Propuesta de la Investigación

1. Título

Desarrollo de una herramienta para la detección *in silico* de neoantígenos a partir de datos genómicos en el marco de desarrollo de vacunas personalizadas para tratar el cancer.

2. Problema (5k)

El cáncer representa el mayor problema de salud mundial [15]. Además, según el instituto de investigación del cáncer del Reino Unido, se ha registrado más de 18 millones de nuevos casos y 10 millones de muertes en el 2020 [16]. Más alarmante aún, se predice que habrá 28 millones de nuevos casos por año alrededor del 2040, si la incidencia se mantiene estable y el crecimiento de la población y el envejecimiento continúan de acuerdo con las tendencias recientes [17]. Esto representa un aumento del 54.9 % con respecto a 2020 y se espera que sea mayor en hombres (aumento del 60.6 %) que en mujeres (aumento del 48.8 %). A todo esto, se sabe que los métodos tradicionales basados en cirugías, radioterapias y quimioterapias tienen baja efectividad y adversos efectos secundarios [18]. En este contexto, surge el desarrollo de la inmunoterapia de cáncer, que tiene como objetivo estimular el sistema inmunológico de un paciente [19]. Existen varios tratamientos como: vacunas personalizadas; terapias de células T adoptivas; e inhibidores de puntos de control inmunológico. De estos, las vacunas basadas en **neoantígenos** han demostrado un gran potencial, al potenciar las respuestas de las células T y es considerada la de mayor probabilidad de éxito [19]. También, los neantígenos son utilizados en la terapia de bloqueo de puntos de control inmunológico. En este sentido, los neoantígenos son considerados biomarcadores predictivos y objetivos de tratamiento sinérgico en la inmunoterapia del cáncer [20].

A pesar de varios esfuerzos en el desarrollo de *pipelines* y algoritmos, menos del 5 % de neoantígenos detectados activan el sistema immune [1–5]. Según los autores de los *pipelines* las razones son:

1. La no inclusión en conjunto de varias fuentes de información como DNA-seq, RNA-seq, y datos de *Mass Spectrometry* (MS) [6]. Por ejemplo, la mayoria de propuestas no utiliza datos de MS; en la actualidad, existe una creciente información de estos datos y se estan aplicado a varios campos de la Bioinformática.

2. Uso herramientas de bajo desempeño para la predicción del enlace péptido-MHC (pMHC). La mayoria de aplicaciones, se basa en el uso de MHCFlurry [7] y NetMHCpan4.1 [8]. Sin embargo, actualmente, se cuenta con herramientas de mejor desempeño como: MixMHCpred [9], Anthem [10], Acme [11] y ESM-GAT [12].

3. Para la etapa 3.2 de la Figura 1, los autores no consideran la predicción del enlace pMHC al TCR (pMHC-TCR) , varios autores consideran incluir esta tarea en trabajos futuros [13].

4. Finalmente y quizas la mas importante es no utilizar información de eventos de *alternative splicing*, variaciones estructurales en el ADN y las mutaciones de fusión de genes, está información esta fuertemente relacionada con varios tipos de cancer [14].

3. Estado del arte o antecedentes (10k)

Falta incluir ¿Cual es el aporte cientifico del proyecto al area tematica de acuerdo a la sub area o disciplina a la que cooresponda su propuesta en comparacion con otros estudios de la literatura publicada?

El desarrollo de vacunas personalizadas contra el cáncer es un proceso largo y depende de la correcta detección de neoantígenos (ver Figura 1). Estos neoantígenos son péptidos que solo están presentes en las células cancerosas. De esta forma, el objetivo de un tratamiento basado en vacunas personalizadas, es entrenar a los linfocitos del paciente (células T) para reconocer los neoantígenos y activar el sistema inmunológico [1, 18]. El proceso se resume en la Figura 1 y consiste en:

1. Obtener muestras de tejido canceroso y saludable, Luego se secuencia ambos tejidos para obtener el ADN y/o ARN. Algunas propuestas incluyen información inmunopeptidoma de *Mass Spectrometry* (MS).

2. Etapa *in-silico*, aquí realiza alineamiento de secuencias, se desarrolla un *llamado de variantes* para detectar las variantes y/o mutaciones; y se anotan dichas variantes (detección de posibles neoantígenos). Esta etapa cuenta con varias herramientas con buen desempeño.

3. En esta etapa *in-silico* se priorizan neoantígenos. Esta etapa es crucial y ha tenido bastante investigación los últimos años debido a su complejidad y la baja efectividad de propuestas actuales. Aquí, se toman los neoantígenos candidatos (péptidos) de la etapa anterior y se predice su afinidad con el *Major Histocompatibility Complex* (MHC), este problema se conoce como *pMHC binding*. Luego, se evalúa la afinidad del pMHC para enlazarse al T-cell Receptor (TCR). Al finalizar esta etapa, se obtienen los neoantígenos.

