Identificación de neoantígenos a partir de datos de ultrasecuenciación y aplicación posterior en análisis de supervivencia en pacientes con cáncer colorrectal

Marina Gómez Rey
Máster en Bioinformática y Bioestadística
Estudio genómico del cáncer

Rebeca Sanz Pamplona
José Antonio Morán Moreno

02/01/2018

FICHA DEL TRABAJO FINAL

Título del trabajo:	*Identificación de neoantígenos a partir de datos de ultrasecuenciación y aplicación posterior en análisis de supervivencia en pacientes con cáncer colorrectal.*
Nombre del autor:	*Marina Gómez Rey*
Nombre del consultor/a:	*Rebeca Sanz Pamplona*
Nombre del PRA:	*José Antonio Morán Moreno*
Fecha de entrega:	01/2018
Titulación::	*Máster en Bioinformática y Bioestadística*
Área del Trabajo Final:	*Estudio genómico del cáncer*
Idioma del trabajo:	*Español*
Palabras clave	*Inmuno-oncologia, neoatígenos, NGS*

Resumen del Trabajo:

El cáncer es considerado una enfermedad génica muy compleja y caracterizada, entre otras cosas, por acumular un gran número de mutaciones. El reconocimiento de dichas mutaciones puede proporcionar una vía a alcanzar mejoras en los tratamientos como también en su predicción y prevención. Los neoantígenos, desarrollan un papel fundamental en el sistema inmunitario relacionado con el cáncer, pueden desencadenar una respuesta en el sistema inmunitario activando la destrucción de las células tumorales. Es por este motivo, que su caracterización e identificación son fundamentales. El NGS ha generado una gran oportunidad en el mundo de la investigación, ya que ofrece elevadas cantidades de datos y proporciona una base sólida en el avance de las investigaciones. En este proyecto, se desarrolla una predicción y posterior análisis de supervivencia de neoantígenos en cáncer colorrectal. Con el fin de alcanzar los propósitos, se han obtenido los datos de NGS a partir del portal de TCGA y se ha realizado su posterior análisis mediante herramientas bioinformáticas y servidores web especializados. El resultado obtenido ha sido un listado de neoantígenos diferenciados según su afinidad de unión al complejo MHC-I y un análisis de supervivencia con los neoantígenos como variable predictor. Se concluye que el proyecto se ha realizado con éxito obteniendo un método más estricto y fiable de predicción de neoantígenos.

Abstract:

Cancer is considered the most complex genetic disease and its main characteristic is the large number of mutations that characterize it. The recognition of these mutations can provide a way to achieve improvements in treatments as well as in their prediction and prevention. The neoantigens, develop a fundamental role in the immune system related to cancer, can tigger a response in the immune system activating the destruction of tumor cells. For this reason, its characterization and identification are fundamental. The NGS has generated a great opportunity in the world of research, as it offers large amounts of data and provides a solid foundation in the advancement of research. In this project, a prediction and subsequent survival analysis of neoantigens in colorectal cancer is developed. In order to achieve the purposes, the NGS data have been obtained from the TCGA portal and its subsequent analysis has been carried out using bioinformatics tools and specialized web servers. The result obtained was a list of differentiated antigens according to their binding affinity to the MHC-I complex and a survival analysis with neoantigens as a variable predictor. It is concluded that the project has been carried out successfully obtaining a more strict and reliable method of predicting neoantigens.

Índice

Lista de figuras

1. Introducción

1.1 Contexto y justificación del Trabajo

La genética es una disciplina cuyo objetivo principal se basa en entender la herencia y variación de los organismos vivos. La herencia de los organismos vivos actúa principalmente a través del ácido desoxirribonucleico (DNA). El DNA del genoma humano contiene más de tres mil millones de pares de bases ordenados en paquetes llamados cromosomas. De estos, 22 están presentes en dos copias y van acompañados de los cromosomas sexuales, en el caso de las mujeres XX y en el caso de los hombres XY. Tan solo una pequeña parte del DNA forma parte de las regiones que codifican proteínas.

El dogma central de la biología molecular, establece que la información genética discurre en sentido DNA, RNA y, por último, proteínas. Las regiones del DNA que se transcriben a proteínas se llaman genes y contienen una estructura típica compuesta por exones (partes codificantes de proteína) e intrones (partes no codificantes de proteína). Cuando el DNA se transcribe en mRNA, tiene lugar un proceso de maduración, en el que los intrones son eliminados de la secuencia y da lugar a un mRNA maduro. Éste, se transporta del núcleo al citosol celular, una vez ahí, los ribosomas traducen las bases nitrogenadas en aminoácidos, que forman las proteínas. La traducción se realiza a partir de tripletes de nucleótidos (codones) y cada combinación corresponde a uno de los 20 aminoácidos posibles [1].

El cáncer, considerado una enfermedad génica, tiene como característica el crecimiento de las células tumorales iniciadas por mutaciones que activan los controladores de crecimiento celular u oncogénico [2]. Las alteraciones génicas que originan este proceso pueden ser debidas a factores externos, producidas por un error durante la replicación, o duplicación, del DNA o causadas por un virus. Estas mutaciones, provocan que muchos genes, que controlan mecanismos celulares de manera muy precisa, contengan errores y que estos mecanismos que controlan puedan dejar de funcionar de manera correcta. Se trata de un proceso que implica la pérdida gradual de la regulación sobre el crecimiento y las capacidades funcionales de las células normales [3].

En muchos cánceres, la oncogénesis viene acompañada de una gran acumulación de mutaciones. Antiguamente, se creía que ésta gran acumulación de mutaciones provocaba cánceres muy virulentos y, a su vez, altamente mortíferos. No obstante, estudios recientes han asegurado que el alto número de mutaciones en los tumores hace que la célula tumoral diverja mucho de una célula normal o sana y, por lo tanto, puede provocar un rápido reconocimiento de éstos por parte del sistema inmunitario como célula extraña [2].

El sistema inmunitario tiene un papel complejo en el cáncer. Las células inmunes pueden actuar tanto de supresores del tumor como de promotores de la proliferación, infiltración y metástasis del cáncer.

Dentro del ambiente tumoral (células, moléculas y vasos sanguíneos que rodean y alimentan un tumor), hay descritas varias células que coinciden en

prácticamente en todos los tumores, estas son células inmunitarias innatas como son los macrófagos, los mastocitos, neutrófilos, células dendríticas, células supresoras y *natural killers* (NK) como células inmunitarias adaptativas como son los linfocitos o células T y B [3].

Las células continuamente muestran pequeños péptidos (fragmentos de proteína) en el exterior de la membrana celular. Este proceso se inicia en el interior de la célula, donde los proteosomas (grandes complejos proteicos) rompen las proteínas en pequeños péptidos. Dichos péptidos, son trasladados al retículo endoplasmático donde se encuentra el Complejo Mayor de Histocompatibilidad (MHC). Una vez reconocido el péptido por el MHC, se genera una unión y complejo estable (péptido-MHC). En el momento en el que se efectúa la unión, el MHC es trasladado a la superficie celular donde se expone durante varias horas.

Cabe destacar, que dependiendo de la naturaleza del antígeno, la respuesta del MHC es diferente. Hay presentes dos isoformas del MHC, la de tipo I y la de tipo II, los antígenos endógenos, se presentan por las MHC-I a células T tipo CD8 (los linfocitos T citotóxicos) en cambio, las MHC-II presentan un tipo más restrictivo de péptidos a las células T tipo CD4 (linfocitos T colabodatores o *helpers),* se trata de antígenos exógenos, que provienen de virus y/o bacterias [4].

Por lo tanto, para que un neoantígeno (péptido mutado surgido de un proceso tumoral) sea reconocido por las células T CD8, éste debe ser presentado por las MHC-I en la superficie celular tumoral [5].

En la siguiente imagen, se puede observar como el proteosoma rompe en pequeños péptidos la proteína. Estos péptidos son transportados al retículo endoplasmático y donde se unen al complejo MHC-I generando una unión estable. Una vez generada dicha unión, el complejo sale del retículo hacia la pared celular, donde es reconocida por el receptor TCR del linfocito T CD8.

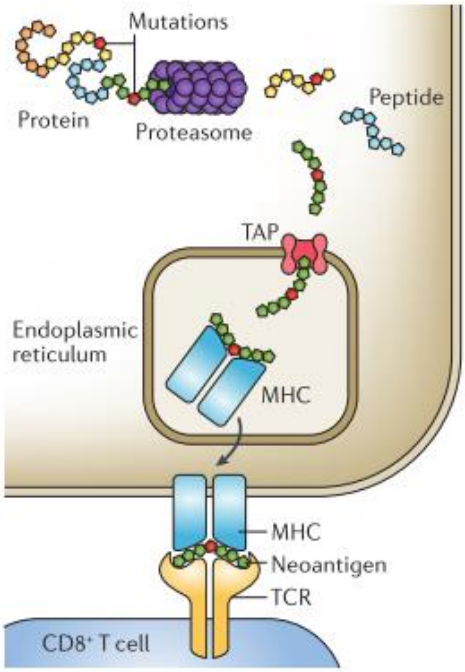


Ilustración 1: Generación de neoantígenos. (Extraído de: [6]).

El *Next generation sequencing* (NGS) proporciona grandes herramientas para la identificación de las mutaciones características del cáncer y esto, conduce al conocimiento en profundidad de la enfermedad y acelera el descubrimiento de técnicas para el diagnóstico, prevención y tratamiento. Proyectos como *The Cancer Genome Atlas* (TCGA) [7], se encargan de secuenciar el genoma de las células cancerosas.

La necesidad que surge en identificar y caracterizar los neoantígenos se basa en la posibilidad de poder realizar terapias inmunogénicas dirigidas a pacientes con cáncer. La inmunoterapia consiste en la inmunización de un individuo mediante anticuerpos específicos.

Para la realización de la inmunoterapia dirigida en forma de vacuna se debe realizar la predicción de neoantígenos, esta predicción se basa en la secuenciación de células tumorales y normales del paciente y su posterior comparación para la identificación de las mutaciones, a continuación se obtiene el péptido mutado con el fin de poder definir cuál es la probabilidad de unión con la molécula MHC-I. Para la realización de las vacunas, se seleccionan los neoantígenos con una unión fuerte a MHC-I, con el fin de poder desencadenar una respuesta inmunitaria más fuerte [8].

El resultado esperado de este proyecto es la caracterización e identificación de neoantígenos para un grupo de pacientes con cáncer extraídos del TCGA en relación al tipo de unión con el MHC-I y su posterior análisis de supervivencia.

1.2 Objetivos del Trabajo

Los principales objetivos que se desean alcanzar con este proyecto es generar un script para predecir neoantígenos en pacientes con Cáncer colorrectal (a partir de datos de ultrasecuenciación) y comprobar su utilidad como marcador pronóstico.

Los objetivos se componen de la siguiente manera:

- Predicción de neoantígenos

Traducción de mutaciones missence a péptidos mutados mediante técnicas bioinformáticas. Uso de programas bioinformáticaos para predecir péptidos inmunogénicos; y comparar diferentes resultados. Análisis bioestadístico descriptivo de los resultados obtenidos.

- Análisis de supervivencia

Análisis bioestadístico descriptivo y de supervivencia utilizando los datos obtenidos a partir de la realización de la predicción de neoantígenos.

1.3 Enfoque y método seguido

Las posibles estrategias para llevar a cabo el trabajo son múltiples, a continuación, se listarán posibles estrategias según los objetivos y se razonará cuál es la escogida y su motivo:

- *Predicción de neoantígenos*, es necesario tener datos de secuenciación masiva. Una manera de poder acceder a dichos datos sería poder colaborar con equipos de investigación para que los facilitaran o bien, otra opción sería poder obtenerlos de manera libre.

En este caso, se llevará a cabo la segunda opción. Por un lado, se evitan problemas de confidencialidad ya que dichos datos son accesibles para toda la comunidad científica. Por otro lado, resulta interesante explotar y reciclar datos masivos que precisamente han sido generados con esa finalidad [7].

- *Análisis de datos y análisis bioestadístico*, existen numerosas herramientas como por ejemplo por ejemplo SPSS, S-Plus, Python y R.

El estudio estadístico se realizará en R ya que es el que se considera que puede proporcionar más herramientas bioestadísticas y el más completo para éste estudio en específico [9].

- *Predicción de péptidos inmunogénicos,* existen diversos servidores para la predicción de la unión de péptidos a las moléculas del Complejo Mayor de Histocompatibilidad (MHC). Algunos ejemplos son el servidor NetMHC, que es un método basado en redes neuronales (ANN) y entrenado con 94 alelos de MHC, el servidor NetMHCpan también basado en ANN y entrenado con más de 115000 datos de unión y el método PickPocket que está basado en una matriz de similitudes y que ha sido entrenado con 94 alelos diferentes de MHC.

En el estudio, se utilizará el servidor NetMHCcons 1.1, el cual predice la unión de péptidos a cualquier molécula conocida de MHC de clase I. Éste método es muy restrictivo ya que se basa en el consenso de los tres servidores NetMHC, NetMHCpan y PickPocket, por lo tanto, se considera que puede ser el más adecuado para la predicción [10].

- *Análisis de supervivencia*, también se realizará en R, utilizando como variable predictora el número de neoantigenos de cada paciente.

1.4 Planificación del Trabajo

A continuación se hará una descripción de los recursos que han sido necesarios para poder realizar el trabajo en función de la planificación temporal:

1. Predicción de neoantígenos:

1.1 Traducción de mutaciones missense a péptidos mutados mediante técnicas bioinformáticas.

Tarea 1: Obtención de datos. (2 días)
Obtención de datos del TCGA a partir de los servidores Broad Institute y cBioPortal.

Tarea 2: Análisis de datos. (3 días)

Tarea 3: Selección de datos (datos missense). (1 día)

Tarea 4: Procesamiento de datos. (7 días)

Tarea 5: Traducción de mutaciones missense a neoantígenos. (10 días)
Tareas 2, 3, 4 y 5: mediante la herramienta RStudio y sus respectivas librerías.

1.2 Uso de programas bioinformáticos para predecir péptidos inmunogénicos; y comparar los diferentes resultados.

Tarea 6: Predicción de péptidos inmunogénicos. (2 días)
Herramienta web NetMHCcons.

Tarea 7: Comparación entre diferentes resultados. (6 días)
Herramienta RStudio y sus respectivas librerías.

1.3 Análisis bioestadístico descriptivo de los resultados obtenidos.

Tarea 8: Análisis bioestadístico descriptivo. (6 días)
Herramienta RStudio y sus respectivas librerías.

2. Análisis de supervivencia:

2.1 Análisis bioestadístico descriptivo y de supervivencia utilizando los datos generados a partir de la realización de la predicción de neoantígenos.

Tarea 9: Procesamiento de datos (4 días)

Tarea 10: Análisis bioestadístico descriptivo y de supervivencia. (9 días)
Tareas 9 y 10: mediante la herramienta RStudio y sus respectivas librerías.

En la siguiente imagen se puede observar un diagrama de Gantt, muy útil en el momento de realizar una planificación temporal.

DIAGRAMA DE GANTT

	Nombre de tarea	Duración	Comienzo	Fin	Predecesoras
1	**Traducción de mutaciones missense a neoantígenos**	**25 días**	**lun 16/10/17**	**lun 20/11/17**	
2	Obtención datos	2 días	lun 16/10/17	mar 17/10/1	
3	Análisis datos	3 días	mié 18/10/1	vie 20/10/17	2
4	Selección datos	3 días	lun 23/10/17	mié 25/10/1	3
5	Procesamiento de datos	7 días	jue 26/10/17	vie 03/11/17	4
6	Traducción de mutaciones missense a neoantígenos	10 días	lun 06/11/17	vie 17/11/17	5
7	Entrega desarrollo fase I	0 días	lun 20/11/17	lun 20/11/17	
8	**Predicción de péptidos y comparación de resultados**	**6 días**	**lun 20/11/17**	**lun 27/11/17**	
9	Uso servidor NetMHCcons 1.1	2 días	lun 20/11/17	mar 21/11/1	6
10	Comparación de resultados	4 días	mié 22/11/1	lun 27/11/17	9
11	**Análisis bioestadístico**	6 días	mar 28/11/1	mar 05/12/1	10
12	**Análisis bioestadístico descriptivo y de supervivencia**	**15 días**	**mar 28/11/1**	**lun 18/12/17**	
13	Procesamiento de datos	4 días	mar 28/11/1	vie 01/12/17	10
14	Análisis descriptivo y de supervivencia	9 días	mié 06/12/1	lun 18/12/17	11
15	Entrega desarrollo fase II	0 días	lun 18/12/17	lun 18/12/17	
16	**Redacción de la memória**	11 días	mar 19/12/1	mar 02/01/1	14

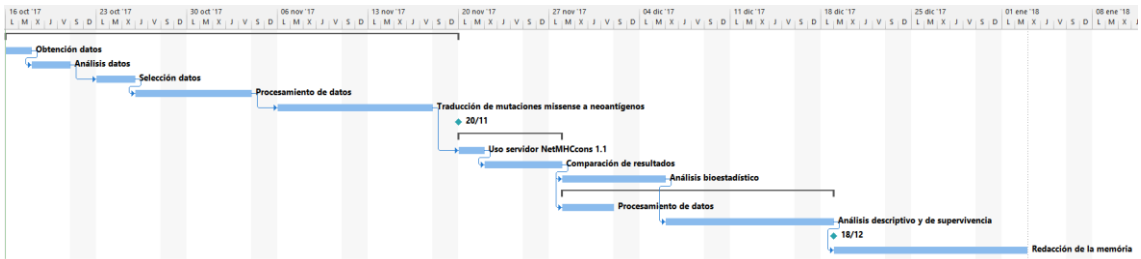


Ilustración 2: Listado de tareas y diagrama de GANTT.

