

Propuesta de la Investigación

1. Título

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

2. Problema

A pesar de varios esfuerzos en el desarrollo de *pipelines* y algoritmos, menos del 5 % de neoantígenos detectados activan el sistema immune (Mattos et al., 2020; Mill et al., 2022; Bulik-Sullivan et al., 2019; Bassani-Sternberg et al., 2015; Yadav et al., 2014). 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) (Kim et al., 2018). Por ejemplo, la mayoría de propuestas no utiliza datos de MS; en la actualidad, existe una creciente información de estos datos y se están aplicando a varios campos de la Bioinformática.
2. Uso de herramientas de bajo desempeño para la predicción del enlace péptido-MHC (pMHC) (etapa 3.1 de la Figura 1b). La mayoría de aplicaciones, se basa en el uso de MHCFlurry (O'Donnell et al., 2020) y NetMHCpan4.1 (Reynisson et al., 2020). Sin embargo, actualmente, se cuenta con herramientas de mejor desempeño basadas en *transformers* (Arceda, 2023).
3. Para la etapa 3.2 de la Figura 1b, los autores no consideran la predicción del enlace pMHC al TCR (pMHC-TCR), varios autores consideran incluir esta tarea en trabajos futuros (Rubinsteyn et al., 2018).
4. Finalmente y quizás la más importante es no utilizar información de eventos de *alternative splicing*, variaciones estructurales en el ADN y las mutaciones de fusión de genes, esta información está fuertemente relacionada con varios tipos de cancer (Wood et al., 2020).

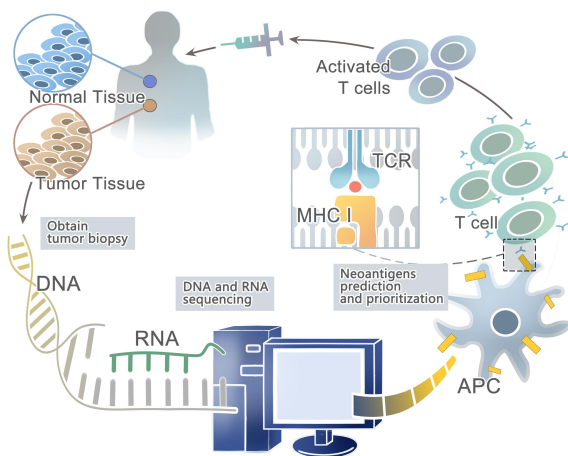
3. Estado del arte o antecedentes

El cáncer representa el mayor problema de salud mundial (Siegel et al., 2023). 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 (UK, 2023b). 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 (UK, 2023a). 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 (Peng et al., 2019). En este contexto, surge el desarrollo de la inmunoterapia de cáncer, que tiene como objetivo estimular el sistema inmunológico de un paciente (Borden et al., 2022). 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 (Borden et al., 2022). También, los neoantígenos 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 (Fang et al., 2022).

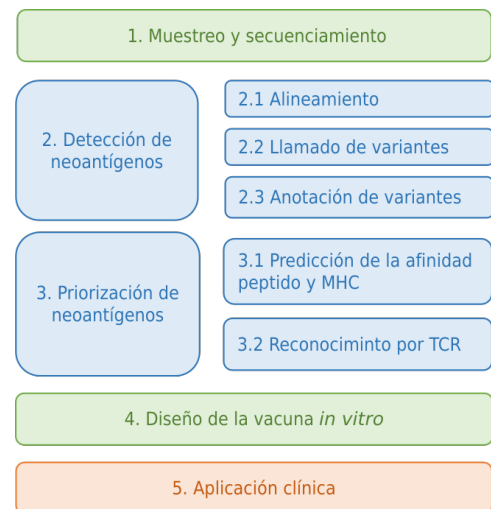
El desarrollo de vacunas personalizadas contra el cáncer es un proceso largo y depende de la correcta detección de neoantígenos (ver Figura 1). Estos neoantígenos son péptidos que solo están presentes en las células cancerosas. De esta forma, el objetivo de un tratamiento basado en vacunas personalizadas, es entrenar a los linfocitos del paciente (células T) para reconocer los neoantígenos y activar el sistema inmunológico (Mattos et al., 2020; Peng et al., 2019). El proceso se resume en la Figura 1b y consiste en:

1. Obtener muestras de tejido canceroso y saludable, Luego se secuencian 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.