4. En esta etapa *in-vitro*, se induce en laboratorio a las células T del paciente a reconocer los neoantígenos. Aquí, se desarrollan las vacunas. Generalmente, esta etapa es desarrollada por biotecnólogos y biólogos.

5. Finalmente, el médico oncólogo realiza la evaluación clínica de la vacuna.

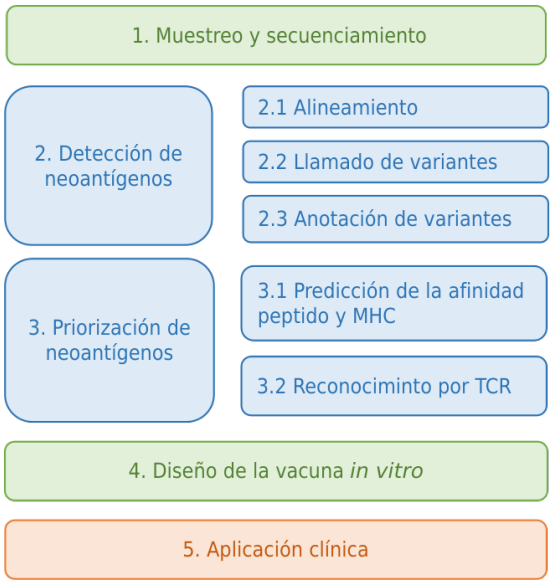

Figura 1: Marco de desarrollo para la elaboración de vacunas personalizadas contra el cáncer basadas en neoantígenos. Se detalla cada fase enfatizando el desarrollo *in-silico*. Fuente: Elaboración propia.

La detección *in-silico* de neoantígenos se basa en la segunda y tercera etapa de la Figura 1. En este contexto, debido a la complejidad del proceso y la cantidad de métodos existentes, se han desarrollado software y *pipelines* para facilitar el uso de estas herramientas. En la Tabla 1, presentamos los *pipelines* publicados a partir del 2018. Estos *pipelines* utilizan diferentes tipos de información como entrada, así PGV Pipeline [13] y PEPPRMINT [22] utilizan DNA-seq; sin embargo, otras herramientas como PGNneo [23], NAP-CNB [24], NaoANT-HILL [25], ProGeo-neo [26], ScanNeo [27] y Neopepse [6] utilizan RNA-seq porque estas secuencias encapsulan mejor la información de mutaciones y *non-coding regions* de ADN [23].

Con el objetivo de reducir la complejidad de los *pipelines*, otras propuestas han optado por utilizar Variant Calling Format (VCF), como entrada. Estos archivos, contienen información de las mutaciones y son obtenidas a partir de métodos de alineamiento y llamado de mutaciones (etapas 2.1 y 2.2 de la Figura 1). De esta forma, herramientas como Valid-Neo [28], HLA3D [29], Neoepiscope [14] , pVACtools [30] y NeoPredPipe [31], reducen la cantidad de herramientas utilizadas en la deteccion de neoantígenos; sin embargo, los resultados obtenidos, pueden ser inferiores comparado con herramientas que usan DNA-seq y RNA-seq.

Adicionalmente, para una correcta detección de neoantígenos, es necesario contar con la secuenciación de proteínas Major Histocompatibility Complex (MHC) o Human Leukocyte Antigens (HLA). Es necesario contar con estas proteínas porque, son utilizadas para predecir la unión entre posibles neoantígenos al MHC (pMHC: etapa 3.1 de la Figura 1). Estas proteínas son codificadas por genes altamente polimórficos, esto proporciona una variación sustancial en la unión de péptidos (neoantígenos), influyendo de esta manera en el conjunto de péptidos presentados a las células T. [32]. En este contexto, los *pipelines* Valid-NEO [28] y NeoPredPipe [31] y Neopepsee [6] solicitan como entrada estas proteínas (HLA); mientras que las otras predicen esta información a partir de DNA-seq. Desde un punto de vista de usabilidad, obtener los tipos de HLA, implica un esfuerzo inncesario para el usuario.

Es así que se propone el desarrollo de un *pipeline*, que incorpora el uso de MS y fusión de genes. Además, la propuesta incluirá el desarrollo de un modelo *transformers* para la predicción del enlace pMHC y pMHC-TCR. Cabe resaltar, que este proyecto es una continuación del proyecto "Principales estrategias y métodos basados en deep learning para la detección de neo antígenos en el marco del desarrollo de vacunas personalizadas en la inmunoterapia del cáncer" financiado por ULaSalle-UCSP, donde se obtuvo dos publicaciones [33, 34].

Tabla 1: Lista de *pipelines* desarrollados desde el 2018 hasta la actualidad para la detección de neoantígenos.GN: Expresión de genes, VA: anotación de variantes.

