), 339–344. https://doi.org/10.1182/blood-2006-09-049189

[13] Major, A.; Smith, S. M. DA-R-EPOCH vs R-CHOP in DLBCL: How Do We Choose? *Clinical Advances in Hematology & Oncology* **2021**, *19* (11), 698. https://pmc.ncbi.nlm.nih.gov/articles/PMC9036549/

[14] MMP14 Matrix Metallopeptidase 14 [Homo sapiens (human)]—Gene—NCBI. n.d. https://www.ncbi.nlm.nih.gov/gene/4323

[15] NCBI. ORFfinder—NCBI. n.d. https://www.ncbi.nlm.nih.gov/orffinder/

[16] Nordquist, H.; Jamil, R. T. Biochemistry, HLA Antigens. In *StatPearls*; StatPearls Publishing: Treasure Island, FL, **2021**. https://www.ncbi.nlm.nih.gov/books/NBK546662/

[17] Rozanov, D. V.; Ghebrehiwet, B.; Postnova, T. I.; Eichinger, A.; Deryugina, E. I.; Strongin, A. Y. The Hemopexin-Like C-Terminal Domain of Membrane Type 1 Matrix Metalloproteinase Regulates Proteolysis of a Multifunctional Protein, gC1qR. *Journal of Biological Chemistry* **2002**, *277* (11), 9318–9325. https://doi.org/10.1074/jbc.m110711200

[18] Simonson, A.; Flores, J.; Garcia, J.; Forofontov, P.; Moreno, C.; Wiggins, E.; Lopez, S.; Rojas, A.; Lopez, M.; McCormick, K.; Artis, R.; Fredrickson, A.; Lund, J.; Nelson, C.; Sturgill, I.; Jamison, A.; Pleass, R.; Mercer, G.; Papadopoulos, K.; Patnaik, A. Abstract B061: In Vivo Benchmarking of Trastuzumab Deruxtecan (T-DXd) Activity in Panel of 400 XPDX Models. *Molecular Cancer Therapeutics* **2025**, *24* (10_Supplement), B061. https://doi.org/10.1158/1535-7163.targ-25-b061

[19] Szolek, A.; Schubert, B.; Mohr, C.; Sturm, M.; Feldhahn, M.; Kohlbacher, O. OptiType: Precision HLA Typing from Next-Generation Sequencing Data. *Bioinformatics* **2014**, *30* (23), 3310–3316. https://doi.org/10.1093/bioinformatics/btu548

[20] TargetScanHuman 8.0 Predicted Targeting of Human C1QBP. **2019**. https://www.targetscan.org/cgi-bin/targetscan/vert_80/view_gene.cgi?rs=ENST00000225698.4&taxid=9606&showcnc=0&shownc=0&shownc_nc=&showncf1=&showncf2=&subset=1

[21] Venturutti, L.; Melnick, A. M. The Dangers of Déjà Vu: Memory B Cells as the Cells of Origin of ABC-DLBCLs. *Blood* **2020**, *136* (20), 2263–2274. https://doi.org/10.1182/blood.2020005857

[22] Wang, L.; Li, L.-R. R-CHOP Resistance in Diffuse Large B-Cell Lymphoma: Biological and Molecular Mechanisms. *Chinese Medical Journal* **2020**, *134* (3), 253–260. https://doi.org/10.1097/cm9.0000000000001294

[23] Wickström, S. L.; Lövgren, T.; Volkmar, M.; Reinhold, B.; Duke-Cohan, J. S.; Hartmann, L.; Rebmann, J.; Mueller, A.; Melief, J.; Maas, R.; Ligtenberg, M.; Hansson, J.; Offringa, R.; Seliger, B.; Poschke, I.; Reinherz, E. L.; Kiessling, R. Cancer Neoepitopes for Immunotherapy: Discordance between Tumor-Infiltrating T Cell Reactivity and Tumor MHC Peptidome Display. *Frontiers in Immunology* **2019**, *10*. https://doi.org/10.3389/fimmu.2019.02766

[24] Weeder, B. R.; Wood, M. A.; Li, E.; Nellore, A.; Thompson, R. F. Pepsickle Rapidly and Accurately Predicts Proteasomal Cleavage Sites for Improved Neoantigen Identification. *Bioinformatics* **2021**, *37* (21), 3723–3733. https://doi.org/10.1093/bioinformatics/btab628

[25] Sherman, B. T.; Hao, M.; Qiu, J.; Jiao, X.; Baseler, M. W.; Lane, H. C.; Imamichi, T.; Chang, W. DAVID: A Web Server for Functional Enrichment Analysis and Functional Annotation of Gene Lists (2021 Update). *Nucleic Acids Research* **2022**, *50* (W1), W216–W221. https://doi.org/10.1093/nar/gkac194

[26] Cradick, T. J.; Qiu, P.; Lee, C. M.; Fine, E. J.; Bao, G. COSMID: A Web-Based Tool for Identifying and Validating CRISPR/Cas Off-Target Sites. *Molecular Therapy: Nucleic Acids* **2014**, *3* (12), e214. https://doi.org/10.1038/mtna.2014.64

[27] Castillo, M. B.; Rankothgedera, S.; Thevasagayampillai, S.; Kandasamy, A.; Lewis, J.; Woody, C.; Vaz de Freitas, M.; Antunes, D. A.; El-Zein, R.; Gunaratne, P. H. Identification of Immunogenic KIF5B-RET Fusion Neopeptides Driving Immune Stimulation in Tumor Specific CD8+ T Cells. *Frontiers in Immunology* **2025**, *16*, 1635810. https://doi.org/10.3389/fimmu.2025.1635810

[28] Mistretta, B.; Rankothgedera, S.; Castillo, M.; Rao, M.; Holloway, K.; Bhardwaj, A.; El Noafal, M.; Albarracin, C.; El-Zein, R.; Rezaei, H.; Su, X.; Akbani, R.; Shao, X. M.; Czerniecki, B. J.; Karchin, R.; Bedrosian, I.; Gunaratne, P. H. Chimeric RNAs Reveal Putative Neoantigen Peptides for Developing Tumor Vaccines for Breast Cancer. *Frontiers in Immunology* **2023**, *14*, 1188831. https://doi.org/10.3389/fimmu.2023.1188831