(a) Proceso de desarrollo de vacunas.



(b) Pipeline para el desarrollo de vacunas.

Figura 1: Marco de desarrollo para la elaboración de vacunas personalizadas contra el cáncer basadas en neoantígenos. (a) muestra un panorama general de cada etapa. (b) detalla cada fase enfatizando el desarrollo *in-silico*. Modificado de Han et al. (2020).

La detección *in-silico* de neoantígenos se basa en la segunda y tercera etapa de la Figura 1b. 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 (Rubinsteyn et al., 2018) y PEPPRIMINT (Zhou et al., 2023) utilizan DNA-seq; sin embargo, otras herramientas como PGNNeo (Tan et al., 2023), NAP-CNB (Wert-Carvajal et al., 2021), NaoANT-HILL (Coelho et al., 2020), ProGeo-neo (Li et al., 2020), ScanNeo (Wang et al., 2019) y Neopepse (Kim et al., 2018) utilizan RNA-seq porque estas secuencias encapsulan mejor la información de mutaciones y *non-coding regions* de ADN (Tan et al., 2023).

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 1b). De esta forma, herramientas como Valid-Neo (Terai et al., 2022), HLA3D (Li et al., 2022), Neoepiscopes (Wood et al., 2020), pVACtools (Hundal et al., 2020) y NeoPredPipe (Schenck et al., 2019), reducen la cantidad de herramientas utilizadas en la detección 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 1b). 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. (Abualrous et al., 2021). En este contexto, los *pipelines* Valid-NEO (Terai et al., 2022) y NeoPredPipe (Schenck et al., 2019) y Neopepsee (Kim et al., 2018) 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 innecesario 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 (Machaca et al., 2023; Arceda, 2023).

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	Zhou et al. (2023)	DNA-seq	Neoantígenos
PGNneo	2023	Tan et al. (2023)	VCF, RNA-seq y MS data	Neoantígenos
Valid-NEO	2022	Terai et al. (2022)	VCF y HLA	Neoantígenos
HLA3D	2022	Li et al. (2022)	VCF, HLA, SMG y HBV	Neoantígenos
NAP-CNB	2021	Wert-Carvajal et al. (2021)	RNA-seq	Neoantígenos
NeoANT-HILL	2020	Coelho et al. (2020)	RNA-seq y VCF	Neoantígenos,GE
Neopepsee	2020	Wood et al. (2020)	VCF y BAM	Neoantígenos
ProGeo-neo	2020	Li et al. (2020)	RNA-seq y VCF	Neoantígenos
pVACtools	2020	Hundal et al. (2020)	VCF	Neoantígenos
NeoPredPipe	2019	Schenck et al. (2019)	VCF y HLA	Neoantígenos,VA
ScanNeo	2019	Wang et al. (2019)	RNA-seq	Neoantígenos
Neopepsee	2018	Kim et al. (2018)	RNA-seq, VCF, HLA	Neoantígenos,GE
PGV Pipeline	2018	Rubinsteyn et al. (2018)	DNA-seq	Neoantígenos

4. Resultados o avances previos

5. Justificación

El cáncer es el mayor problema de salud mundial; sin embargo, los métodos tradicionales basados en cirugías, radioterapias y quimioterapias tienen baja efectividad (Peng et al., 2019). En este contexto, los neoantígenos son factores clave en el desarrollo de vacunas contra el Cáncer (Borden et al., 2022; Chen et al., 2021; Gopanenko et al., 2020). Si se logra desarrollar un método con un buen desempeño, la inmunoterapia del cáncer basada en el desarrollo de vacunas personalizadas, podría utilizarse como alternativa a otros métodos como radioterapias y quimioterapias.

El proyecto tiene dos contribuciones: CONTRIBUCIÓN 01: En el área de ciencia de la computación se va a desarrollar un modelo basado en *transformers* para la predicción del enlace pMHC, actualmente se tienen resultados previos que superan a otros del estado del arte. CONTRIBUCIÓN 02: En el área de la Bioinformática: el desarrollo propio del *pipeline*, representa un reto resolviendo problemas de integración, alto costo computacional, heterogeneidad y modularidad. Además, el *pipeline* utilizará datos de *Mass Spectrometry* (MS) y fusión de genes para obtener mejores resultados que otros métodos del estado del arte.