Nombre	**Año**	**Ref.**	**Entrada**	**Salida**
PEPPRMINT	2023	[22]	DNA-seq	Neoantígenos
PGNneo	2023	[23]	VCF, RNA-seq y MS data	Neoantígenos
Valid-NEO	2022	[28]	VCF y HLA	Neoantígenos
HLA3D	2022	[29]	VCF, HLA, SMG y HBV	Neoantígenos
NAP-CNB	2021	[24]	RNA-seq	Neoantígenos
NeoANT-HILL	2020	[25]	RNA-seq y VCF	Neoantígenos,GE
Neoepiscope	2020	[14]	VCF y BAM	Neoantígenos
ProGeo-neo	2020	[26]	RNA-seq y VCF	Neoantígenos
pVACtools	2020	[30]	VCF	Neoantígenos
NeoPredPipe	2019	[31]	VCF y HLA	Neoantígenos,VA
ScanNeo	2019	[27]	RNA-seq	Neoantígenos
Neopepsee	2018	[6]	RNA-seq, VCF, HLA	Neoantígenos,GE
PGV Pipeline	2018	[13]	DNA-seq	Neoantígenos

4. Resultados o avances previos (4k)

La propuesta del proyecto representa la continuación de una serie de iniciativas y publicaciones. Se inició con el PROYECTO 01: "Principales estrategias y métodos basados en deep learning para la detección de neoantígenos en el marco del desarrollo de vacunas personalizadas en la inmunoterapia del cáncer", financiado por las Universidades La Salle y Católica San Pablo. Este proyecto resultó en dos publicaciones: "Deep Learning and Transformers in MHC-Peptide Binding and Presentation Towards Personalized Vaccines in Cancer Immunology: A Brief Review" [33] y "Neoantigen Detection Using Transformers and Transfer Learning in the Cancer Immunology Context" [34].

Recientemente, hemos completado la ejecución del PROYECTO 02: "Desarrollo de una Aplicación Web para la Detección de Neoantígenos en el Marco de Desarrollo de Vacunas Personalizadas para Tratar el Cáncer". En este proyecto, hemos creado una aplicación centrada en la detección de neoantígenos, con un enfoque en la predicción del enlace pMHC mediante el uso de modelos Transformers y Transfer Learning. Además, hemos enviado para revisión el artículo "Fine-tuning Transformers for Peptide-MHC Class I Binding Prediction". Además, esta propuesta de PROCIENCIA también representa el trabajo futuro de la tesis de doctorado en Ciencia de la Computación del investigador principal, Vicente Machaca Arceda, titulada: "Detección *in Silico* de Neoantígenos Utilizando Transformers y Transfer Learning en el Marco de Desarrollo de Vacunas Personalizadas para Tratar el Cáncer". La tesis aborda los mismos objetivos que el PROYECTO 02 y ha logrado la aceptación del artículo "Transformers Meets Neoantigen Detection: A Systematic Literature Review.[en] el Journal of Integrative Bioinformatics (Q2).

Actualmente, estamos llevando a cabo el PROYECTO 03 "NeoArgos-tools: Un Pipeline de Detección In-silico de Neoantígenos de Cáncer para el Desarrollo de Vacunas Personalizadas", con fecha de finalización en junio de este año y financiamiento de las Universidades La Salle y UCSP. Este proyecto implica el desarrollo de la versión inicial de NeoArgos-tools (la postulación a PROCIENCIA corresponde a la versión 2). En la primera versión, utilizamos archivos Variant Calling File (VCF) como entrada y mejoramos el módulo de predicción del enlace pMHC según investigaciones previas de proyectos anteriores.

Es relevante destacar que en la versión 2 de NeoArgos-tools, las mejoras propuestas incluyen el uso de datos genómicos como entrada al pipeline (RNS-seq y DNA-seq), la incorporación de Mass Spectrometry (MS) para refinar la

detección de neoantígenos, la inclusión de información sobre variantes estructurales y fusión de genes, y el desarrollo de una interfaz gráfica.

5. Justificación (4k)

El cáncer constituye el principal desafío de salud a nivel global; no obstante, las técnicas convencionales basadas en cirugías, radioterapias y quimioterapias presentan una eficacia limitada [18]. En este escenario, los neoantígenos emergen como elementos cruciales en la concepción de vacunas contra el cáncer [19, 35, 36]. Si se logra desarrollar un enfoque altamente efectivo, la inmunoterapia del cáncer, fundamentada en la creación de vacunas personalizadas, podría posicionarse como una alternativa a procedimientos más tradicionales, como radioterapias y quimioterapias.