1.5 Breve sumario de productos obtenidos

Los productos obtenidos a partir de la primera fase del proyecto son un total de 6.455 neoantígenos. Se puede observar que de los 223 pacientes 206 presentan neoantígenos con unión fuerte a MHC-I y 216 presentan neoantígenos con unión débil a MHC-I. También se ha realizado una descriptiva estadística.

En la segunda fase del proyecto se han obtenido diferentes gráficas de supervivencia en función del número de neoantígenos. En primer lugar se ha realizado una gráfica de supervivencia en función del total de neoantígenos sin discriminar en su tipo de unión, en segundo lugar se ha analizado la supervivencia en función del tipo de unión fuerte y, en tercer lugar se ha realizado la supervivencia en función del tipo de unión débil.

1.6 Breve descripción de los otros capítulos de la memoria

A continuación se detallará el contenido de cada capítulo y su relación con el trabajo global:

- Capítulo 2: Elección de Herramientas Bioinformáticas

Dada la relevancia que supone la elección de las diferentes herramientas bioinformáticas en el momento de realizar un estudio, en el Capítulo 2 se ha realizado un análisis de las herramientas que han sido necesarias para la elaboración del proyecto. También se ha llevado a cabo una comparativa sobre las diferentes herramientas especificando el motivo de la elección.

En el Anexo 1, se puede observar el script con el que se ha desarrollado el proyecto presentado en formato Rmarkdown.

- Capítulo 3: Predicción de neoantígenos

En el Capítulo 3, se ha detallado el procedimiento a seguir para alcanzar el primer objetivo, la predicción de neoantígenos. Se ha especificado las características de los datos en los que se basa el estudio, los procedimientos y materiales empleados además de los resultados obtenidos.

- Capítulo 4: Análisis de supervivencia

En el Capítulo 4, se ha detallado el procedimiento a seguir para alcanzar el segundo objetivo, el análisis de supervivencia. Se ha especificado las características de los datos en los que se basa el estudio, los procedimientos y materiales empleados además de los resultados obtenidos.

2. Elección de Herramientas Bioinformáticas

El NGS genera grandes cantidades de datos que en muchos casos un ordenador convencional no puede procesar. Es por este motivo que cada vez se va generando más *software* y servidores web que facilitan dicho procesamiento. No obstante, su elección, procesamiento e interpretación requieren de conocimientos tanto biológicos como bioinformáticos. A pesar del gran número y diversidad de herramientas que existen su elección representa un gran desafío.

Dado que el procesamiento requiere una división de los procesos en subprocesos para obtener los resultados intermedios, es necesaria una cantidad significativa de herramientas.

En los apartados siguientes se detallarán las herramientas que han sido necesarias para el procesamiento de los datos.

2.1. Software para la anotación de variantes

La identificación de variantes génicas, en este estudio, se realiza a partir de los datos de NGS obtenidos a partir del servidor del TCGA (en el siguiente apartado se detallan dichos datos de partida). Éstos datos nos proporcionan los listados de variantes candidatas, no obstante, debemos de anotarlas. Identificar el gen donde se encuentra, la posición y el cambio de variante. Aunque la anotación puede realizarse variante a variante extrayendo la información de bases de datos especializadas, existen herramientas bioinformáticas que facilitan dicha anotación.

Para la realización de anotación de variantes existen diversas herramientas, entre las cuales destacan: ANNOVAR que realiza una anotación de SNP e indels, la herramienta GATK que también detecta SNP e indeles pero a su vez realiza un pequeño control de calidad de los datos proporcionados, la herramienta MuTect que realiza una predicción de SNP a muy alta sensibilidad, la herramienta Strelka que realiza detección de SNP e indeles [11].

En nuestro caso, se ha realizado con la herramienta Annovar, en concreto wAnnovar que se trata de una herramienta versátil y ampliamente usada para la anotación funcional de variantes. Además de obtener la variante, también obtendremos el nombre del gen donde se encuentra, se trata de una herramienta muy útil para, de éste modo poder obtener la secuencia de proteína [12].

2.2. Obtención de secuencias proteicas

Para poder predecir los neoantígenos, en primer lugar debemos de obtener la secuencia proteica donde se hallan. Para realizar dicha tarea, existen diversas herramientas en los que introduciendo el nombre del gen, puedes obtener la secuencia proteica final.

Entre ellos cabe destacar Ensambl, RefSep y Uniprot. Todos ellos funcionan de una manera muy similar, introduciendo el nombre del gen obtienes, entre otros, la secuencia de la proteína en cuestión.

En nuestro caso, se ha hecho servir la librería de Bioconductor 'Uniprot.ws', ésta librería permite una interacción con los servicios web de Uniprot (http://www.uniprot.org/), está destinado a permitir la extracción de una manera fácil y sencilla de los datos Uniprot [13].

2.3. Predicción de la afinidad de unión Neoantígeno-MHC

Una vez obtenidas las secuencias y posteriormente los péptidos candidatos, se procede a seleccionar los que tienen una unión con MHC y, por lo tanto, se consideran neoantígenos.

La predicción se puede realizar en función al tipo de MHC al que el neoantígeno tiene unión. Tal y como sabemos, hay presentes dos tipos principales de MHC, los de tipo I (presentadores de péptidos endógenos) y los de tipo II (presentadores de péptidos exógenos). Los algoritmos de predicción de unión al MHC-II son menos precisos que los de predicción de unión al MHC-I. Por lo tanto, cabe destacar que la predicción la realizaremos en base a la unión de MHC-I, que son los presentadores de péptidos tumorales.

Para predecir dicha unión, existen diversas herramientas vía web. A continuación se detallarán las más relevantes:

- IEDB: El *Immune Epitope Database Analysis Resource (http://tools.iedb.org/mhci/)*, se trata de un recurso que ofrece una búsqueda en su propia base de datos de anticuerpos y neoantígenos de células T estudiados previamente en humanos y otras espécies (http://www.iedb.org/). Entre todas las herramientas disponibles hay la posibilidad de realizar una predicción de unión a MHC de tipo I y II.

  Una de las limitaciones principales de esta herramienta, se encuentra en que dicha unión debe haber sido estudiada con anterioridad, factor que limita el estudio.

- NetMHC: se trata de una herramienta web proporcionada por el Centro de Análisis de Secuencia Biológica de la Universidad Técnica de Dinamarca (http://www.cbs.dtu.dk/services/NetMHC/). Se trata de un método basado en redes neuronales artificiales (ANN) que se ha entrenado con un total de 94 alelos de MHC.

  Una de las limitaciones que tiene este método, es que se trata de un método alelo-específico y, por lo tanto, solo predice la unión de las moléculas en las que se ha entrenado [10].

- NetMHCpan: también se trata de una herramienta web proporcionada por el Centro de Análisis de Secuencia Biológica de la Universidad Técnica de Dinamarca (http://www.cbs.dtu.dk/services/NetMHCpan/). Ésta herramienta, genera predicciones cuantitativas de la afinidad de cualquier interacción péptido-MHC sin la necesidad de haber sido estudiada y caracterizadas anteriormente [14].

El servidor predice la unión de péptidos a cualquier molécula de MHC de secuencia conocida usando ANN. El método se entrena combinando más de 180.000 datos de unión a MHC y de ligando, generando de este modo un conjunto de red neuronal que permite hacer predicciones para otras moléculas en comparación con las que se entrenó el método.

- PickPocket: facilitada también por el Centro de Análisis de Secuencia Biológica de la Universidad Técnica de Dinamarca (http://www.cbs.dtu.dk/services/PickPocket/). Predice la unión de péptidos a cualquier molécula de MHC conocida usando matrices de similitud. El método se entrena con un total de 150.000 datos de unión a 150 moléculas de MHC diferentes [15].

- NetMHCcons: facilitada también por el Centro de Análisis de Secuencia Biológica de la Universidad Técnica de Dinamarca (http://www.cbs.dtu.dk/services/NetMHCcons/). Esta herramienta se trata de un consenso de las herramientas NetMHC, NetMHCpan y PickPocket.

  NetMHCcons predice la unión de péptidos a cualquier molécula conocida de MHC-I. Los resultados, se dan en forma de enlaces fuertes o débiles en función de la afinidad pronosticada por el servidor [10].

Se considera que la opción más adecuada para llevar a cabo el pronóstico de neoantígenos se trata del servidor NetMHCcons, ya que se considera una herramienta muy restrictiva dada la necesidad de consenso entre tres métodos diferentes y, por lo tanto, con unos resultados muy precisos.

3. Predicción de neoantígenos

Para realizar la predicción de neoantígenos, se partió de datos de cáncer colorectal. Para dicha predicción es necesario tener datos de secuenciación masiva, concretamente mutaciones, que se obtuvieron a partir de portales web especializados.

También es necesario el uso de herramientas bioinformáticas para procesar los datos de secuenciación masiva. En concreto se usó RStudio 1.1.383, sus respectivos paquetes y Bioconductor. Como también servidores web públicos que facilitan el procesamiento de datos.

A continuación, se detallará el procedimiento seguido para la realización de la predicción de neoantígenos así como también los resultados obtenidos.

En la siguiente imagen (Ilustración 3), se puede observar el flujo de trabajo seguido para poder alcanzar con éxito el proyecto. Partiendo de datos proporcionados por la plataforma del TCGA y mediante los servidores Broad Institute y cBioPortal se procedió a su descarga. Una vez obtenidos, se procedió a la Selección del tipo de mutación de estudio. A continuación, se procedió a la obtención de las Variantes génicas a partir del servidor web wAnnovar y con la librería proporcionada por Bioconductor a la obtención de las secuencias proteicas. Una vez obtenidas dichas secuencias, se procedió a la generación de una función en forma de bucle con el fin de obtener los péptidos candidatos a neoantígenos. A continuación, mediante el servidor web NetMHCcons se procedió a la obtención de la lista de neoantígenos con unión al complejo MHC-I y su posterior análisis bioinformático y de supervivencia.

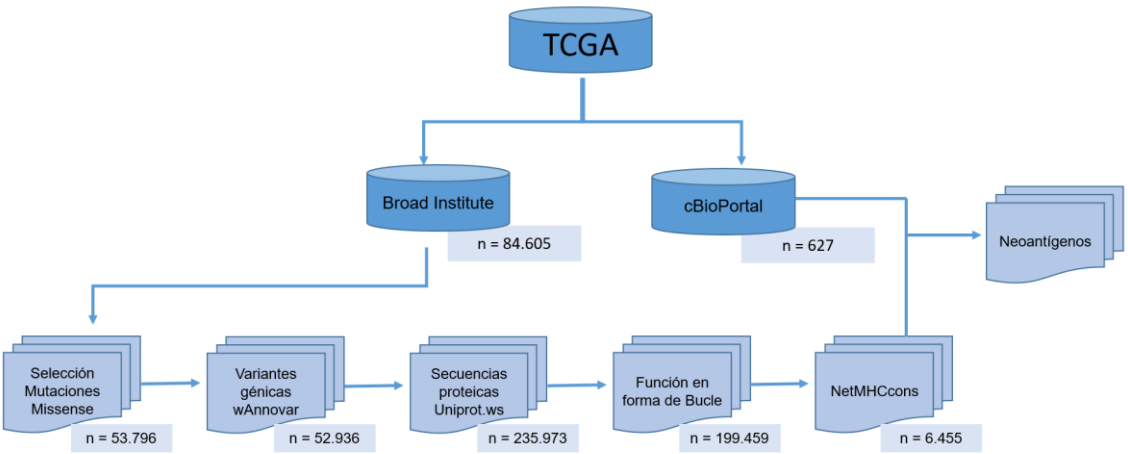


Ilustración 3: Diagrama de flujo.

3.1 Características de los datos

Para iniciar el proyecto, partimos de dos tipos de datos distintos, en primer lugar, los datos clínicos de los pacientes y, en segundo lugar, los datos mutacionales. Ambos, proporcionados por TCGA.

- The Cancer Genome Atlas:

El *National Institute of Health* (NIH) desarrolló su proyecto TCGA con el fin de crear perfiles genómicos completos sobre el cáncer para catalogar y descubrir las principales alteraciones genómicas causantes del cáncer en más de 30 tipos diferentes de tumores. Con la finalidad de proporcionar un análisis completo del genoma los datos se analizaros con tecnologías de alto rendimiento basados en microarrays y métodos de NGS [7].

Para la obtención de dichos datos, existían diversas alternativas, a continuación se detallaran las posibles y el motivo de su elección:

- TCGA: El portal da la opción de descargar los datos, tanto mutacionales como clínicos, no obstante, ésta opción fue descartada dada su complejidad para obtener dichos datos.

- Broad Institute: Eli & Edthe Broad Institute de MIT y Harvard se creó a partir de la colaboración durante décadas de ambas instituciones y diversos hospitales (https://www.broadinstitute.org/about-us). Su proyecto "FireHose" pretende facilitar el acceso a los datos del portal TCGA facilitando su visualización, análisis y descarga.

  Los datos mutacionales se descargaron manualmente del Broad Institute (TCGA data versión 2016_01_28) a través del sitio web firebrowse.org.

  Dado que los datos se encontraban en formato MD5, se procedió a crear una cuenta en GenomeSpace (http://www.genomespace.org/). Se trata de un entorno con diversas herramientas y aplicaciones de genómica que pretende facilitar la transferencia de datos a través de su rápida conversión. Una vez introducidos los datos en MD5, se pueden descargar en un formato en el que su procesamiento sea más inteligible.

  En éste caso, se procedió su descarga en archivo en formato '.maf'.

- cBioPortal: éste portal facilita también la visualización, análisis y descarga de los datos proporcionados por el TCGA [16][17]. Además de proporcionar información génica, cBioPortal proporciona acceso al resumen de información clínica sobre cada uno de los pacientes.

  Se procedió a descargar la información clínica de dicho portal ya que se consideró que la información que contenía era más relevante que en el caso de la información presentada por Broad Institute. En concreto, los datos de cBioPortal se encuentran mejor estructurados y a su vez

contienen información más característica para realizar el análisis de supervivencia.

3.2. Procedimientos y materiales empleados

Como se ha comentado en el apartado anterior, el procedimiento se ha desglosado en diferentes tareas, cada una de ellas se detallará a continuación como también los materiales empleados para realizarlas con éxito.

Traducción de mutaciones missense a péptidos mutados mediante técnicas bioinformáticas

- Obtención y análisis de datos:

Utilizando RStudio 1.1.383, sus respectivos paquetes y Bioconductor, se procedió a leer los datos de cBioPortal con la función “read.csv”. Para la lectura de archivos mutacionales de Broad Institute, se requirió del paquete “maftools” de Bioconductor.

Se realizó un emparejamiento de los datos clínicos y los datos mutacionales con la función “merge”, con la finalidad de mantener únicamente la información clínica de los pacientes de los que se tenía información mutacional. Manteniendo de éste modo la información clínica de 223 pacientes. Además, con la finalidad de reducir el tamaño de los datos y el número de componentes de interés solo se mantuvo un subconjunto de columnas de los datos clínicos para el análisis de supervivencia (Véase apartado 4). Estos incluyen el código del paciente, el sexo, el estado vital, la edad, recaída en el tumor, promedio de supervivencia en años y meses y el diagnostico histológico. Se procedió a guardar los datos para el posterior análisis de supervivencia.

En la Figura 6, se puede observar la frecuencia de mutaciones en función del diagnóstico histológico del cáncer y también en función del sexo. Los tumores que tienen una mayor recurrencia de mutaciones en mujeres son los de Adenocarcinoma en colon no obstante, los que presentan un mayor número de mutaciones en hombres son los de Adenocarcinoma rectales. En la Figura 7, se puede observar la carga mutacional en función de la edad y sexo.