Finalmente, el contar con una aplicación de detección de neoantígenos desarrollada por la Universidad La Salle y la Universidad Católica San Pablo, realza el nombre de ambas universidades y demuestra que gracias al trabajo en conjunto se puede desarrollar aplicaciones de gran envergadura aplicadas a campos multidisciplinarios y nos alinea a instituciones como el *European Molecular Biology Laboratory* (EMBL) y el *National Institutes of Health* (NIH).

6. Porqué es una investigación básica

7. Pregunta de investigación

8. Objetivos de la investigación

8.1. Objetivo general

Desarrollar el *pipeline* NeoArgos-tools de detección *in-silico* de neoantígenos de cáncer para el desarrollo de vacunas personalizadas.

8.2. Objetivos específicos

1. Analizar qué fuentes de información o datos de entrada recibirá el *pipeline*. Se evaluará DNA-seq, RNA-seq, VCF y como integrar esta información con datos de MS.
2. Analizar qué herramientas se utilizarán para la primera etapa del *pipeline*, 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* (Uhrig et al., 2021) y *FusionQ* (Liu et al., 2013).
4. 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 (Arceda, 2023).
5. Implementar un modelo basado en *transformers* para la predicción del enlace del enlace pMHC al TCR (pMGC-TCR).
6. Integrar las herramientas evaluadas en un contenedor Docker.
7. Comparar el desempeño del *pipeline* con otras herramientas del estado del arte.

9. Metodología

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.

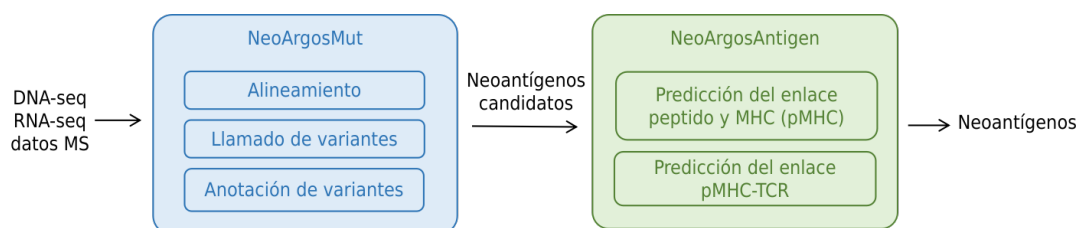


Figura 2: Representación de NeoArgosMut y NeoArgosAntigen para la detección de neoantígenos

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 (Rubinsteyn et al., 2018). Como salida a esta etapa, se obtiene 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 Zhou et al. (2023) y Rubinsteyn et al. (2018). Como salida, se obtienen archivos

VCF. Adicionalmente a otros *pipelines*, utilizaremos información sobre la fusión de genes que se obtendrán de las herramientas *Arriba* (Uhrig et al., 2021) y *FusionQ* (Liu et al., 2013). 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 (Wood et al., 2020). 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 esta 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 priorización la realiza en base a la predicción del enlace de los neoantígenos al MHC y posteriormente 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 $p = \{A, \dots, Q\}$ y $q = \{A, N, \dots, Q, E, G\}$. 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 Zhang et al. (2022) y Gfeller et al. (2023). 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 Zhang et al. (2022). En esta etapa también, se va a evaluar el desempeño de varios modelos BERT pre-entrenados como: TAPE (Rao et al., 2019), ProtBERT-BFD (Elnaggar et al., 2021) y ESM2 (Lin et al., 2023) 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 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 (Arceda, 2023).

La misma arquitectura de la Figura 3, se utilizará para la predicción del enlace pMHC y TCR (pMHC-TCR) según recomendaciones de Li et al. (2020) y Myronov et al. (2023). Sin embargo, se va reentrenar el modelo para adaptarse a este nuevo problema, se utilizarán muestras de Li et al. (2020) y la base de datos de VDJdb (Shugay et al., 2018). Al finalizar esta etapa, se obtendrán los neoantígenos.

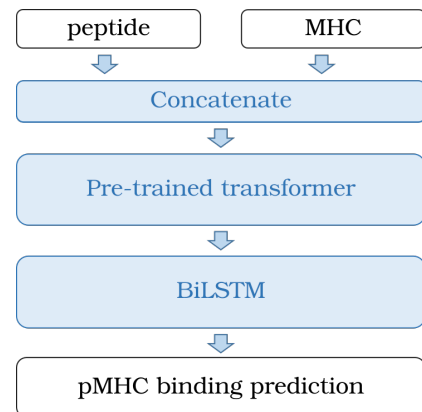


Figura 3: Modelo *transformer* seguido de BiLSTM para predecir el enlace pMHC.