En el proyecto se propone realizar dos contribuciones significativas: CONTRIBUCIÓN 01: En el ámbito de la ciencia de la computación, se llevará a cabo el desarrollo de un modelo basado en *Transformers* y Transfer Learning para la predicción del enlace pMHC, con resultados previos que demuestran superar a otras propuestas en el estado del arte. CONTRIBUCIÓN 02: En el ámbito de la Bioinformática, la implementación del *pipeline* plantea un desafío al abordar problemas de integración, alto costo computacional, heterogeneidad y modularidad. Además, este *pipeline* incorporará datos de *Mass Spectrometry* (MS) y fusión de genes con el objetivo de obtener resultados más sobresalientes que otros métodos presentes en el estado del arte.

Con la conclusión de este proyecto, se abrirá paso a la segunda fase, que implica una colaboración interdisciplinaria para llevar a cabo el desarrollo *in vitro* de las vacunas de neoantígenos. En una tercera etapa, se contemplarán pruebas clínicas, marcando así un avance integral en la búsqueda de soluciones efectivas en la lucha contra el cáncer.

6. Porqué es una investigación básica (3k)

7. Pregunta de investigación

¿El desarrollo de una herramienta in silico para la detección de neonatígenos a partir de datos genómicos es capaz de detectar correctmente geoantígenos de cancer?

8. Objetivos de la investigación

8.1. Objetivo general

Desarrollar una herramienta para la detección *in silico* de neoantígenos de cáncer a partir de datos genómicos.

8.2. Objetivos específicos

1. Analizar qué fuentes de información o datos de entrada recibirá la herramienta. Se evaluará DNA-seq, RNA-seq, *Variant Calling File* (VCF) y como integrar esta información con datos de *Mass Spectrometry* (MS).
2. Analizar qué herramientas se utilizarán para la primera etapa de la herramienta, referente al alineamiento de secuencias, llamado de variantes y anotación de variantes (predicción de posibles neoantígenos).
3. Analizar el uso de información de variaciones estructurales del ADN y mutaciones de fusión de genes. Se evaluará el desempeño de *Arriba* [37] y FusionQ [38].
4. Evaluar métodos de detección de eventos de *alternative splicing* y analizar la aplicación de estos al integrarse a la herramienta para la detección de neoantígenos.
5. Implementar un modelo basado en *transformers* para la predicción del enlace de los neoantígenos al MHC (pMHC). Ya se cuenta con resultados previos de una propuesta que es superior a otras del estado del arte [34].
6. Implementar un modelo basado en *transformers* para la predicción del enlace pMHC al TCR (pMHC-TCR).
7. Implementar una interfaz gráfica que sirva como panel de administración para configurar y escoger las herramientas en cada fase del *pipeline*.
8. Comparar el desempeño de la herramienta propuesta con otras herramientas del estado del arte.

9. Metodología (5k)

FALTA DETALAR LA METODOLIGA POR CADA OBJETIVO ESPECÍFICO

Hemos dividido la propuesta en dos módulos: NeoArgosMut y NeoArgosAntigen. NeoArgosMut, se enfoca en el llamado y anotación de variantes, obteniéndose como salida neoantígenos candidatos. Luego, NeoArgosAntigen, prioriza estos antígenos, al predecir su afinidad al MHC (pMHC) y luego la afinidad del pMHC al TCR (pMHC-TCR). En la Figura 2, mostramos estos módulos.

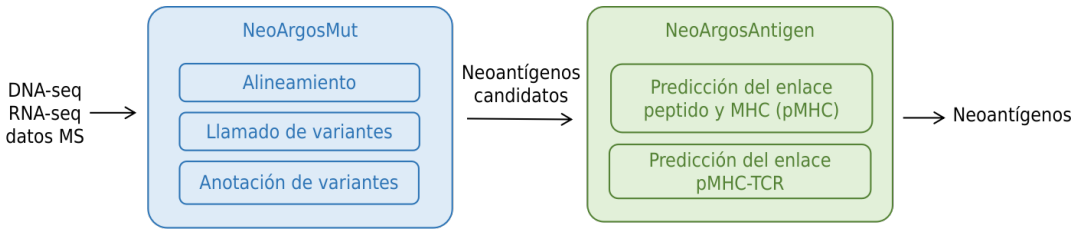

Figura 2: Representación de NeoArgosMut y NeoArgosAntigen para la detección de neoantígenos

9.1. NeoArgosMut

NeoArgosMut, se encarga de recibir como entrada datos de DNA-seq, RNA-seq y Mass Spectrometry (MS). Luego se plantea alinear dichas secuencias con uso de las herramientas como BWA-MEM y Bowtie2. Adicionalmente, se usará STAR, porque alinea mejor muestras tumorales [13]. Como salida a esta etapa, se optiene archivos de alineamiento BAM.