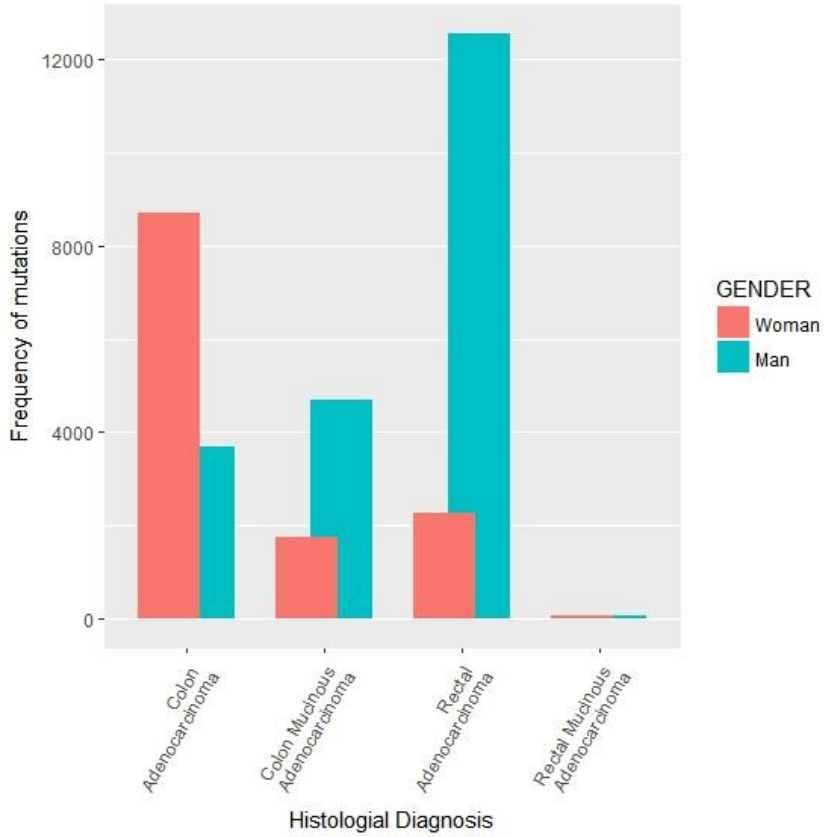


Ilustración 4: Histograma de Frecuencia mutacional en función de la histología tumoral.

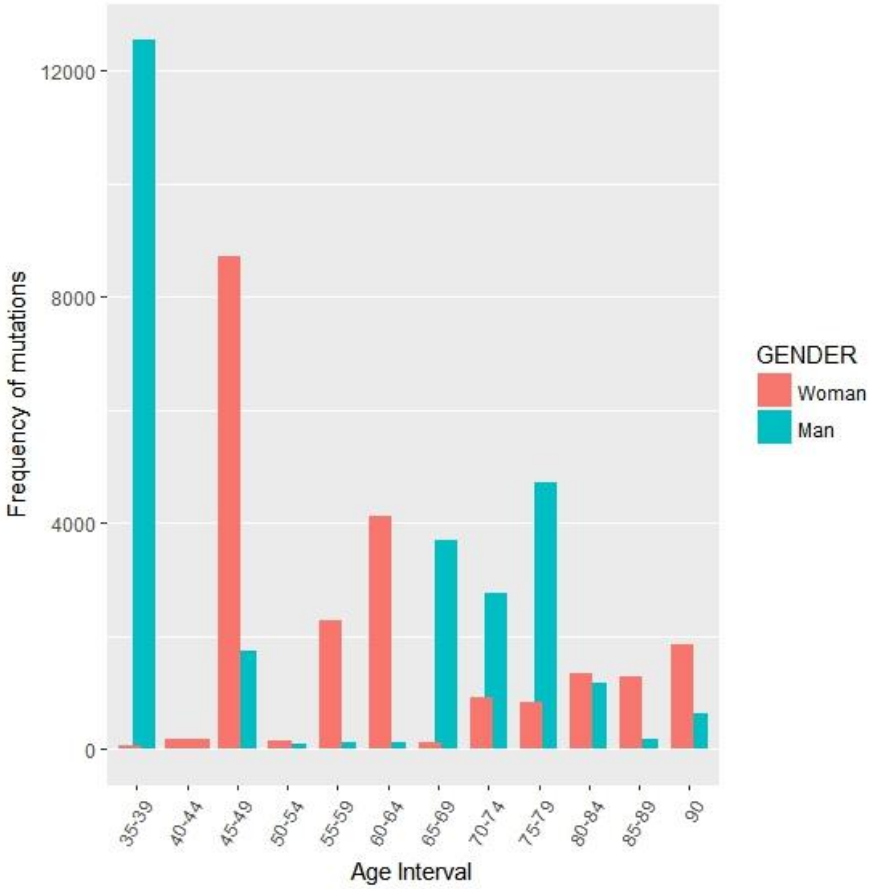


Ilustración 5: Histograma de Frecuencia mutacional en función de los grupos de edad.

- Selección de mutaciones de interés:

A continuación, se seleccionaron los datos mutacionales de interés, en nuestro caso mutaciones missense. Obteniendo de éste modo un total de 223 pacientes con 53.796 datos mutacionales.

Cabe destacar que la elección de este tipo de mutación se realiza debido a que las mutaciones missense generan un cambio de base que tiene como consecuencia un cambio de aminoácido. Dicho cambio de aminoácido hace que la maquinaria celular reconozca la proteína como aberrante y la lleve a degradar por el proteasoma. Producto de dicha degradación se generan pequeños péptidos, alguno de ellos incluyendo el aminoácido mutante y por lo tanto susceptible de ser presentado por el complejo mayor de histocompatibilidad MHC-I.

En la siguiente imagen, se puede observar como el tipo de mutación seleccionada para la realización de la predicción de neoantígenos (mutaciones missense) representa el más abundante entre los diferentes que hay.

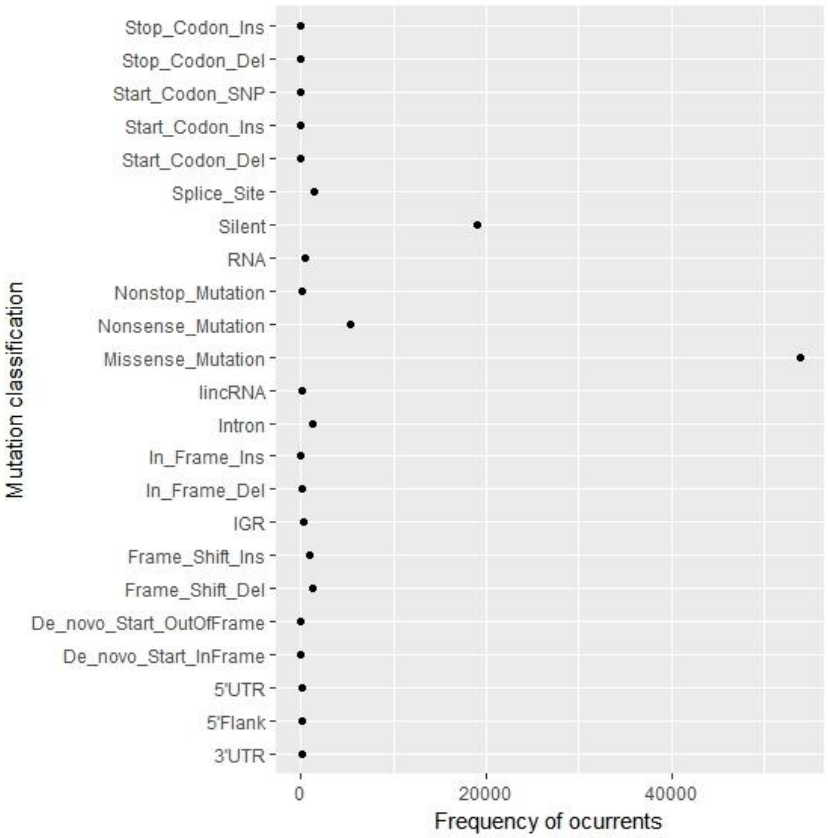


Ilustración 6: Diagrama de tipo mutacional y frecuencia.

- Procesamiento de datos:

De los datos mutacionales obtenidos en el apartado anterior, se seleccionaron las columnas del número de Cromosoma, la posición de inicio, la posición final, el alelo de referencia y el alelo mutado. Una vez obtenida dicha tabla, se procedió a introducirla en el servidor de Wannovar.

Wannovar (http://wannovar.wglab.org/) se trata de una alternativa al software Annovar creado por Wang Genomics Lab de la Universidad de Columbia. Éste servidor web pretende ser una alternativa más eficiente al software, aunque solamente contenga las funciones más relevantes de Annovar, proporciona una manera eficaz de procesar las variantes génicas. Dado que nosotros queremos consultar el cambio específico de aminoácido, esta herramienta nos resulta más eficiente que el software Annovar, ya que no requiere de instalación y su uso es más sencillo. Este es el principal motivo de su elección.

Una vez introducidos los datos y obtenido el output, los cargamos en RStudio. Se procedió a depurar los datos, ya que wAnnovar, proporciona los cambios específicos de aminoácido más probables y seleccionamos la primera opción, ya que se trata de la de mayor probabilidad.

Se obtuvo información sobre un total de 52.936 mutaciones, con información sobre la posición del gen y el cambio específico de aminoácido. El cambio específico de aminoácido nos da la información sobre el aminoácido original, la posición en la proteína en la que se produce la mutación y el aminoácido mutado. Información esencial para poder realizar la traducción de mutaciones missense a neoantígenos.

A continuación, se procedió a la obtención de las secuencias de proteínas utilizando el código de Entrez Gene y el paquete “Uniprot.ws” de Bioconductor. Éste paquete, permite entrar a la base de datos de Uniprot a través del software RStudio y, de este modo, poder obtener la información sobre proteínas que se desee. Una vez realizado el proceso, obtuvimos un total de 47.442 secuencias de proteínas.

Para finalizar, se realizó un emparejamiento de datos con el fin de obtener un conjunto de datos en el que estuvieran presentes el cambio de aminoácido y la secuencia de proteínas. Obteniendo en total de 235.973 proteínas.

- Traducción de mutaciones missense a neoantígenos:

A partir de la información de cambio específico de aminoácido y la secuencia proteica, se procedió a la obtención de los péptidos.

Para la obtención se generó una función en RStudio en forma de bucle. Cabe destacar, que la realización de funciones en RStudio se realizan de una manera muy lenta, por lo tanto, para agilizar la función, fue necesaria una preparación previa de los datos.

El cambio específico de aminoácido, se presenta de la siguiente forma: YpX, en la que ‘Y’ es el aminoácido original, ‘p’ es la posición dentro de la proteína y ‘X’ es aminoácido mutado. Se procedió a separar la información en tres nuevas columnas: en la primera de ellas se introdujo el aminoácido original, en la segunda el aminoácido mutado y en la tercera la posición dentro de la proteína.

A continuación, se generó la función para la obtención de los péptidos. Se trata de un bucle repetitivo que se ejecuta para cada una de las filas de la columna. La manera de ejecutarse en la siguiente: selecciona la posición en la que debería de encontrarse el aminoácido mutado dentro de la proteína, consulta si en dicha posición se encuentra el aminoácido original de la secuencia. En caso positivo, el aminoácido original es substituido por el aminoácido mutado y se seleccionan 9 posiciones anteriores y 9 posteriores a dicha posición. Éste nuevo péptido obtenido de 19 aminoácidos se guarda en una nueva columna. En caso negativo, en que el aminoácido original no se encuentre en la posición de la mutación, en la nueva columna se guarda el valor “NA”.

Una vez depurada la tabla y eliminando las filas que contengan “NA” y los casos duplicados que se traten de la misma mutación obtenemos un total de 199.459 péptidos.

A continuación, se puede ver un ejemplo de cómo funciona la función en forma de secuencia.

En ‘aa_change_l2’ se encuentra el cambio de aminoácido que posteriormente se ha separado en tres columnas distintas. En la columna ‘SEQUENCE’ se puede observar la cadena proteica original, en ‘newseq’ la secuencia mutada y en ‘mutseq’ el péptido obtenido que servirá como input para la herramienta NetMHCcons.

```
aa_change_l2 aa_original aa_mutado aa_position
S216F                S         F         216

SEQUENCE
MGRKSLYLLIVGILIAYYIYTPLPDNVEEPWRMMWINAHLKTIQNLATFVELLGLHHFMDSFKVVGSFDE
VPPTSDENVTVTETKFNNILVRVYVPKRKSEALRRGLFYIHGGGWCVGSAALSGYDLLSRWTADRLDAVV
VSTNYRLAPKYHFPIQFEDVYNALRWFLRKKVLAKYGVNPERIGISGDSAGGNLAAAVTQQLLDDPDVKI
KLKIQSLIYPALQPLDVDLPSYQENSNFLFLSKSLMVRFWSEYFTTDRSLEKAMLSRQHVPVESSHLFKF
VNWSSLLPERFIKGHVYNNPNYGSSELAKKYPGFLDVRAAPLLADDNKLRGLPLTYVITCQYDLLRDDGL
MYVTRLRNTGVQVTHNHVEDGFHGAFSFLGLKISHRLINQYIEWLKENL

newseq
MGRKSLYLLIVGILIAYYIYTPLPDNVEEPWRMMWINAHLKTIQNLATFVELLGLHHFMDSFKVVGSFDE
VPPTSDENVTVTETKFNNILVRVYVPKRKSEALRRGLFYIHGGGWCVGSAALSGYDLLSRWTADRLDAVV
VSTNYRLAPKYHFPIQFEDVYNALRWFLRKKVLAKYGVNPERIGISGDSAGGNLAAAVTQQLLDDPDVKI
KLKIQFLIYPALQPLDVDLPSYQENSNFLFLSKSLMVRFWSEYFTTDRSLEKAMLSRQHVPVESSHLFKF
VNWSSLLPERFIKGHVYNNPNYGSSELAKKYPGFLDVRAAPLLADDNKLRGLPLTYVITCQYDLLRDDGL
MYVTRLRNTGVQVTHNHVEDGFHGAFSFLGLKISHRLINQYIEWLKENL

mutseq
DVKIKLKIQFLIYPALQPL
```

Uso de programas bioinformáticos para predecir péptidos inmunogénicos; y comparación de los diferentes resultados

- Predicción de péptidos inmunogénicos

La predicción de péptidos inmunogénicos se ha realizado con la herramienta web NetMHCcons (http://www.cbs.dtu.dk/services/NetMHCcons/) [10].

Para la realización de la predicción de la predicción se generó un input en formato FASTA. El formato FASTA empieza con una descripción de única línea inferior a los 80 caracteres y con un símbolo mayor que (“>”), seguida de líneas de datos de secuencia [18].

Una vez realizado el formato FASTA, se procedió a dividir los datos en 4 inputs ya que NetMHCcons tiene una restricción de número de líneas posibles, en total 10000. En el momento de introducir los inputs en NetMHCcons se seleccionaron las siguientes opciones; *Peptide length (several lengths are possible): 9mer peptides, Select Species/loci: HLA supertype representative*. Dado a que no conocemos el tipo de HLA que tienen nuestros péptidos debemos de seleccionar el general y más representativo.

Se procedió a descargar también los datos de predicción de neoantígenos proporcionados por el portal cBioPortal con la intención de poder realizar una comparativa entre ambos resultados.

Una vez obtenidos ambos grupos de datos (los procedentes de NetMHCcons y de cBioPortal), se leyeron en R. Los datos obtenidos a partir de NetMHCcons se depuraron para eliminar los péptidos que no eran neoantígenos y para generar dos nuevos grupos, los neoantígenos con unión al MHC-I débil “WB” (de “weak binding”) y los neoantígenos con unión al MHC-I fuerte “SB” (de “strong binding”.

En la siguiente tabla se pueden observar parcialmente los resultados obtenidos a partir de la herramienta web NetMHCcons. Se puede observar cómo se indica la posición dentro del péptido introducido, el alelo, el péptido, la identidad del paciente, la probabilidad de unión, la afinidad de unión en nanómetros, el rango y el tipo de afinidad de unión.