10. Limitaciones

11. Describa el equipamiento e infraestructura disponible

12. Resultados esperados

13. Cronograma de Actividades

En la Tabla 2, presentamos el cronograma de actividades por mes. Adicionalmente, se tendrán reuniones semanales con el equipo de trabajo para comunicar resultados y compartir conocimiento.

Tabla 2: Cronograma de actividades por mes.

Actividades	Entregable	1	2	3	4	5	6	7	8
HITO I (NeoArgosMut)	Informe								
Revisión de la literatura		x	x	x	x	x	x		
Reuniones con todo el equipo de trabajo para el control de actividades		x	x	x	x	x	x	x	x
Capacitación a los miembros del equipo en Bioinformática e Immunoinformática		x	x						
Análisis y evaluación de herramientas para alineamiento de secuencias.		x	x						
Análisis y evaluación de herramientas para la detección de mutaciones y variantes.		x	x						
Análisis y evaluación de métodos para integrar datos MS con DNA-seq y RNA-seq.		x	x	x					
Análisis y evaluación de herramientas para la detección de variantes estructurales y fusión de genes.			x	x	x				
Análisis y evaluación de herramientas para la anotación de variantes.					x	x			
Análisis y evaluación de herramientas para determinar el tipo de HLA a partir de DNA-seq y RNA-seq.					x	x			
Integración de todas las herramientas y desarrollo de NeoArgosMut.							x		
HITO II (NeoArgosAntigen)	Informe								
Implementación y entrenamiento de <i>transformers</i> para la predicción pMHC.		x	x	x	x				
Entrenamiento de <i>transformers</i> para la predicción pMHC-TCR.				x	x	x			
Evaluación de los modelos y comparación con otros del estado del arte.						x	x		
Integración de NeoArgosMut con NeoArgosAntigen.							x	x	x
Redacción del artículo científico						x	x	x	x
Difusión de resultados									x

14. Sostenibilidad de la propuesta

15. Impacto Científico de la propuesta

Referencias

- Abualrous, E. T., Sticht, J., and Freund, C. (2021). Major histocompatibility complex (mhc) class i and class ii proteins: impact of polymorphism on antigen presentation. *Current Opinion in Immunology*, 70:95–104.
- Arceda, V. E. M. (2023). Neoantigen detection using transformers and transfer learning in the cancer immunology context. In *International Conference on Practical Applications of Computational Biology & Bioinformatics*, pages 97–102. Springer.
- Bassani-Sternberg, M., Pletscher-Frankild, S., Jensen, L. J., and Mann, M. (2015). Mass spectrometry of human leukocyte antigen class i peptidomes reveals strong effects of protein abundance and turnover on antigen presentation*[s]. *Molecular & Cellular Proteomics*, 14(3):658–673.
- Borden, E. S., Buetow, K. H., Wilson, M. A., and Hastings, K. T. (2022). Cancer neoantigens: Challenges and future directions for prediction, prioritization, and validation. *Frontiers in Oncology*, 12.
- Bulik-Sullivan, B., Busby, J., Palmer, C. D., Davis, M. J., Murphy, T., Clark, A., Busby, M., Duke, F., Yang, A., Young, L., et al. (2019). Deep learning using tumor hla peptide mass spectrometry datasets improves neoantigen identification. *Nature biotechnology*, 37(1):55–63.
- Chen, I., Chen, M., Goedegebuure, P., and Gillanders, W. (2021). Challenges targeting cancer neoantigens in 2021: a systematic literature review. *Expert Review of Vaccines*, 20(7):827–837.
- Coelho, A. C. M., Fonseca, A. L., Martins, D. L., Lins, P. B., da Cunha, L. M., and de Souza, S. J. (2020). neoant-hill: an integrated tool for identification of potential neoantigens. *BMC Medical Genomics*, 13(1):1–8.
- Elnaggar, A., Heinzinger, M., Dallago, C., Rehawi, G., Wang, Y., Jones, L., Gibbs, T., Feher, T., Angerer, C., Steinegger, M., et al. (2021). Prottrans: Toward understanding the language of life through self-supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, 44(10):7112–7127.
- Fang, X., Guo, Z., Liang, J., Wen, J., Liu, Y., Guan, X., and Li, H. (2022). Neoantigens and their potential applications in tumor immunotherapy. *Oncology Letters*, 23(3):1–9.
- Gfeller, D., Schmidt, J., Croce, G., Guillaume, P., Bobisse, S., Genolet, R., Queiroz, L., Cesbron, J., Racle, J., and Harari, A. (2023). 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*, 14(1):72–83.
- Gopanenko, A. V., Kosobokova, E. N., and Kosorukov, V. S. (2020). Main strategies for the identification of neoantigens. *Cancers*, 12(10):2879.
- Han, X.-J., Ma, X.-l., Yang, L., Wei, Y.-q., Peng, Y., and Wei, X.-w. (2020). Progress in neoantigen targeted cancer immunotherapies. *Frontiers in Cell and Developmental Biology*, 8:728.
- Hundal, J., Kiwala, S., McMichael, J., Miller, C. A., Xia, H., Wollam, A. T., Liu, C. J., Zhao, S., Feng, Y.-Y., Graubert, A. P., et al. (2020). pvactools: a computational toolkit to identify and visualize cancer neoantigens. *Cancer immunology research*, 8(3):409–420.
- Kim, S., Kim, H. S., Kim, E., Lee, M., Shin, E.-C., and Paik, S. (2018). Neopepsee: accurate genome-level prediction of neoantigens by harnessing sequence and amino acid immunogenicity information. *Annals of Oncology*, 29(4):1030–1036.
- Li, X., Lin, X., Mei, X., Chen, P., Liu, A., Liang, W., Chang, S., and Li, J. (2022). Hla3d: an integrated structure-based computational toolkit for immunotherapy. *Briefings in bioinformatics*, 23(3):bbac076.