Para el llamado de variantes se utilizará MuTect y Strelka. Luego, se utilizará la unión de la información de ambos métodos tal como lo hizo [22] y [13]. Como salida, se obtienen archivos VCF. Adicionalmente a otros *pipelines*, utilizaremos información sobre la fusión de genes que se obtendrán de las herramientas *Arriba* [37] y FusionQ [38]. Esta forma parte de la contribución de este trabajo, porque se sabe que la mayoría de *pipelines* tienen un bajo desempeño debido ausencia de información en su procesos de variantes estructurales y fusión de genes [14]. Finalmente, también se va a utilizar MaxQuant para identificar las mutaciones a nivel de péptidos con ayuda de información de Mass Spectrometry (MS), esto también forma parte de la contribución del trabajo al incluir fuentes adicionales de información como MS.

Luego corresponde a la anotación de variantes, en esta etapa se toman los archivos en formato VCF y se obtienen los péptidos generados a partir de estas variaciones o mutaciones. Estos péptidos representan los posibles neoantígenos. Para está tarea se va a utilizar Isovar y ANNOVAR.

Finalmente, para obtener el tipo de HLA del paciente se va a utilizar la herramienta OptiType. Otros *pipelines* optan por solicitar al usuario la información del tipo de HLA; sin embargo, obtener el HLA a partir de las mismas secuencias de ADN, mejora considerablemente el desempeño general del pipeline y la accesibilidad del usuario. Al finalizar esta etapa, se va a obtener los neoantígenos candidatos.

9.2. NeoArgosAntigen

NeoArgosAntigen, prioriza los neoantígenos detectados previamente por NeoArgosMut. Esta priroización la realiza en base a la predicción del enlace de los neoantígenos al MHC y posteriormete al TCR. El módulo se divide en dos partes: la predicción del enlace pMHC y la afinidad del pMHC al TCR. Ambas toman como entrada dos secuencias de proteínas, luego se necesita predecir su afinidad (regresión) o el enlace (clasificación). En resumen, las proteínas se pueden representar como $p = \{A, ..., Q\}$ y $q = \{A, N, ..., Q, E, G\}$. Luego, tenemos que predecir la probabilidad del enlace o afinidad entre p y q.

Para el problema de predicción del enlace pMHC se va a utilizar modelos BERT pre-entrenados y se realizará *fine-tuning* agregando un bloque de capas BiLSTM. Luego se volverá a entrenar estos modelos con una base de datos compuesta por muestras de [39] y [9]. Se propone la arquitectura de la Figura 3. Como se puede ver, la entrada son dos secuencias de proteínas: el péptido y el MHC. Luego, el modelo basado en transformers está compuesto por un modelo pre-entrenado y un bloque de capas BiLSTM, esta propuesta se basó en el trabajo de [39]. En esta etapa también. se va a evaluar el desempeño de varios modelos BERT pre-entrenados como: TAPE [40], ProtBERT-BFD [41] y ESM2 [42] cada una con 92 millones, 420 millones, 650 millones parámetros respectivamente. Adicionalmente, TAPE fue entrenado con 30 millones de proteínas, ProtBERT-BFD con 2122 millones de proteínas y 60 millones de proteínas

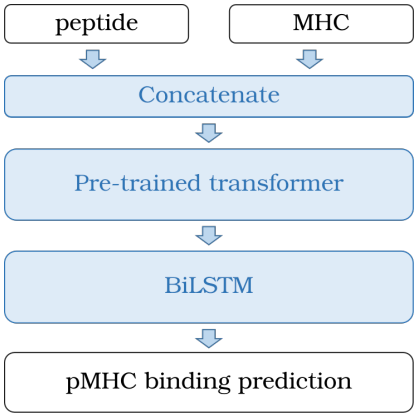

Figura 3: Modelo *transformer* seguido de BiLSTM para predecir el enlace pMHC.

para ESM-2. En base a trabajos anteriores propios, sabemos que el uso de TAPE y el modelo más pequeño de ESM2 tienen buenos resultados [34].

La misma arquitectura de la Figura 3, se utilizará para la predicción del enlace pMHC y TCR (pMHC-TCR) según recomendaciones de [26] y [43]. Sin embargo, se va reentrenar el modelo para adaptarse a este nuevo problema, se utilizarán muestras de [26] y la base de datos de VDJdb [44]. Al finalizar esta etapa, se obtendrán los neonatígenos.

10. Limitaciones (5K)

lIMITACIONES Y ESTRATEGIAS ABORADADAS

11. Resumen (5K)

12. Describa el equipamiento e infraestructura disponible (5k)

13. Resultados esperados

14. Cronograma de Actividades

15. Sostenibilidad de la propuesta (5k)

16. Impacto Científico de la propuesta (5k)

Referencias

[1] L Mattos, M Vazquez, F Finotello, R Lepore, E Porta, J Hundal, P Amengual-Rigo, CKY Ng, A Valencia, J Carrillo, et al., "Neoantigen prediction and computational perspectives towards clinical benefit: recommendations from the esmo precision medicine working group," *Annals of oncology*, vol. 31, no. 8, pp. 978–990, 2020.