Tabla 1: Resultados obtenidos a partir de la herramienta NetMHCcons.

pos	Allele	peptide	Identity	X1.log50k.aff.	Affinity.nM.	X.Rank	BindingLevel
0	HLA-A02:01	IQNKKITEV	2672_APOB_A1105	0.464	330.10	4.00	<=WB
0	HLA-A02:01	RLISTLENL	2672_BLM_Y784C	0.671	34.96	1.50	<=SB
4	HLA-A02:01	SLWIGHSFV	2672_COL4A1_Y15	0.730	18.57	0.80	<=SB
8	HLA-A02:01	AITEGLTCV	2672_CTRL_L149I	0.605	72.18	2.00	<=WB
3	HLA-A02:01	YVSDIRCFL	2672_DFFB_R87C	0.635	52.18	2.00	<=WB
6	HLA-A02:01	SLHVIQESL	2672_DMD_L1381V	0.451	379.96	4.00	<=WB
9	HLA-A02:01	VIQESLTFI	2672_DMD_L1381V	0.577	97.73	3.00	<=WB
1	HLA-A02:01	KLQYLFSQA	2672_DRP2_V519A	0.467	319.56	4.00	<=WB
4	HLA-A02:01	YLFSQAANS	2672_DRP2_V519A	0.474	296.25	4.00	<=WB
7	HLA-A02:01	AAYPLIAPA	2672_PHC2_H291Y	0.447	398.91	4.00	<=WB
1	HLA-A02:01	YIAPAKEPL	2672_NR3C2_S722	0.491	246.48	4.00	<=WB
6	HLA-A02:01	NQLEYLTRL	2672_MYO1C_K336	0.623	58.77	2.00	<=WB
0	HLA-A02:01	SMLDVALFM	2672_NINJ2_S45N	0.752	14.55	0.80	<=SB
0	HLA-A02:01	GVLQGSFTV	2672_SLC5A5_M42	0.699	25.97	1.00	<=SB
1	HLA-A02:01	VLQGSFTVV	2672_SLC5A5_M42	0.608	69.50	2.00	<=WB
7	HLA-A02:01	SVMLQVLSL	2672_TCN2_T333M	0.485	263.01	4.00	<=WB
8	HLA-A02:01	VMLQVLSLL	2672_TCN2_T333M	0.675	33.48	1.50	<=SB
9	HLA-A02:01	SLCHFFFNI	2672_SLC34A2_A4	0.746	15.61	0.80	<=SB
5	HLA-A02:01	LLVTDAYEI	2672_TMED10_G56	0.695	27.11	1.00	<=SB
5	HLA-A02:01	AVFIVLEYV	2672_DOLK_F372V	0.661	39.38	1.50	<=SB
4	HLA-A02:01	VMLVLVWKL	2672_DCUN1D4_A1	0.753	14.48	0.80	<=SB
0	HLA-A02:01	ALEALINFA	2672_KLHL18_Y97	0.458	352.24	4.00	<=WB
1	HLA-A02:01	KQLTLAQGV	2672_SMG5_A961V	0.534	154.78	3.00	<=WB
10	HLA-A02:01	YLDNPNALT	2672_KANK2_K603	0.485	263.01	4.00	<=WB

Cabe destacar que se ha obtenido un total de 6455 neoantígenos. De los 223 pacientes, 206 presentan neoantígenos con unión fuerte a MHC-1 y 216 presentan neoantígenos con unión débil a MHC-I.

En las siguiente imagen (Ilustración 7), se puede observar como hay pacientes con una alta carga de neoantígenos, probablemente coincidiendo con aquellos pacientes que tienen elevada carga mutacional (cáncer de colon con inestabilidad de microsatélites).

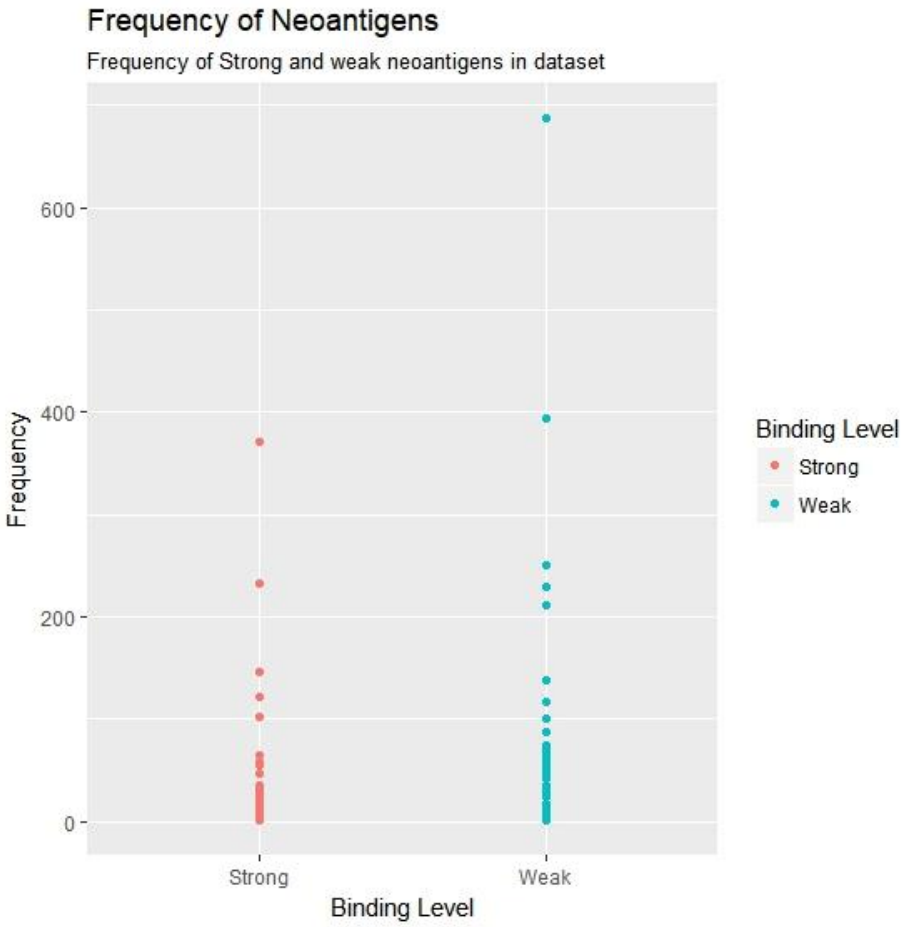


Ilustración 7: Frecuencia de neoantígenos en función al tipo de unión de MHC.

Como se puede comprobar, la cantidad de neoantígenos de unión débil es superior a la cantidad de neoantígenos de unión fuerte (Ilustración 8).

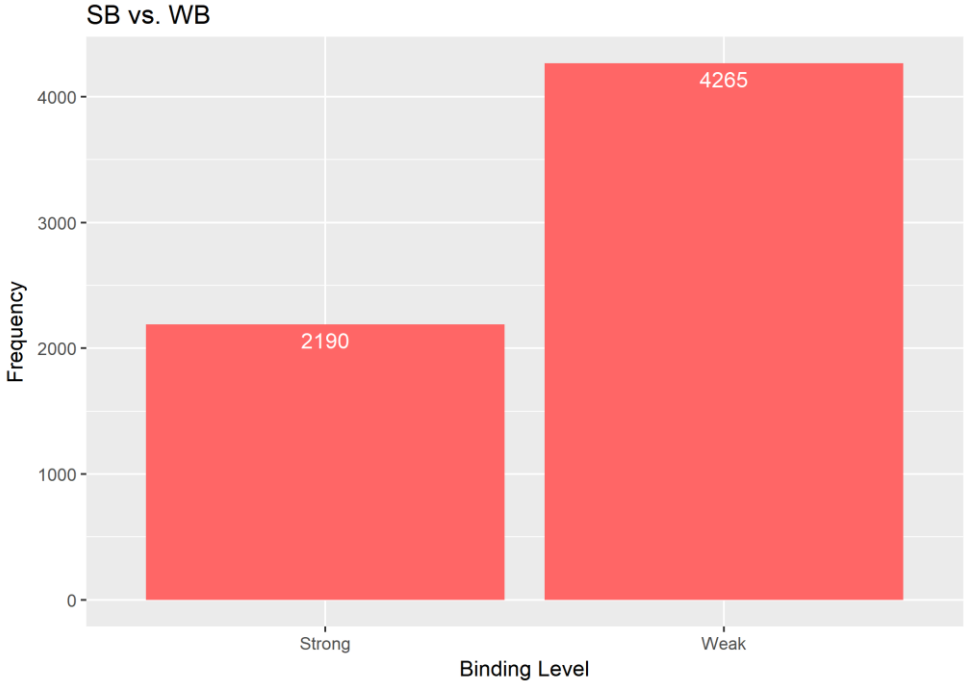


Ilustración 8: Histograma cantidad de neoantígenos.

- Comprobación de los resultados obtenidos

El portal *The Cancer Immunome Atlas* (TCIA) facilita información sobre 20 tipos de cáncer obtenida a través del portal TCGA. Concretamente, facilita información sobre aspectos relacionados con la inmunología del cáncer y, concretamente, una predicción sobre neoantígenos.

Se procedió a realizar una descarga de dichos datos con la finalidad de realizar una comparativa entre dichos datos y los obtenidos en el presente trabajo. Con la librería de RStudio "ggplot2" se generaron diversas gráficas.

3.3. Resultados obtenidos

A continuación, se puede observar gráficamente el resultado de la comparación de los resultados obtenidos en este proyecto (TFM Data) y los obtenidos a partir del portal del TCIA. En total hay 51 pacientes en común entre ambos estudios. Se puede observar que la frecuencia de neoantígenos es más elevada en los obtenidos por el portal TCIA, no obstante, se puede observar que hay una gran correlación (Figura 11) entre ambos tipos de datos, R^2=0.804. La gran diferencia de concentración de neoantígenos en función del paciente puede ser debida a que los métodos de predicción de neoantígenos en este estudio han sido muy estrictos ya que se han realizado con NetMHCcons, que se trata de un consenso entre 3 predicciones diferentes.

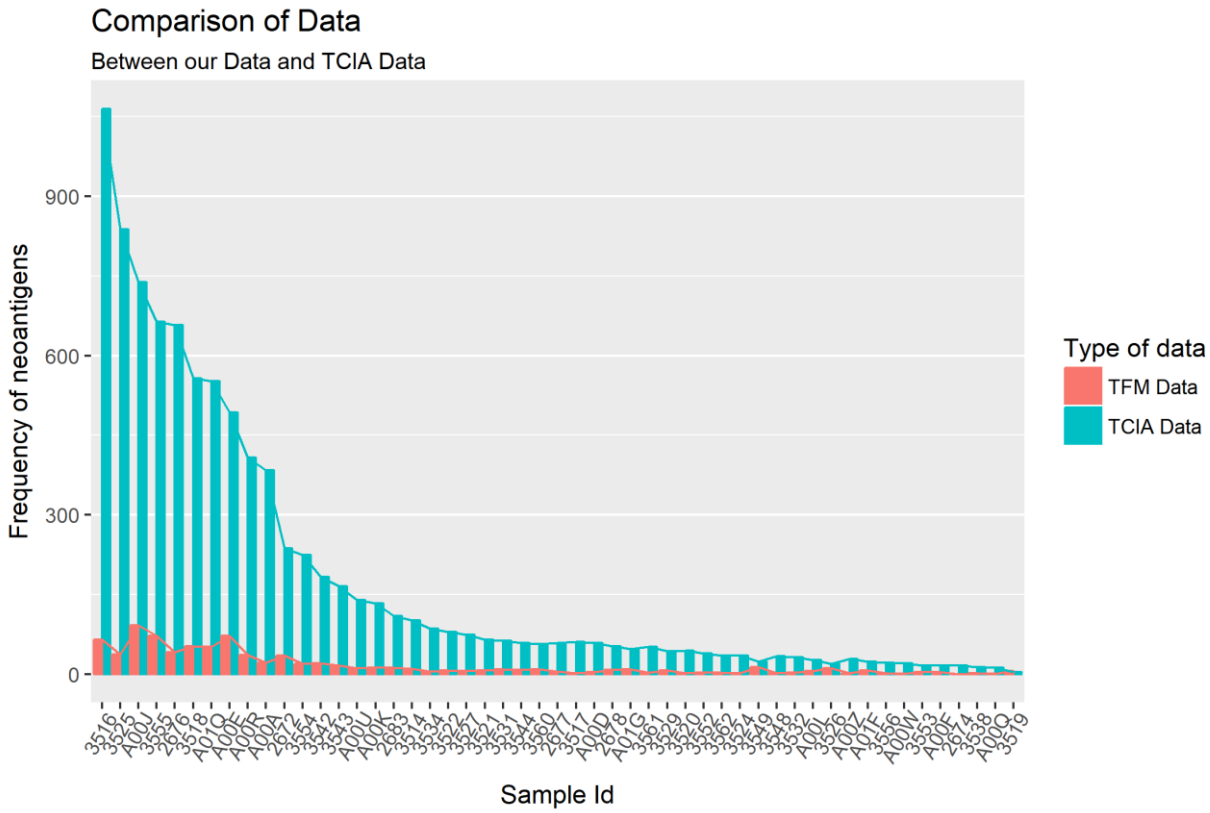


Ilustración 9: Frecuencia de neoantígenos entre los datos obtenidos y los del servidor TCIA.

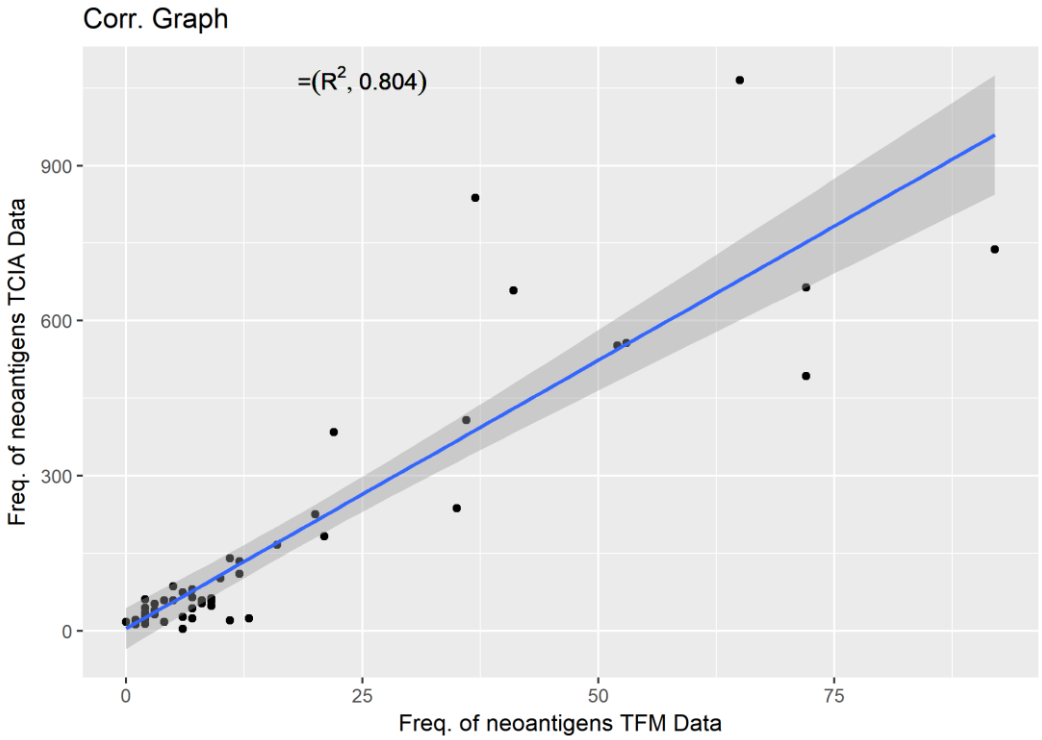


Ilustración 10: Correlación de la frecuencia de neoantígenos entre los datos obtenidos y los del servidor TCIA.

4. Análisis de supervivencia

Para la realización del análisis de supervivencia, fueron necesarios los datos obtenidos a partir de la predicción de neoantígenos ya que dicha predicción se realizó en función de los neoantígenos presentes en cada uno de los pacientes, separándolos por unión a MHC-I débil o fuerte. Se ha realizado un análisis univariante, ya que la única variable que ha sido tenida en cuenta ha sido "cantidad de neoantígenos".

Han sido necesarios librerías de RStudio como 'Survfit' que hace una estimación de la curva de supervivencia para los datos introducidos usando el método de Kaplan-Meier.

A continuación, para la realización de las gráficas de supervivencia se ha usado la librería de 'ggplot2' 'ggsurvplot'.

4.1. Características de los datos

Como ya se ha comentado anteriormente, partimos de los datos obtenidos a partir de la predicción de neoantígenos.

Estos datos se separaron en dos subgrupos según su afinidad con el MHC-I. Un subgrupo con unión débil y otro subgrupo con unión fuerte.