- Li, Y., Wang, G., Tan, X., Ouyang, J., Zhang, M., Song, X., Liu, Q., Leng, Q., Chen, L., and Xie, L. (2020). Progeo-neo: a customized proteogenomic workflow for neoantigen prediction and selection. *BMC medical genomics*, 13(5):1–11.
- Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., Verkuil, R., Kabeli, O., Shmueli, Y., et al. (2023). Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130.
- Liu, C., Ma, J., Chang, C. J., and Zhou, X. (2013). Fusionq: a novel approach for gene fusion detection and quantification from paired-end rna-seq. *BMC bioinformatics*, 14(1):1–11.
- Machaca, V. E., Goyzueta, V., Cruz, M., and Tupac, Y. (2023). 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*, pages 14–23. Springer.
- Mattos, L., Vazquez, M., Finotello, F., Lepore, R., Porta, E., Hundal, J., Amengual-Rigo, P., Ng, C., Valencia, A., Carrillo, J., et al. (2020). Neoantigen prediction and computational perspectives towards clinical benefit: recommendations from the esmo precision medicine working group. *Annals of oncology*, 31(8):978–990.
- Mill, N. A., Bogaert, C., van Crielinge, W., and Fant, B. (2022). neoms: Attention-based prediction of mhc-i epitope presentation. *bioRxiv*.
- Myronov, A., Mazzocco, G., Krol, P., and Plewczynski, D. (2023). Bertrand-peptide: Tcr binding prediction using bidirectional encoder representations from transformers augmented with random tcr pairing. *bioRxiv*, pages 2023–06.
- O'Donnell, T. J., Rubinsteyn, A., and Laserson, U. (2020). Mhcflurry 2.0: improved pan-allele prediction of mhc class i-presented peptides by incorporating antigen processing. *Cell systems*, 11(1):42–48.
- Peng, M., Mo, Y., Wang, Y., Wu, P., Zhang, Y., Xiong, F., Guo, C., Wu, X., Li, Y., Li, X., et al. (2019). Neoantigen vaccine: an emerging tumor immunotherapy. *Molecular cancer*, 18(1):1–14.
- Rao, R., Bhattacharya, N., Thomas, N., Duan, Y., Chen, P., Canny, J., Abbeel, P., and Song, Y. (2019). Evaluating protein transfer learning with tape. *Advances in neural information processing systems*, 32.
- Reynisson, B., Alvarez, B., Paul, S., Peters, B., and Nielsen, M. (2020). Netmhcpa-4.1 and netmhciipa-4.0: improved predictions of mhc antigen presentation by concurrent motif deconvolution and integration of ms mhc eluted ligand data. *Nucleic acids research*, 48(W1):W449–W454.
- Rubinsteyn, A., Kodysh, J., Hodes, I., Mondet, S., Aksoy, B. A., Finnigan, J. P., Bhardwaj, N., and Hammerbacher, J. (2018). Computational pipeline for the pgv-001 neoantigen vaccine trial. *Frontiers in immunology*, 8:1807.
- Schenck, R. O., Lakatos, E., Gatenbee, C., Graham, T. A., and @miscNCIdictionary2022, author = NCI, title = National Cancer Institute Dictionary, year = 2022, url = <https://www.cancer.gov/publications/dictionaries/genetics-dictionary>, urldate = 2022-03-20 Anderson, A. R. (2019). Neopredpipe: high-throughput neoantigen prediction and recognition potential pipeline. *BMC bioinformatics*, 20(1):1–6.
- Shugay, M., Bagaev, D. V., Zvyagin, I. V., Vroomans, R. M., Crawford, J. C., Dolton, G., Komech, E. A., Sycheva, A. L., Koneva, A. E., Egorov, E. S., et al. (2018). Vdjdb: a curated database of t-cell receptor sequences with known antigen specificity. *Nucleic acids research*, 46(D1):D419–D427.
- Siegel, R. L., Miller, K. D., Wagle, N. S., and Jemal, A. (2023). Cancer statistics, 2023. *Ca Cancer J Clin*, 73(1):17–48.
- Tan, X., Xu, L., Jian, X., Ouyang, J., Hu, B., Yang, X., Wang, T., and Xie, L. (2023). Pgnneo: A proteogenomics-based neoantigen prediction pipeline in noncoding regions. *Cells*, 12(5):782.
- Terai, Y. L., Huang, C., Wang, B., Kang, X., Han, J., Douglass, J., Hsiue, E. H.-C., Zhang, M., Purohit, R., deSilva, T., et al. (2022). Valid-neo: A multi-omics platform for neoantigen detection and quantification from limited clinical samples. *Cancers*, 14(5):1243.