[2] Nil Adell Mill, Cedric Bogaert, Wim van Criekinge, and Bruno Fant, "neoms: Attention-based prediction of mhc-i epitope presentation," *bioRxiv*, 2022.

[3] Brendan Bulik-Sullivan, Jennifer Busby, Christine D Palmer, Matthew J Davis, Tyler Murphy, Andrew Clark, Michele Busby, Fujiko Duke, Aaron Yang, Lauren Young, et al., "Deep learning using tumor hla peptide mass spectrometry datasets improves neoantigen identification," *Nature biotechnology*, vol. 37, no. 1, pp. 55–63, 2019.

[4] Michal Bassani-Sternberg, Sune Pletscher-Frankild, Lars Juhl Jensen, and Matthias Mann, "Mass spectrometry of human leukocyte antigen class i peptidomes reveals strong effects of protein abundance and turnover on antigen presentation*[s]," *Molecular & Cellular Proteomics*, vol. 14, no. 3, pp. 658–673, 2015.

[5] Mahesh Yadav, Suchit Jhunjhunwala, Qui T Phung, Patrick Lupardus, Joshua Tanguay, Stephanie Bumbaca, Christian Franci, Tommy K Cheung, Jens Fritsche, Toni Weinschenk, et al., "Predicting immunogenic tumour mutations by combining mass spectrometry and exome sequencing," *Nature*, vol. 515, no. 7528, pp. 572–576, 2014.

[6] Sora Kim, Han Sang Kim, Eunyoung Kim, MG Lee, E-C Shin, and S Paik, "Neopepsee: accurate genome-level prediction of neoantigens by harnessing sequence and amino acid immunogenicity information," *Annals of Oncology*, vol. 29, no. 4, pp. 1030–1036, 2018.

[7] Timothy J O'Donnell, Alex Rubinsteyn, and Uri Laserson, "Mhcflurry 2.0: improved pan-allele prediction of mhc class i-presented peptides by incorporating antigen processing," *Cell systems*, vol. 11, no. 1, pp. 42–48, 2020.

[8] Birkir Reynisson, Bruno Alvarez, Sinu Paul, Bjoern Peters, and Morten Nielsen, "Netmhcpan-4.1 and netmhciipan-4.0: improved predictions of mhc antigen presentation by concurrent motif deconvolution and integration of ms mhc eluted ligand data," *Nucleic acids research*, vol. 48, no. W1, pp. W449–W454, 2020.

[9] David Gfeller, Julien Schmidt, Giancarlo Croce, Philippe Guillaume, Sara Bobisse, Raphael Genolet, Lise Queiroz, Julien Cesbron, Julien Racle, and Alexandre Harari, "Improved predictions of antigen presentation and tcr recognition with mixmhcpred2. 2 and prime2. 0 reveal potent sars-cov-2 cd8+ t-cell epitopes," *Cell Systems*, vol. 14, no. 1, pp. 72–83, 2023.

[10] Shutao Mei, Fuyi Li, Dongxu Xiang, Rochelle Ayala, Pouya Faridi, Geoffrey I Webb, Patricia T Illing, Jamie Rossjohn, Tatsuya Akutsu, Nathan P Croft, et al., "Anthem: a user customised tool for fast and accurate prediction of binding between peptides and hla class i molecules," *Briefings in Bioinformatics*, vol. 22, no. 5, pp. bbaa415, 2021.

[11] Yan Hu, Ziqiang Wang, Hailin Hu, Fangping Wan, Lin Chen, Yuanpeng Xiong, Xiaoxia Wang, Dan Zhao, Weiren Huang, and Jianyang Zeng, "Acme: pan-specific peptide–mhc class i binding prediction through attention-based deep neural networks," *Bioinformatics*, vol. 35, no. 23, pp. 4946–4954, 2019.

[12] Nasser Hashemi, Boran Hao, Mikhail Ignatov, Ioannis Ch Paschalidis, Pirooz Vakili, Sandor Vajda, and Dima Kozakov, "Improved prediction of mhc-peptide binding using protein language models," *Frontiers in Bioinformatics*, vol. 3, 2023.

[13] Alex Rubinsteyn, Julia Kodysh, Isaac Hodes, Sebastien Mondet, Bulent Arman Aksoy, John P Finnigan, Nina Bhardwaj, and Jeffrey Hammerbacher, "Computational pipeline for the pgv-001 neoantigen vaccine trial," *Frontiers in immunology*, vol. 8, pp. 1807, 2018.