Partimos de un total de 2.190 neoantígenos con unión fuerte y 4.265 neoantígenos con unión débil

Los datos necesarios para realizar el análisis de supervivencia son: el tipo de unión a MHC-I, el promedio de supervivencia en meses y el estado vital.

4.2. Procedimientos y materiales empleados

Análisis bioestadístico descriptivo y de supervivencia

- Procesamiento de datos:

La tarea de procesamiento de datos se realizó con el fin de tener los datos depurados y listos para realizar el análisis bioestadístico y de supervivencia.

Esta preparación de datos consistió en generar las columnas necesarias para realizar el análisis de supervivencia. Uno de los pasos esenciales para realizar con éxito este análisis consiste en poner como variable binaria la ocurrencia o no del evento. Como se trata de un análisis de supervivencia, la ocurrencia del evento se considera que la persona ha fallecido (1) y la no ocurrencia que sigue con vida (0).

Una vez realizada dicha fase, se procedió a realizar el análisis bioestadístico y de supervivencia.

- Análisis bioestadístico y de supervivencia

Para la realización del análisis han sido necesarias diversas herramientas de RStudio.

En primer lugar, se realizó un análisis de supervivencia basado en el estimador de Kaplan-Meier. A continuación se realizaron las gráficas con los resultados obtenidos del anterior análisis.

Se realizaron un seguido de análisis y gráficas de supervivencia:

- Unión débil + unión fuerte: en primer lugar se realizó el estudio con los dos tipos de unión. Considerando un nivel elevado de neoantígenos la mediana de los datos, en concreto 8.
- Unión fuerte: el estudio se realizó considerando únicamente el tipo de unión a MHC-I fuerte. Considerando también un nivel elevado de neoantígenos la mediana de los datos, en concreto 3.
- Unión débil: por último lugar, se realizó el estudio considerando únicamente el tipo de unión a MHC-I débil. Considerando también un nivel elevado de neoantígenos la mediana de los datos, en concreto 5.

4.3. Resultados obtenidos

Los resultados obtenidos son las gráficas que se muestran a continuación:

En la siguiente gráfica se puede observar la probabilidad de supervivencia en función del número de neoantígenos totales presentes en el tumor:

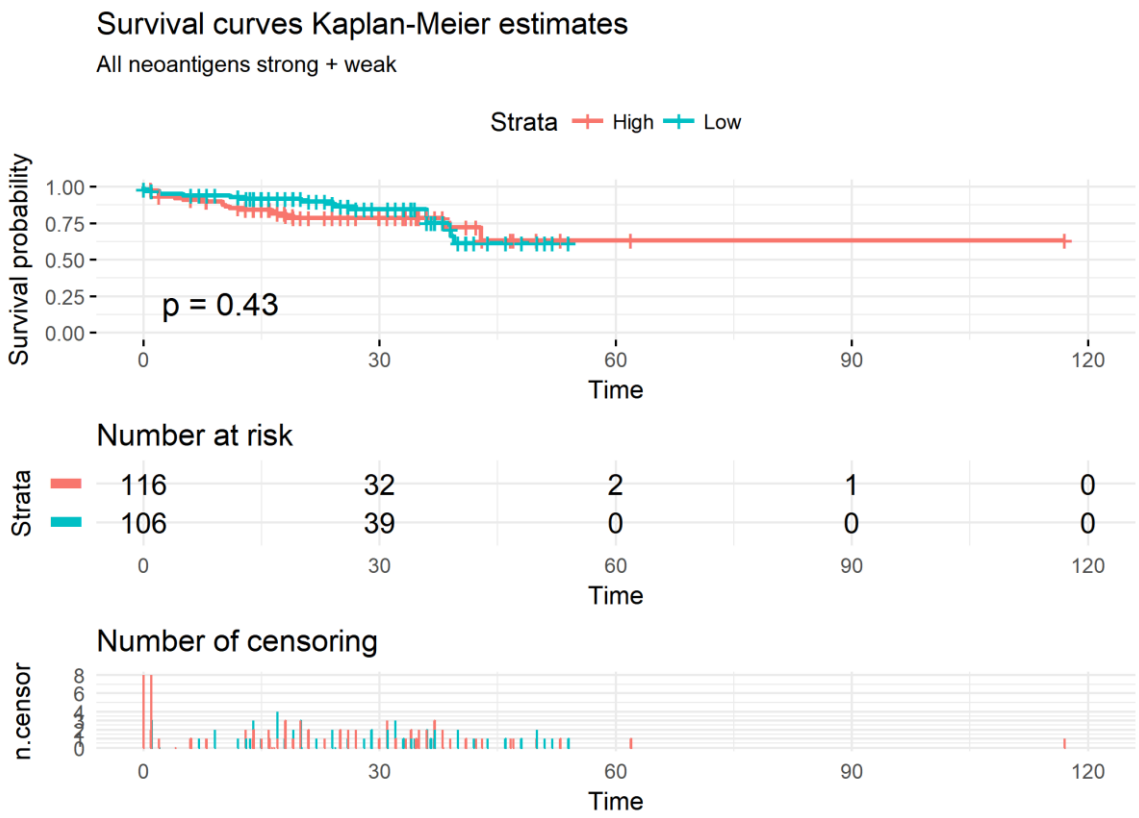


Ilustración 11: Curva de supervivencia neoantígenos totales.

En la siguiente gráfica se puede observar la probabilidad de supervivencia en función del número de neoantígenos con unión fuerte al complejo MHC-I presentes en el tumor:

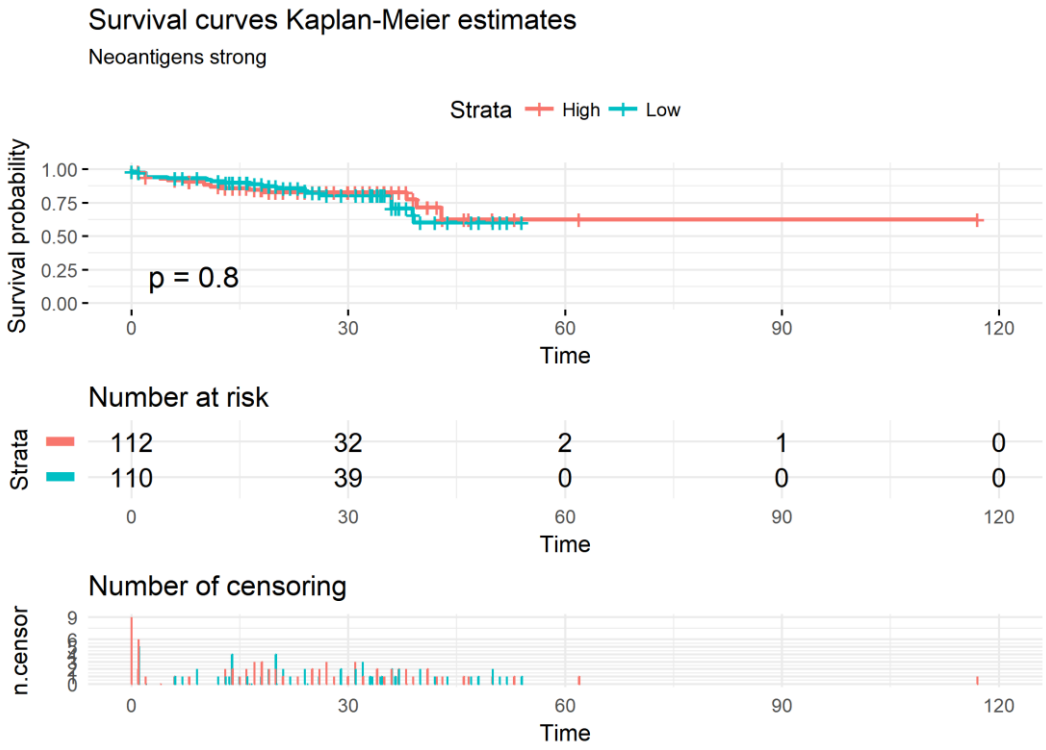


Ilustración 12: Curva de supervivencia neoantígenos unión fuerte.

En la siguiente gráfica se puede observar la probabilidad de supervivencia en función del número de neoantígenos con unión débil al complejo MHC-I presentes en el tumor:

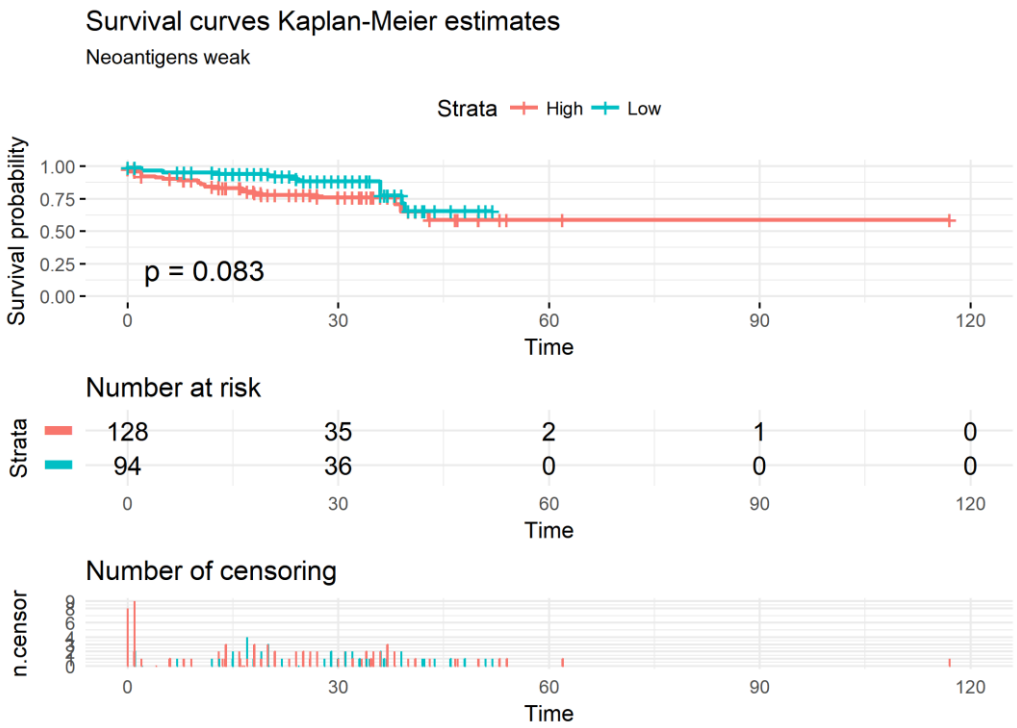


Ilustración 13: Curva de supervivencia neoantígenos unión débil.

Tal y como podemos observar en las gráficas de supervivencia, se puede concluir de que en ninguno de los casos se observa una relación entre el número de neoantígenos y una mayor probabilidad de supervivencia.

5. Análisis final sobre la viabilidad

A continuación se realizará un análisis de viabilidad del proyecto realizado.

Se considera que técnicamente el proyecto es viable ya que tanto el software, los servidores y los datos en los que se ha basado el análisis son de acceso libre y han sido evaluados por personal cualificado.

Dado que se trata de datos de dominio público no existe ningún problema de confidencialidad para poder reproducir dicho estudio.

Para concluir, cabe destacar que la validación del proyecto la ha realizado la Dra. Rebeca Sanz. A partir del script y de los datos de partida ha podido realizar el proyecto con éxito.

6. Conclusiones

En el presente proyecto se ha realizado la identificación de neoantígenos a partir de datos de ultrasecuenciación y, posteriormente se han aplicado dichos neoantígenos para implementar un análisis de supervivencia en pacientes con cáncer colorrectal.

A continuación se listarán las conclusiones extraídas a partir de la realización del proyecto:

1) El procedimiento y la ejecución del script han sido validados externamente asegurando, de este modo, un correcto funcionamiento.

2) Se han cumplido los objetivos propuestos al inicio del proyecto, así como también se ha seguido la planificación y metodología prevista.

3) El método desarrollado en el presente proyecto, es adecuado para la predicción de neoantígenos de una manera fiable y rápida a partir de datos de ultrasecuenciación.

4) Con el método desarrollado se hayan menos neoantígenos que con otros métodos. Esto es debido, muy probablemente a que se utilizan métodos mucho más restrictivos. De hecho, la correlación entre neoantígenos encontrados por ambos métodos es muy alta; siendo los mismos individuos los que presentan una carga de neoantígenos más elevada.

5) En cáncer colorrectal, las mujeres tienen un elevado número de neoantígenos respecto a los hombres. Cabe destacar, que la frecuencia de mujeres que padecen cáncer colorrectal es inferior al número de hombres.

6) No se ha encontrado una relación entre el número de neoantígenos y la supervivencia de pacientes con cáncer colorrectal usando la mediana como punto de corte.

7) Las futuras líneas de investigación se podrían focalizar en la comparación de los neoantígenos entre pacientes y ver si tienen en común. También la caracterización de neoantígenos en otros tipos de cáncer como puede ser el de endometrio.

7. Glosario

Antígeno: substáncia desencadenante de respuesta inmunitaria.

MAF: Mutation Annotation Format

MHC: Major Histocompatibility Complex

mRNA: messenger Ribonucleic Acid

NGS: Next Generation Sequencing

Neoantígneo: antígeno en forma de péptido presentado por el complejo MHC-I procedente de células tumorales.

Rmarkdown: librería de RStudio que permite insertar el código de R en un documento en una amplia variedad de formatos.

RStudio: Software alternativo a R con editor de código y herramientas para la depuración y visualización de datos.

TCGA: The Cancer Genome Atlas

8. Bibliografía

[1] W. S. Klug, *Conceptos de genética*. Madrid : Prentice Hall, 2006.

[2] D. S. Chen and I. Mellman, “Elements of cancer immunity and the cancer-immune set point,” *Nature*, vol. 541, no. 7637, pp. 321–330, 2017.

[3] J. L. Markman and S. L. Shiao, “Impact of the immune system and immunotherapy in colorectal cancer,” *J. Gastrointest. Oncol.*, vol. 6, no. 2, pp. 208–223, 2015.

[4] J. B. Zabriskie, *Essential Clinical Immunology*. 2009.

[5] S. D. Brown, R. L. Warren, E. A. Gibb, S. D. Martin, J. J. Spinelli, B. H. Nelson, and R. A. Holt, “Neo-antigens predicted by tumor genome meta-analysis correlate with increased patient survival,” pp. 743–750, 2014.

[6] H. Hackl, P. Charoentong, F. Finotello, and Z. Trajanoski, “Computational genomics tools for dissecting tumour-immune cell interactions,” *Nat. Rev. Genet.*, vol. 17, no. 8, pp. 441–458, 2016.

[7] K. Tomczak, P. Czerwińska, and M. Wiznerowicz, “The Cancer Genome Atlas (TCGA): An immeasurable source of knowledge,” *Wspolczesna Onkol.*, vol. 1A, pp. A68–A77, 2015.

[8] X. Zhang, P. K. Sharma, S. Peter Goedegebuure, and W. E. Gillanders, “Personalized cancer vaccines: Targeting the cancer mutanome,” *Vaccine*, vol. 35, no. 7, pp. 1094–1100, 2017.

[9] “Comparison of statistical packages.” .

[10] E. Karosiene, C. Lundegaard, O. Lund, and M. Nielsen, “NetMHCcons: A consensus method for the major histocompatibility complex class i predictions,” *Immunogenetics*, vol. 64, no. 3, pp. 177–186, 2012.

[11] and B. J. R. Li Ding, Michael C. Wendl, Joshua F. McMichael, “Expanding the computational toolbox for mining cancer genomes,” *Nat Rev Genet*, vol. 15, no. 8, pp. 556–570, 2014.

[12] K. Wang, M. Li, and H. Hakonarson, “ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data,” *Nucleic Acids Res.*, vol. 38, no. 16, pp. 1–7, 2010.

[13] M. Carlson, “UniProt . ws : A package for retrieving data from the UniProt web service,” *R Packag. version 2.18.0.*, pp. 1–3, 2012.

[14] I. Hoof, B. Peters, J. Sidney, L. E. Pedersen, A. Sette, O. Lund, S. Buus, and M. Nielsen, “NetMHCpan, a method for MHC class i binding prediction beyond humans,” *Immunogenetics*, vol. 61, no. 1, pp. 1–13, 2009.

[15] H. Zhang, O. Lund, and M. Nielsen, “The PickPocket method for predicting binding specificities for receptors based on receptor pocket similarities: Application to MHC-peptide binding,” *Bioinformatics*, vol. 25, no. 10, pp. 1293–1299, 2009.

[16] J. Gao, B. A. Aksoy, U. Dogrusoz, G. Dresdner, B. Gross, S. O. Sumer, Y. Sun, A. Jacobsen, R. Sinha, E. Larsson, E. Cerami, C. Sander, and N. Schultz, “Integrative Analysis of Complex Cancer Genomics and Clinical Profiles Using the cBioPortal.”