- Uhrig, S., Ellermann, J., Walther, T., Burkhardt, P., Fröhlich, M., Hutter, B., Toprak, U. H., Neumann, O., Stenzinger, A., Scholl, C., et al. (2021). Accurate and efficient detection of gene fusions from rna sequencing data. *Genome research*, 31(3):448–460.
- UK, C. R. (2023a). Worldwide cancer incidence statistics.
- UK, C. R. (2023b). Worldwide cancer statistics.
- Wang, T.-Y., Wang, L., Alam, S. K., Hoepfner, L. H., and Yang, R. (2019). Scanneo: identifying indel-derived neoantigens using rna-seq data. *Bioinformatics*, 35(20):4159–4161.
- Wert-Carvajal, C., Sánchez-García, R., Macías, J. R., Sanz-Pamplona, R., Pérez, A. M., Alemany, R., Veiga, E., Sorzano, C. Ó. S., and Muñoz-Barrutia, A. (2021). Predicting mhc i restricted t cell epitopes in mice with nap-cnb, a novel online tool. *Scientific reports*, 11(1):1–10.
- Wood, M. A., Nguyen, A., Struck, A. J., Ellrott, K., Nellore, A., and Thompson, R. F. (2020). Neoepiscopes improves neopeptide prediction with multivariant phasing. *Bioinformatics*, 36(3):713–720.
- Yadav, M., Jhunjhunwala, S., Phung, Q. T., Lupardus, P., Tanguay, J., Bumbaca, S., Franci, C., Cheung, T. K., Fritsche, J., Weinschenk, T., et al. (2014). Predicting immunogenic tumour mutations by combining mass spectrometry and exome sequencing. *Nature*, 515(7528):572–576.
- Zhang, Y., Zhu, G., Li, K., Li, F., Huang, L., Duan, M., and Zhou, F. (2022). Hlab: learning the bilstm features from the protbert-encoded proteins for the class i hla-peptide binding prediction. *Briefings in Bioinformatics*.
- Zhou, L. Y., Zou, F., and Sun, W. (2023). Prioritizing candidate peptides for cancer vaccines through predicting peptide presentation by hla-i proteins. *Biometrics*, 79(3):2664–2676.