[14] Mary A Wood, Austin Nguyen, Adam J Struck, Kyle Ellrott, Abhinav Nellore, and Reid F Thompson, "Neoepiscope improves neoepitope prediction with multivariant phasing," *Bioinformatics*, vol. 36, no. 3, pp. 713–720, 2020.

[15] Rebecca L Siegel, Kimberly D Miller, Nikita Sandeep Wagle, and Ahmedin Jemal, "Cancer statistics, 2023," *Ca Cancer J Clin*, vol. 73, no. 1, pp. 17–48, 2023.

[16] Cancer Research UK, "Worldwide cancer statistics," 2023.

[17] Cancer Research UK, "Worldwide cancer incidence statistics," 2023.

[18] Miao Peng, Yongzhen Mo, Yian Wang, Pan Wu, Yijie Zhang, Fang Xiong, Can Guo, Xu Wu, Yong Li, Xiaoling Li, et al., "Neoantigen vaccine: an emerging tumor immunotherapy," *Molecular cancer*, vol. 18, no. 1, pp. 1–14, 2019.

[19] Elizabeth S Borden, Kenneth H Buetow, Melissa A Wilson, and Karen Taraszka Hastings, "Cancer neoantigens: Challenges and future directions for prediction, prioritization, and validation," *Frontiers in Oncology*, vol. 12, 2022.

[20] Xianzhu Fang, Zhiliang Guo, Jinqing Liang, Jiao Wen, Yuanyuan Liu, Xiumei Guan, and Hong Li, "Neoantigens and their potential applications in tumor immunotherapy," *Oncology Letters*, vol. 23, no. 3, pp. 1–9, 2022.

[21] Xue-Jiao Han, Xue-lei Ma, Li Yang, Yu-quan Wei, Yong Peng, and Xia-wei Wei, "Progress in neoantigen targeted cancer immunotherapies," *Frontiers in Cell and Developmental Biology*, vol. 8, pp. 728, 2020.

[22] Laura Y Zhou, Fei Zou, and Wei Sun, "Prioritizing candidate peptides for cancer vaccines through predicting peptide presentation by hla-i proteins," *Biometrics*, vol. 79, no. 3, pp. 2664–2676, 2023.

[23] Xiaoxiu Tan, Linfeng Xu, Xingxing Jian, Jian Ouyang, Bo Hu, Xinrong Yang, Tao Wang, and Lu Xie, "Pgnneo: A proteogenomics-based neoantigen prediction pipeline in noncoding regions," *Cells*, vol. 12, no. 5, pp. 782, 2023.

[24] Carlos Wert-Carvajal, Rubén Sánchez-García, José R Macías, Rebeca Sanz-Pamplona, Almudena Méndez Pérez, Ramon Alemany, Esteban Veiga, Carlos Óscar S Sorzano, and Arrate Muñoz-Barrutia, "Predicting mhc i restricted t cell epitopes in mice with nap-cnb, a novel online tool," *Scientific reports*, vol. 11, no. 1, pp. 1–10, 2021.

[25] Ana Carolina MF Coelho, André L Fonseca, Danilo L Martins, Paulo BR Lins, Lucas M da Cunha, and Sandro J de Souza, "neoant-hill: an integrated tool for identification of potential neoantigens," *BMC Medical Genomics*, vol. 13, no. 1, pp. 1–8, 2020.

[26] Yuyu Li, Guangzhi Wang, Xiaoxiu Tan, Jian Ouyang, Menghuan Zhang, Xiaofeng Song, Qi Liu, Qibin Leng, Lanming Chen, and Lu Xie, "Progeo-neo: a customized proteogenomic workflow for neoantigen prediction and selection," *BMC medical genomics*, vol. 13, no. 5, pp. 1–11, 2020.

[27] Ting-You Wang, Li Wang, Sk Kayum Alam, Luke H Hoeppner, and Rendong Yang, "Scanneo: identifying indel-derived neoantigens using rna-seq data," *Bioinformatics*, vol. 35, no. 20, pp. 4159–4161, 2019.

[28] Yuri Laguna Terai, Chun Huang, Baoli Wang, Xiaonan Kang, Jing Han, Jacqueline Douglass, Emily Han-Chung Hsiue, Ming Zhang, Raj Purohit, Taylor deSilva, et al., "Valid-neo: A multi-omics platform for neoantigen detection and quantification from limited clinical samples," *Cancers*, vol. 14, no. 5, pp. 1243, 2022.

[29] Xingyu Li, Xue Lin, Xueyin Mei, Pin Chen, Anna Liu, Weicheng Liang, Shan Chang, and Jian Li, "Hla3d: an integrated structure-based computational toolkit for immunotherapy," *Briefings in bioinformatics*, vol. 23, no. 3, pp. bbac076, 2022.