[17] E. Cerami, J. Gao, U. Dogrusoz, B. E. Gross, S. O. Sumer, B. A. Aksoy, A. Jacobsen, C. J. Byrne, M. L. Heuer, E. Larsson, Y. Antipin, B. Reva, A. P. Goldberg, C. Sander, and N. Schultz, “The cBio Cancer Genomics Portal: An open platform for exploring multidimensional cancer genomics data,” *Cancer Discov.*, vol. 2, no. 5, pp. 401–404, 2012.

[18] “NCBI - BLAST.” [Online]. Available:

https://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Web&PAGE_TYPE=BlastDocs&DOC_TYPE=BlastHelp. [Accessed: 22-Dec-2017].

9. Anexos

Identificación de neoantígenos a partir de datos de ultrasecuenciación y aplicación posterior en análisis de supervivencia en pacientes con cáncer colorrectal

Marina Gómez Rey

02/01/2018

Contents

Predicción de neoantígenos

En primer lugar, se procede a realizar la carga de librerías que serán necesarias para la realización del proyecto:

```
library(data.table)
library(maftools)
library(ggplot2)
library(stringr)
library(plyr)
library(survminer)
```

```
## Loading required package: ggpubr
```

```
## Loading required package: magrittr
```

```
library(survival)
library(UniProt.ws)
```

```
## Loading required package: RSQLite
```

```
## Loading required package: RCurl
```

```
## Loading required package: bitops
```

Obtención y análisis de datos

Procedemos a cargar los datos clínicos obtenidos a partir del portal cBioPortal.

```
clin<-read.csv("/Volumes/MARINA/Practicas/colon/data_bcr_clinical_data_patient.txt", skip=4, header=T,
               as.is=TRUE, sep="\t") #Cargamos los datos
```

A continuación, realizamos una visualizacion global de los datos:

```r
summary(clin)
```

```
##  OTHER_PATIENT_ID    PATIENT_ID        FORM_COMPLETION_DATE
##  Length:627         Length:627         Length:627
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
##  HISTOLOGICAL_DIAGNOSIS PROSPECTIVE_COLLECTION RETROSPECTIVE_COLLECTION
##  Length:627             Length:627             Length:627
##  Class :character       Class :character       Class :character
##  Mode  :character       Mode  :character       Mode  :character
##
##
##
##     GENDER           DAYS_TO_BIRTH          RACE
##  Length:627         Length:627         Length:627
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
##   ETHNICITY         HISTORY_OTHER_MALIGNANCY HISTORY_NEOADJUVANT_TRTYN
##  Length:627         Length:627               Length:627
##  Class :character   Class :character         Class :character
##  Mode  :character   Mode  :character         Mode  :character
##
##
##
##  INITIAL_PATHOLOGIC_DX_YEAR AJCC_STAGING_EDITION AJCC_TUMOR_PATHOLOGIC_PT
##  Min.   :1998               Length:627           Length:627
##  1st Qu.:2007               Class :character     Class :character
##  Median :2009               Mode  :character     Mode  :character
##  Mean   :2008
##  3rd Qu.:2010
##  Max.   :2013
##  AJCC_NODES_PATHOLOGIC_PN AJCC_METASTASIS_PATHOLOGIC_PM
##  Length:627               Length:627
##  Class :character         Class :character
##  Mode  :character         Mode  :character
##
##
##
##  AJCC_PATHOLOGIC_TUMOR_STAGE RESIDUAL_TUMOR     LYMPH_NODES_EXAMINED
##  Length:627                  Length:627         Length:627
##  Class :character            Class :character   Class :character
##  Mode  :character            Mode  :character   Mode  :character
##
##
##
##  LYMPH_NODE_EXAMINED_COUNT LYMPH_NODES_EXAMINED_HE_COUNT
##  Length:627                Length:627
##  Class :character          Class :character
```

```
##  Mode  :character          Mode  :character
##
##
##
##  LYMPH_NODES_EXAMINED_IHC_COUNT VITAL_STATUS       DAYS_TO_LAST_FOLLOWUP
##  Length:627                     Length:627         Length:627
##  Class :character               Class :character   Class :character
##  Mode  :character               Mode  :character   Mode  :character
##
##
##
##  DAYS_TO_DEATH      TUMOR_STATUS       VASCULAR_INVASION_INDICATOR
##  Length:627         Length:627         Length:627
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
##  LYMPHOVASCULAR_INVASION_INDICATOR PERINEURAL_INVASION
##  Length:627                        Length:627
##  Class :character                  Class :character
##  Mode  :character                  Mode  :character
##
##
##
##  KRAS_GENE_ANALYSIS_INDICATOR KRAS_MUTATION
##  Length:627                   Length:627
##  Class :character             Class :character
##  Mode  :character             Mode  :character
##
##
##
##  BRAF_GENE_ANALYSIS_INDICATOR BRAF_GENE_ANALYSIS_RESULT    WEIGHT
##  Length:627                   Length:627                Length:627
##  Class :character             Class :character          Class :character
##  Mode  :character             Mode  :character          Mode  :character
##
##
##
##     HEIGHT          RADIATION_TREATMENT_ADJUVANT
##  Length:627         Length:627
##  Class :character   Class :character
##  Mode  :character   Mode  :character
##
##
##
##  PHARMACEUTICAL_TX_ADJUVANT TREATMENT_OUTCOME_FIRST_COURSE
##  Length:627                 Length:627
##  Class :character           Class :character
##  Mode  :character           Mode  :character
##
##
##
##  NEW_TUMOR_EVENT_AFTER_INITIAL_TREATMENT      AGE
```

```
##  Length:627                            Min.   :31.00
##  Class :character                      1st Qu.:58.00
##  Mode  :character                      Median :68.00
##                                        Mean   :66.28
##                                        3rd Qu.:76.00
##                                        Max.   :90.00
##  PRIMARY_SITE       CLIN_M_STAGE       CLIN_N_STAGE
##  Length:627         Length:627         Length:627
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
##  CLIN_T_STAGE       CLINICAL_STAGE
##  Length:627         Length:627
##  Class :character   Class :character
##  Mode  :character   Mode  :character
##
##
##
##  DAYS_TO_INITIAL_PATHOLOGIC_DIAGNOSIS DAYS_TO_PATIENT_PROGRESSION_FREE
##  Min.   :0                            Length:627
##  1st Qu.:0                            Class :character
##  Median :0                            Mode  :character
##  Mean   :0
##  3rd Qu.:0
##  Max.   :0
##  DAYS_TO_TUMOR_PROGRESSION DISEASE_CODE       EXTRANODAL_INVOLVEMENT
##  Length:627                Length:627         Length:627
##  Class :character          Class :character   Class :character
##  Mode  :character          Mode  :character   Mode  :character
##
##
##
##     ICD_10          ICD_O_3_HISTOLOGY  ICD_O_3_SITE
##  Length:627         Length:627         Length:627
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
##  INFORMED_CONSENT_VERIFIED INITIAL_PATHOLOGIC_DIAGNOSIS_METHOD
##  Length:627                Length:627
##  Class :character          Class :character
##  Mode  :character          Mode  :character
##
##
##
##  PROJECT_CODE       STAGE_OTHER        TISSUE_SOURCE_SITE
##  Length:627         Length:627         Length:627
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
```

```
##
##  TUMOR_TISSUE_SITE   OS_STATUS          OS_MONTHS
##  Length:627         Length:627         Length:627
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
##   DFS_STATUS         DFS_MONTHS
##  Length:627         Length:627
##  Class :character   Class :character
##  Mode  :character   Mode  :character
##
##
##
```

```
dim(clin)
```

```
## [1] 627  64
```

Procederemos a listar el contenido de la carpeta de mutaciones. Debemos de tener en cuenta que estos archivos se encuentran en '.maf' y, por lo tanto, tendremos que hacer servir la herramienta maftools de bioconductor

```
files <- list.files("/Volumes/MARINA/Practicas/colon/mut3/")
length(files)
```

```
## [1] 223
```

```
#la renombramos
id <- files;
#substraemos los que contienen "tcga" #223
id<- substring(id, regexpr("tcga", id));
#Seleccionamos el nombre de la posicion 1 a la posición 12 del nombre de archivo
id<-substr(id,1,12)
```

A continuación, generaremos una tabla para visualizar que cantidad de archivos de mutaciones contienen información clínica, debemos de tenerlo en cuenta para realizar en el último punto el análisis de supervivencia, ya que necesitaremos conocer: tiempo de seguimiento, supervivencia, edad, sexo y la ID.

```
table(id%in%clin$PATIENT_ID)
```

```
##
## TRUE
##  223
```

Procederemos a obtener los archivos de mutaciones:

```
setwd("/Volumes/MARINA/Practicas/colon/mut3/")
dat <- do.call("rbind", lapply(files, fread, header = TRUE, sep="\t", skip=3));
setDF(dat)
dim(dat)#84605
```

```
## [1] 84605   300
```

```
mut<-table(dat$Tumor_Sample_Barcode);
mut<-as.data.frame(mut) #223
id2<-substring(mut$Var1, regexpr("TCGA", id));
id2<-substr(id2, 1, 12);
```

```
mut$id<-id2
```

Visualizaremos el nombre de las columnas de información clínica para escoger los de interés:

```
colnames(clin)
```

```
##  [1] "OTHER_PATIENT_ID"
##  [2] "PATIENT_ID"
##  [3] "FORM_COMPLETION_DATE"
##  [4] "HISTOLOGICAL_DIAGNOSIS"
##  [5] "PROSPECTIVE_COLLECTION"
##  [6] "RETROSPECTIVE_COLLECTION"
##  [7] "GENDER"
##  [8] "DAYS_TO_BIRTH"
##  [9] "RACE"
## [10] "ETHNICITY"
## [11] "HISTORY_OTHER_MALIGNANCY"
## [12] "HISTORY_NEOADJUVANT_TRTYN"
## [13] "INITIAL_PATHOLOGIC_DX_YEAR"
## [14] "AJCC_STAGING_EDITION"
## [15] "AJCC_TUMOR_PATHOLOGIC_PT"
## [16] "AJCC_NODES_PATHOLOGIC_PN"
## [17] "AJCC_METASTASIS_PATHOLOGIC_PM"
## [18] "AJCC_PATHOLOGIC_TUMOR_STAGE"
## [19] "RESIDUAL_TUMOR"
## [20] "LYMPH_NODES_EXAMINED"
## [21] "LYMPH_NODE_EXAMINED_COUNT"
## [22] "LYMPH_NODES_EXAMINED_HE_COUNT"
## [23] "LYMPH_NODES_EXAMINED_IHC_COUNT"
## [24] "VITAL_STATUS"
## [25] "DAYS_TO_LAST_FOLLOWUP"
## [26] "DAYS_TO_DEATH"
## [27] "TUMOR_STATUS"
## [28] "VASCULAR_INVASION_INDICATOR"
## [29] "LYMPHOVASCULAR_INVASION_INDICATOR"
## [30] "PERINEURAL_INVASION"
## [31] "KRAS_GENE_ANALYSIS_INDICATOR"
## [32] "KRAS_MUTATION"
## [33] "BRAF_GENE_ANALYSIS_INDICATOR"
## [34] "BRAF_GENE_ANALYSIS_RESULT"
## [35] "WEIGHT"
## [36] "HEIGHT"
## [37] "RADIATION_TREATMENT_ADJUVANT"
## [38] "PHARMACEUTICAL_TX_ADJUVANT"
## [39] "TREATMENT_OUTCOME_FIRST_COURSE"
## [40] "NEW_TUMOR_EVENT_AFTER_INITIAL_TREATMENT"
## [41] "AGE"
## [42] "PRIMARY_SITE"
## [43] "CLIN_M_STAGE"
## [44] "CLIN_N_STAGE"
## [45] "CLIN_T_STAGE"
## [46] "CLINICAL_STAGE"
## [47] "DAYS_TO_INITIAL_PATHOLOGIC_DIAGNOSIS"
## [48] "DAYS_TO_PATIENT_PROGRESSION_FREE"
## [49] "DAYS_TO_TUMOR_PROGRESSION"
```

```
## [50] "DISEASE_CODE"
## [51] "EXTRANODAL_INVOLVEMENT"
## [52] "ICD_10"
## [53] "ICD_O_3_HISTOLOGY"
## [54] "ICD_O_3_SITE"
## [55] "INFORMED_CONSENT_VERIFIED"
## [56] "INITIAL_PATHOLOGIC_DIAGNOSIS_METHOD"
## [57] "PROJECT_CODE"
## [58] "STAGE_OTHER"
## [59] "TISSUE_SOURCE_SITE"
## [60] "TUMOR_TISSUE_SITE"
## [61] "OS_STATUS"
## [62] "OS_MONTHS"
## [63] "DFS_STATUS"
## [64] "DFS_MONTHS"
```

```
mut2<-merge(mut, clin[c("PATIENT_ID", "GENDER", "VITAL_STATUS", "AGE",
                        "DFS_STATUS", "OS_STATUS", "OS_MONTHS",
                        "HISTOLOGICAL_DIAGNOSIS")],
            by.x="id", by.y="PATIENT_ID", all.x=FALSE, all.y=FALSE)
```

Guardamos la tabla, para tenera lista para el último punto de análisis de supervivencia.

```
save(mut2, file="/Volumes/MARINA/TFM/clincolon.Rdata")
write.table(mut2, "/Volumes/MARINA/TFM/clincolon.txt", sep="\t")
```

Visualización de datos obtenidos:

Procedemos a preparar los datos para poder realizar las gráficas:

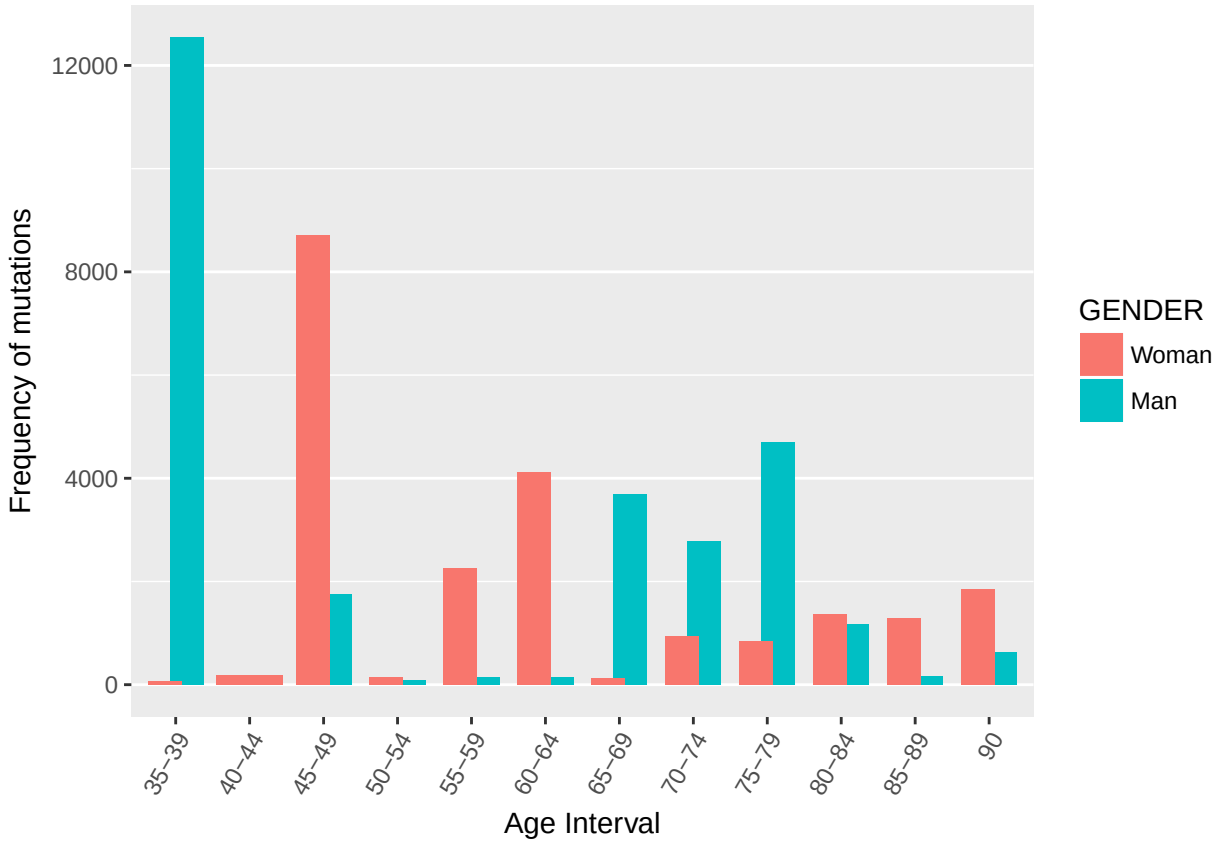
```
mutat<-mut2[!mut2$HISTOLOGICAL_DIAGNOSIS=="[Not Available]",]


saltos_edad <- c(35,40,45,50,55,60,65,70,75,80,85,90,95)
grupos_edad <- c("35-39","40-44","45-49","50-54","55-59","60-64","65-69",
                 "70-74","75-79","80-84","85-89","90")

setDT(mutat)[ , agegroups := cut(AGE,
                                 breaks = saltos_edad,
                                 right = FALSE,
                                 labels = grupos_edad)]

colnames(mutat)
```

```
##  [1] "id"                    "Var1"
##  [3] "Freq"                  "GENDER"
##  [5] "VITAL_STATUS"          "AGE"
##  [7] "DFS_STATUS"            "OS_STATUS"
##  [9] "OS_MONTHS"             "HISTOLOGICAL_DIAGNOSIS"
## [11] "agegroups"
```