[30] Jasreet Hundal, Susanna Kiwala, Joshua McMichael, Christopher A Miller, Huiming Xia, Alexander T Wollam, Connor J Liu, Sidi Zhao, Yang-Yang Feng, Aaron P Graubert, et al., "pvactools: a computational toolkit to identify and visualize cancer neoantigens," *Cancer immunology research*, vol. 8, no. 3, pp. 409–420, 2020.

[31] Ryan O Schenck, Eszter Lakatos, Chandler Gatenbee, Trevor A Graham, and Alexander RA @miscNCIdictionary2022, author = NCI, title = National Cancer Institute Dictionary, year = 2022, url = https://www.cancer.gov/publications/dictionaries/genetics-dictionary, urldate = 2022-03-20 Anderson, "Neopredpipe: high-throughput neoantigen prediction and recognition potential pipeline," *BMC bioinformatics*, vol. 20, no. 1, pp. 1–6, 2019.

[32] Esam T Abualrous, Jana Sticht, and Christian Freund, "Major histocompatibility complex (mhc) class i and class ii proteins: impact of polymorphism on antigen presentation," *Current Opinion in Immunology*, vol. 70, pp. 95–104, 2021.

[33] Vicente Enrique Machaca, Valeria Goyzueta, Maria Cruz, and Yvan Tupac, "Deep learning and transformers in mhc-peptide binding and presentation towards personalized vaccines in cancer immunology: A brief review," in *International Conference on Practical Applications of Computational Biology & Bioinformatics*. Springer, 2023, pp. 14–23.

[34] Vicente Enrique Machaca Arceda, "Neoantigen detection using transformers and transfer learning in the cancer immunology context," in *International Conference on Practical Applications of Computational Biology & Bioinformatics*. Springer, 2023, pp. 97–102.

[35] Ina Chen, Michael Chen, Peter Goedegebuure, and William Gillanders, "Challenges targeting cancer neoantigens in 2021: a systematic literature review," *Expert Review of Vaccines*, vol. 20, no. 7, pp. 827–837, 2021.

[36] Alexander V Gopanenko, Ekaterina N Kosobokova, and Vyacheslav S Kosorukov, "Main strategies for the identification of neoantigens," *Cancers*, vol. 12, no. 10, pp. 2879, 2020.

[37] Sebastian Uhrig, Julia Ellermann, Tatjana Walther, Pauline Burkhardt, Martina Fröhlich, Barbara Hutter, Umut H Toprak, Olaf Neumann, Albrecht Stenzinger, Claudia Scholl, et al., "Accurate and efficient detection of gene fusions from rna sequencing data," *Genome research*, vol. 31, no. 3, pp. 448–460, 2021.

[38] Chenglin Liu, Jinwen Ma, ChungChe Jeff Chang, and Xiaobo Zhou, "Fusionq: a novel approach for gene fusion detection and quantification from paired-end rna-seq," *BMC bioinformatics*, vol. 14, no. 1, pp. 1–11, 2013.

[39] Yaqi Zhang, Gancheng Zhu, Kewei Li, Fei Li, Lan Huang, Meiyu Duan, and Fengfeng Zhou, "Hlab: learning the bilstm features from the protbert-encoded proteins for the class i hla-peptide binding prediction," *Briefings in Bioinformatics*, 2022.

[40] Roshan Rao, Nicholas Bhattacharya, Neil Thomas, Yan Duan, Peter Chen, John Canny, Pieter Abbeel, and Yun Song, "Evaluating protein transfer learning with tape," *Advances in neural information processing systems*, vol. 32, 2019.

[41] Ahmed Elnaggar, Michael Heinzinger, Christian Dallago, Ghalia Rehawi, Yu Wang, Llion Jones, Tom Gibbs, Tamas Feher, Christoph Angerer, Martin Steinegger, et al., "Prottrans: Toward understanding the language of life through self-supervised learning," *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 10, pp. 7112–7127, 2021.

[42] Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, et al., "Evolutionary-scale prediction of atomic-level protein structure with a language model," *Science*, vol. 379, no. 6637, pp. 1123–1130, 2023.

[43] Alexander Myronov, Giovanni Mazzocco, Paulina Krol, and Dariusz Plewczynski, "Bertrand-peptide: Tcr binding prediction using bidirectional encoder representations from transformers augmented with random tcr pairing," *bioRxiv*, pp. 2023–06, 2023.

[44] Mikhail Shugay, Dmitriy V Bagaev, Ivan V Zvyagin, Renske M Vroomans, Jeremy Chase Crawford, Garry Dolton, Ekaterina A Komech, Anastasiya L Sycheva, Anna E Koneva, Evgeniy S Egorov, et al., "Vdjdb: a curated database of t-cell receptor sequences with known antigen specificity," *Nucleic acids research*, vol. 46, no. D1, pp. D419–D427, 2018.