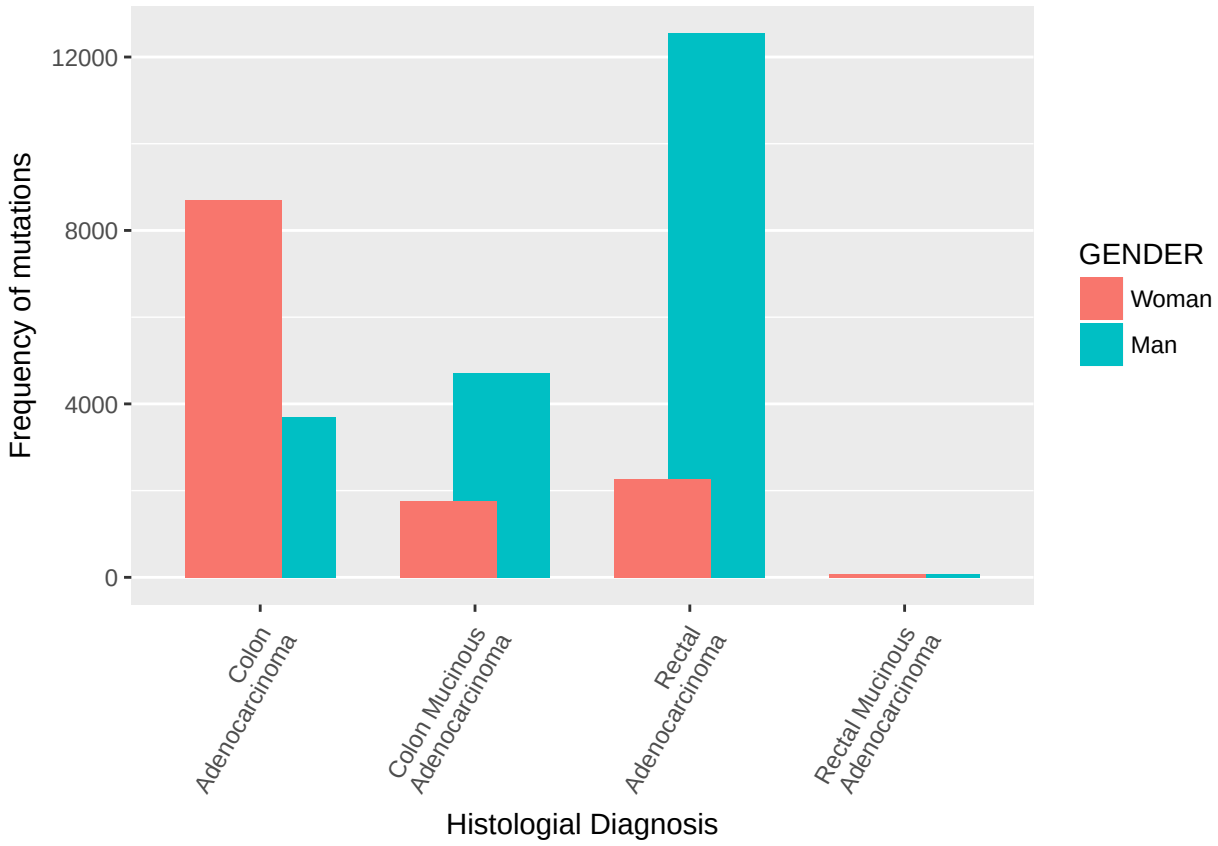
```
colnames(mutat)[11]<-"Age_Interval"


hist_labels<-c("Colon \nAdenocarcinoma", "Colon Mucinous \nAdenocarcinoma",
          "Rectal \nAdenocarcinoma", "Rectal Mucinous \nAdenocarcinoma")
```

Realizamos las gráficas con el fin de poder visualizar los datos:

```
ggplot(mutat, aes(x=Age_Interval, y=Freq, fill=GENDER)) +
  geom_bar(stat="identity", position=position_dodge(width=0.6)) +
  scale_fill_discrete(breaks=c("FEMALE", "MALE"),labels=c("Woman", "Man"))+
  theme(axis.text.x = element_text(angle=60, hjust=1),
        panel.grid.major.x = element_blank()) +
  labs(x = "Age Interval", y="Frequency of mutations")
```




```
ggplot(mutat, aes(x=HISTOLOGICAL_DIAGNOSIS, y=Freq, fill=GENDER)) +
  geom_bar(stat="identity", position=position_dodge(width=0.5)) +
  scale_x_discrete(labels= hist_labels) +
  scale_fill_discrete(breaks=c("FEMALE", "MALE"),labels=c("Woman", "Man"))+
  theme(axis.text.x = element_text(angle=60, hjust=1),
      panel.grid.major.x = element_blank()) +
  labs(x = "Histologial Diagnosis", y="Frequency of mutations")
```



Selección de datos

Debido a que tenemos información clínica sobre todas las mutaciones, procederemos a obtener los datos de interés:

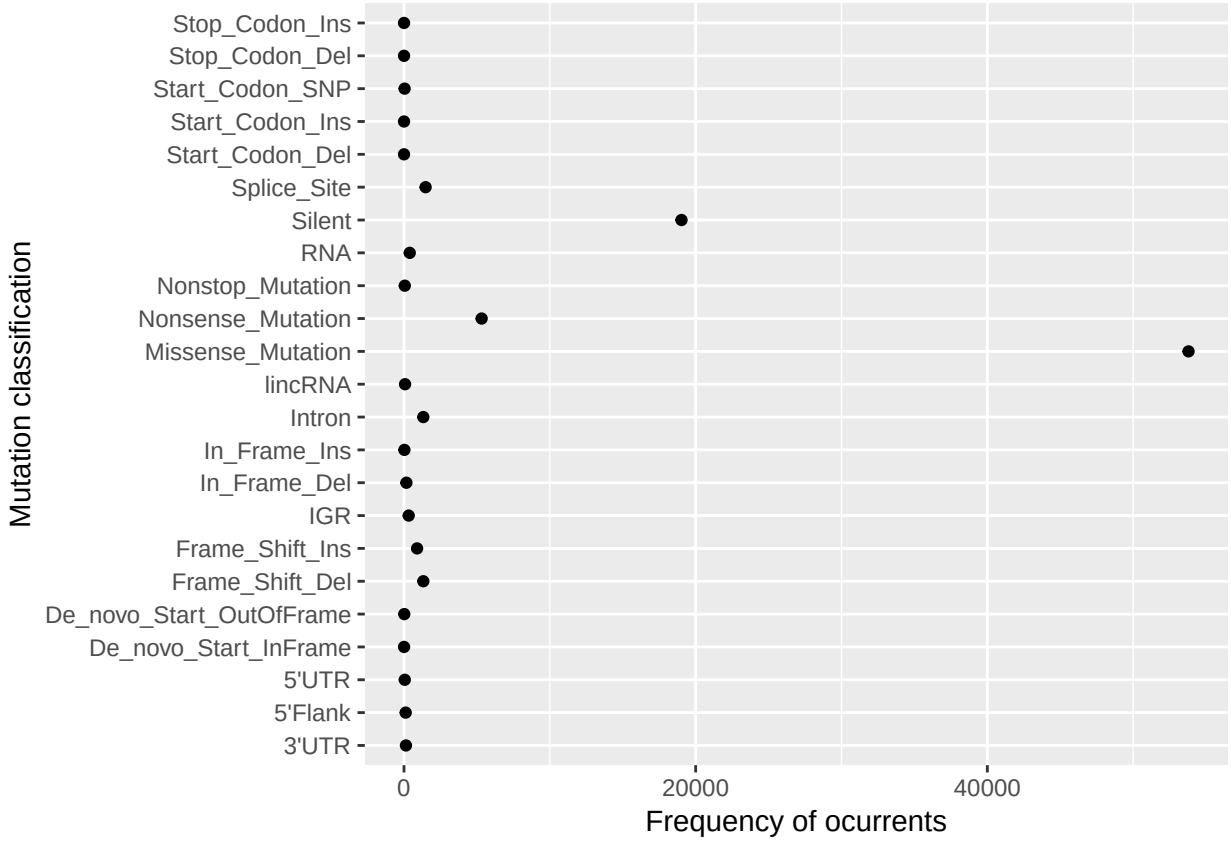
```
table(id%in%clin$PATIENT_ID)
```

```
##
## TRUE
##  223
```

```
in.col <- dat[,c("Hugo_Symbol", "Entrez_Gene_Id","Chromosome","Start_position",
                 "End_position","Variant_Classification","Variant_Type",
                 "Reference_Allele", "Tumor_Seq_Allele2",
                 "Tumor_Sample_Barcode")]
```

A continuación se procede a generar una gráfica con el fin de visualizar el tipo de mutaciones:

```
visual<-table(dat$Variant_Classification)
visual<-as.data.frame(visual)


ggplot(visual, aes(x=Freq, y=Var1)) +
  geom_point() +
  labs(x="Frequency of ocurrents", y="Mutation classification")
```



Seleccionamos las mutaciones de interés

```
mis<-in.col[in.col$Variant_Classification=="Missense_Mutation",]#53796
dim(mis)
```

```
## [1] 53796    10
```

```
table(mis$Variant_Classification)#53796
```

```
##
## Missense_Mutation
##             53796
```

```
save(mis, file="/Volumes/MARINA/TFM/datos_missense.Rdata")
```

Procesamiento de datos

Generamos el input para wannovar

```
input.wa.col <- mis[,c("Chromosome","Start_position","End_position",
                       "Reference_Allele","Tumor_Seq_Allele2")];
write.table(input.wa.col, "/Volumes/MARINA/TFM/input.wa.col.txt", col.names=FALSE,
            row.names=FALSE, sep="\t", quote=FALSE)
```

Una vez obtenidos los datos del wAnnovar, se procede a leerlos:

```
out.wa.col <- read.csv("/Volumes/MARINA/Practicas/datos/exome_summary.csv", header=TRUE);
setDF(out.wa.col)
dim(out.wa.col)
```

```
## [1] 53796    70
```

Comprovamos el numero que coincide con el nombre del gen i el inicio de posición:

```
table(out.wa.col$Gene.wgEncodeGencodeCompV19==mis$Hugo_Symbol);
```

```
##
## FALSE  TRUE
##   860 52936
```

```
table(out.wa.col$Start==mis$Start_position)
```

```
##
##  TRUE
## 53796
```

```
data <- cbind(out.wa.col[,c(1:6,10)], mis)

colnames(data)
```

```
##  [1] "Chr"                              "Start"
##  [3] "End"                              "Ref"
##  [5] "Alt"                              "Func.wgEncodeGencodeCompV19"
##  [7] "AAChange.wgEncodeGencodeCompV19" "Hugo_Symbol"
##  [9] "Entrez_Gene_Id"                   "Chromosome"
## [11] "Start_position"                   "End_position"
## [13] "Variant_Classification"           "Variant_Type"
## [15] "Reference_Allele"                 "Tumor_Seq_Allele2"
## [17] "Tumor_Sample_Barcode"
```

```
#Cambiamos nombre de dos columnas:
colnames(data)[6:7] <- c("gene","AAChange")
```

Procedemos a depurar la columna de "aa_change"

```
#Otorgamos valor a inicio
inicio <- regexpr("p.",data$AAChange)
#Otorgamos valor a final
final <- regexpr(",",data$AAChange)
final[!final==-1] <- final[!final==-1] -1
final[final==-1] <- 1000
#Generamos columnas para procesar la obtención de AAchange
data$aa_change_limpio <- substring(data$AAChange, inicio,final)
data$aa_change_l <- gsub("p.","", data$aa_change_limpio)


inicio <- 1
final <- regexpr(";",data$aa_change_l)
final[!final==-1] <- final[!final==-1] -1
final[final==-1] <- 1000
data$aa_change_l2 <- substring(data$aa_change_l, inicio,final)

#Limpiamos las columnas que no deseamos
data <- data[, -7]
data<-data[,-18]
data<-data[,-17]

#Visualizamos la columna de interés:
```

```
head(data$aa_change_l2)
```

```
## [1] "N305S" "R122Q" "F418S" "I279S" "V28E"  "R55Q"
```

Obtención de la secuencia:

Procederemos a la obtención de la secuencia con la herramienta Uniprot.ws:

```
human_prot <- UniProt.ws(taxId=9606); #TaxId 9606=human
species(human_prot)

#Seleccionamos los datos de humano y Entrez_gene
if(interactive()){
  egs = keys(human_prot, "ENTREZ_GENE")
}

mis_prot <- select(human_prot, keys = c(prot),
                   columns = c("UNIGENE","SEQUENCE"),
                   keytype = "ENTREZ_GENE");

list_prot_colon<-mis_prot
dim(list_prot_colon)

save(list_prot_colon, file="list_prot_colon.Rdata")
write.table(list_prot_colon, "list_prot_colon.txt", col.names=TRUE,
            row.names=FALSE, sep="\t", quote=FALSE)



mis_prot$index <- as.numeric(row.names(mis_prot)) #Ordenamos por row number
mis_prot_2<-mis_prot[order(mis_prot$index), ]
save(mis_prot_2, file="list_prot_colondesorde.Rdata")


mis_prot_3<-format(mis_prot_2$index, nsmall=0)
mis_prot_3<-as.data.frame(mis_prot_3)
mis_prot_3<-unique(mis_prot[order(mis_prot$index), ]) #Seleccionamos los unique

save(mis_prot_3, file="list_prot_unique.Rdata")


names(mis_prot_3)[1]<-paste("Entrez_Gene_Id")
data_seq<-merge(x = mis_prot_3, y = data, by = "Entrez_Gene_Id", all.x = TRUE)
save(data_seq, file="prot_colon.Rdata")
```

Obtención de neoantígenos

Se realiza un merge de la información mutacional y las secuencias obtenidas:

```
load("/Volumes/MARINA/28/list_prot_colondesorde.Rdata")
dim(data)
```

```
## [1] 53796    17
```

```r
dim(mis_prot_2);
```

```
## [1] 47384     4
```

```r
sec <- mis_prot_2;

super <- merge(data, sec, by="Entrez_Gene_Id",
               all.x=TRUE)
dim(super)
```

```
## [1] 235323     20
```

```r
save(super, file="mutcolonmerge.Rdata")


super2<-super[complete.cases(super),]
dim(super2)
```

```
## [1] 203552     20
```

```r
super2["id"]<-substr(super2$Tumor_Sample_Barcode,1,12)


save(super2, file="mutcolonmerge2.Rdata")
```

Separación de la columna de cambio de aminoácido:

A continuación, se procederá a separar el cambio de aminoácido en diferentescolumnas.

```r
#En primer lugar seleccionamos los aminoacidos
super2["aa_only"]<-(gsub("[^a-zA-Z]+", "", super2$aa_change_l2))
#Seleccionamos el aminoácido original
super2["aa_original"]<- substr(super2$aa_only, 1,1)##
#Seleccionamos el aminoácido mutado
super2["aa_mutado"]<- substr(super2$aa_only, 2,2)
#Seleccionamos la posición donde se encuentra
super2["aa_position"] <- as.numeric(str_extract(super2$aa_change_l2, "[0-9]+"))
#Creamos columna para generar la secuencia
super2["newseq"]<- super2$SEQUENCE
```

Generación de la función en forma de bucle para la obtención de neoantígenos:

Se procede a generar la función para la obtención de los neoantígenos.

```r
pos<-super2$aa_position
#Generamos un valor para la columna donde se introducirá el péptido
super2$mutseq <- "kk"
#Generamos el bucle:
for(i in  1:NROW(super2))#para cada fila en la data.frame super 2
{
  print(i)
  if(substr(super2$SEQUENCE, pos,pos)[i]==super2$aa_original[i])  {
    substr(super2$SEQUENCE, pos, pos)[i] <- super2$aa_mutado[i]
    super2$mutseq[i] <-substr(super2$SEQUENCE, pos-9, pos+9)[i]

  }  else {
```

```
    super2$mutseq[i] <- NA
  }
}

#Renombramos el archivo y lo guardamos
super3<-super2
head(super3)
write.table(super3, "peptidos.txt", col.names=TRUE, row.names=FALSE, sep="\t",
            quote=FALSE)
save(super3, file="peptidos.Rdata")
```

A continuación, se procede a la generación del input para la herramienta netMHCcons:

```
super2<-read.csv("/Volumes/MARINA/28/peptidos.txt",  header=T,
               as.is=TRUE, sep="\t")
dim(super2); # 204644
```

```
## [1] 203519     27
```

```
#Eliminamos los duplicados que son de una misma mutación
p <- unique(super2);

dim(p); # 199459
```

```
## [1] 199459     27
```

```
# Ordenamos por paciente:
pp <- p[order(p$id), ];

# generar id con toda la info
pp$tcga_corto <-substr(pp$id,9,12)
pp$fasta <- paste(">",pp$tcga_corto,"_",pp$Hugo_Symbol, "_",
                  pp$aa_change_12, sep="")


# crear input para el NetMHCCons
aux <- pp[,c("fasta","mutseq")]; # 199459
aux <- aux[!is.na(aux$mutseq),]; # 102246
aux <- unique(aux); #  17279

input_MHC <- as.vector(t(aux))
input_MHC <- as.data.frame(input_MHC)
```

Dado que la herramienta tiene límite de dimensión de archivo, se procede a generar 4 inputs:

```
input1<-as.data.frame(input_MHC[1:10000,1])
input2<-as.data.frame(input_MHC[10001:20000,1])
input3<-as.data.frame(input_MHC[20001:30000,1])
input4<-as.data.frame(input_MHC[30001:34558,1])
```

Se procede a guardar los archivos para introducirlos en el servidor. Se debe indicar 'quote=FALSE' ya que de lo contrario, el archivo .txt contendría comillas y el programa NetMHCcons lo consideraría todo como una misma secuencia.

```
write.table(input1, "input1.txt", col.names=FALSE, row.names=FALSE, quote=FALSE)
write.table(input2, "input2.txt", col.names=FALSE, row.names=FALSE, quote=FALSE)
```

```
write.table(input3, "input3.txt", col.names=FALSE, row.names=FALSE, quote=FALSE)
write.table(input4, "input4.txt", col.names=FALSE, row.names=FALSE, quote=FALSE)
```

Una vez obtenidos los datos, procedemos a leerlos:

```
net1 <- read.table("/Volumes/MARINA/Supervivencia/output_MHC/NetMHCcons__1_tab.txt",
                   header=TRUE, as.is=TRUE, sep="\t")

net2 <- read.table("/Volumes/MARINA/Supervivencia/output_MHC/NetMHCcons__2_tab.txt",
                   header=TRUE, as.is=TRUE, sep="\t")

net3 <- read.table("/Volumes/MARINA/Supervivencia/output_MHC/NetMHCcons__3_tab.txt",
                   header=TRUE, as.is=TRUE, sep="\t")

net4 <- read.table("/Volumes/MARINA/Supervivencia/output_MHC/NetMHCcons__4_tab.txt",
                   header=TRUE, as.is=TRUE, sep="\t")

nrow(net1)+nrow(net2)+nrow(net3)+nrow(net4)
```

```
## [1] 6455
```

```
netmhc<- rbind(net1,net2,net3,net4)
netmhc$X<-NULL
dim(netmhc)
```

```
## [1] 6455    8
```

Se procede a guardar los datos:

```
save(netmhc, file="/Volumes/MARINA/Supervivencia/netmhc.Rdata")
```

Comparación entre diferentes resultados

Una vez obtenidos los datos, realizaremos una comparación entre los datos obtenidos en este proyecto y los datos generados por el portal del TCIA.

Se procede a cargar los datos proporcionados por el portal del TCIA y a depurar las columnas:

```
tcia<-read.table(file = '/Volumes/MARINA/Supervivencia/neoantigensAll.tsv',
                 sep = '\t', header = TRUE)
dim(tcia) #77595
```

```
## [1] 77595     5
```

```
colnames(tcia)
```

```
## [1] "patientBarcode" "disease"        "gene"           "peptide"
## [5] "HLA.alleles"
```

```
head(tcia)
```

```
##   patientBarcode disease gene     peptide HLA.alleles
## 1   TCGA-CA-6717     CRC A1CF  YAAKYTLQTL          NA
## 2   TCGA-CA-6717     CRC A1CF FVDEAKTYAAK          NA
## 3   TCGA-G4-6302     CRC A1CF    KLYDILPR          NA
## 4   TCGA-G4-6302     CRC A1CF   KLYDILPRM          NA
## 5   TCGA-G4-6302     CRC A1CF   YDILPRMEL          NA
## 6   TCGA-G4-6302     CRC A1CF   LPRMELTPM          NA
```

```
tcia<-as.data.frame(table(tcia$patientBarcode))
head(tcia)
```

```
##           Var1 Freq
## 1 TCGA-A6-2671   25
## 2 TCGA-A6-2672  237
## 3 TCGA-A6-2674   17
## 4 TCGA-A6-2675   39
## 5 TCGA-A6-2676  658
## 6 TCGA-A6-2677   59
```

```
tcia$Var1<- substr(tcia$Var1, 9,12)
colnames(tcia)<-c("id2", "Freq2")
tcia[, "type"] <- "tcia"
```

Se procede a depurar los datos obtenidos a partir del portal del netMHCcons:

```
#En primer lugar, procedemos a arreglar la tabla:
mutcolon<-mut2
netmhc$BindingLevel <- gsub("<=","",netmhc$BindingLevel)


strong <- netmhc[netmhc$BindingLevel=="SB",];
strong$id <- substring(strong$Identity, 1,4)
n_mut_ind <- as.data.frame(table(strong$id))
colnames(n_mut_ind) <- c("id","neoant_strong")
dim(n_mut_ind)#206
```

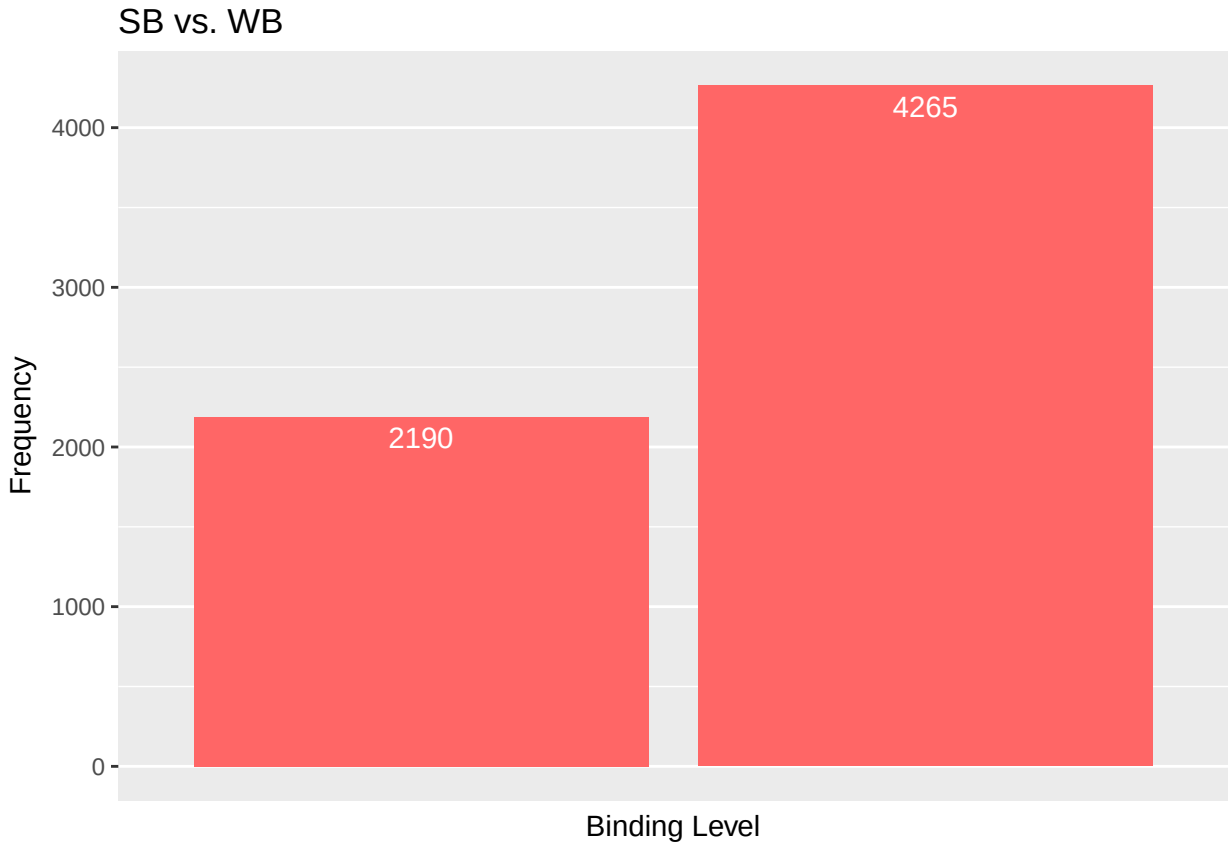
```
## [1] 206   2
```

```
weak <- netmhc[netmhc$BindingLevel=="WB",];
weak$id <- substring(weak$Identity, 1,4)
n_mut_ind_w <- as.data.frame(table(weak$id))
colnames(n_mut_ind_w) <- c("id","neoant_weak")
dim(n_mut_ind_w)#216
```

```
## [1] 216   2
```

Se realiza una gráfica para la observación de los datos:

```
tmhc<-as.data.frame(table(netmhc$BindingLevel))

colnames(tmhc)<-c("Bindinglevel", "Freq")

ggplot(tmhc, aes(x=Bindinglevel, y=Freq))+
  geom_bar(stat="identity", fill="#FF6666") +
  labs(title = "SB vs. WB",
     x = "Binding Level", y="Frequency")+
  scale_x_discrete(breaks=c("<=SB", "<=WB"),
                        labels=c("Strong", "Weak"))+
  geom_text(aes(label=Freq), vjust=1.5, colour="white")
```



A continuación, se procede a depurar los datos de información clínica:

```
clincolon<-mut2
clincolon["id2"]<- substr(clincolon$id, 9,12)
clincolon$id<-NULL
clincolon$Var1<-NULL

head(clincolon)
```

```
##   Freq GENDER VITAL_STATUS AGE          DFS_STATUS OS_STATUS OS_MONTHS
## 1  600 FEMALE        Alive  82         DiseaseFree    LIVING     46.62
## 2   31   MALE        Alive  71 Recurred/Progressed    LIVING     43.73
## 3  839 FEMALE         Dead  75     [Not Available]  DECEASED     42.87
## 4   78 FEMALE         Dead  68     [Not Available]  DECEASED     24.31
## 5   69 FEMALE        Alive  43         DiseaseFree    LIVING     42.25
## 6  107 FEMALE         Dead  57 Recurred/Progressed  DECEASED     16.56
##          HISTOLOGICAL_DIAGNOSIS  id2
## 1          Colon Adenocarcinoma 2672
## 2 Colon Mucinous Adenocarcinoma 2674
## 3          Colon Adenocarcinoma 2676
## 4          Colon Adenocarcinoma 2677
## 5          Colon Adenocarcinoma 2678
## 6          Colon Adenocarcinoma 2683
```

Se realiza el el merge de neoantígenos con la información clínica en este caso, se debe poner 'all.y=TRUE' para no perder ningún paciente:

```
clin_s<-merge(n_mut_ind, clincolon,
              by.x="id", by.y="id2", all.x=FALSE,all.y=TRUE)
```

```r
dim(clin_s)#223
```

```
## [1] 223  10
```

```r
colnames(n_mut_ind_w)
```

```
## [1] "id"          "neoant_weak"
```

```r
colnames(n_mut_ind_w)[1] <- "id2"
```

Se debe indicar 'all.x=TRUE' para no perder ningún dato y mantener los 223 ya que habrá pacientes que tienen mutaciones con unián a MHC-I débiles pero no fuertes y al contrario:

```r
datos_sup<-merge(clin_s, n_mut_ind_w,
                 by.x="id", by.y="id2", all.x=TRUE, all.y=FALSE)
dim(datos_sup)#223
```

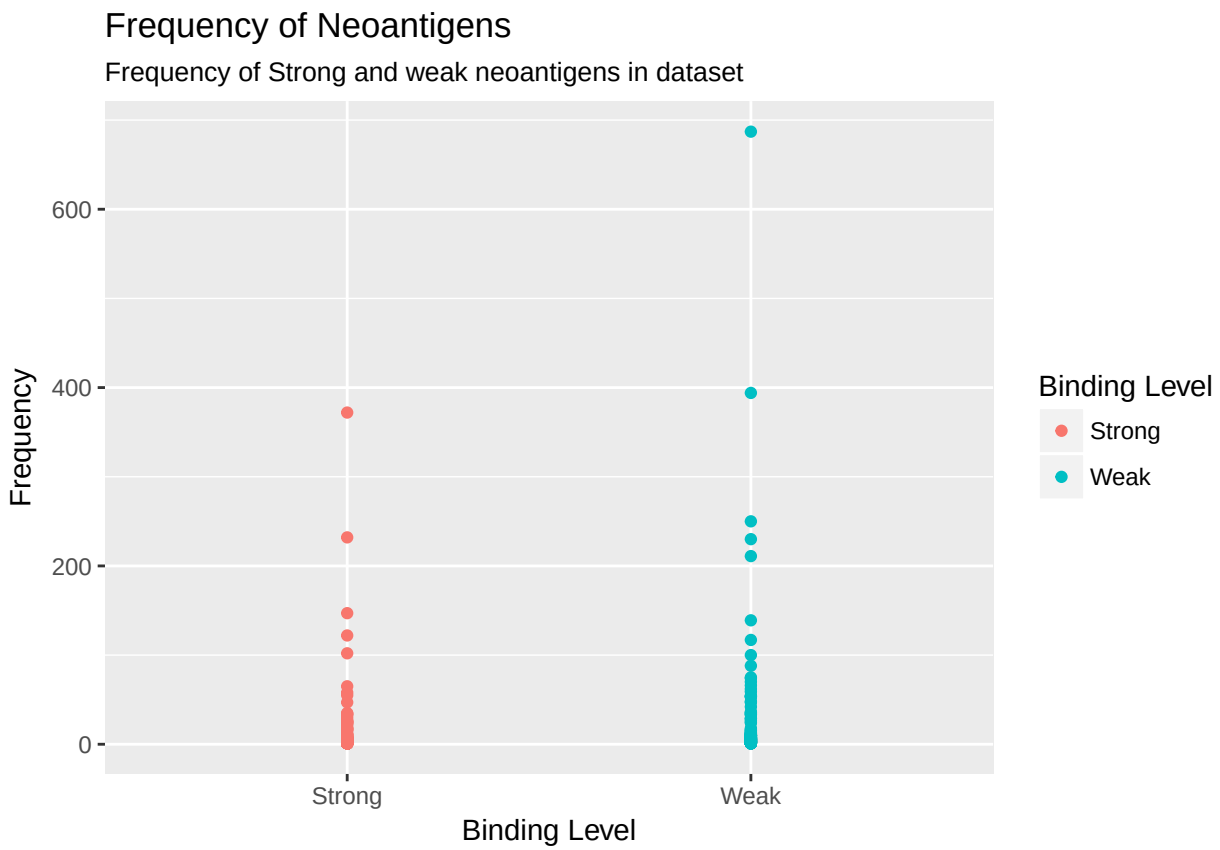
```
## [1] 223  11
```

```r
save(datos_sup, file="datos_supervivencia.Rdata")
```

A continuación se procede a la realización de las gráficas de comparación entre diferentes resultados:

```r
graph_neoant<-subset(datos_sup, select=c("id", "neoant_strong", "neoant_weak"))

graph_neoant_long <- melt(graph_neoant, id="id")


ggplot(data=graph_neoant_long, aes(x=variable, y=value, colour=variable)) +
  geom_point() +
  scale_x_discrete(breaks=c("neoant_strong", "neoant_weak"),
                   labels=c("Strong", "Weak"))+
  scale_colour_discrete(name=("Binding Level"),
                        breaks=c("neoant_strong", "neoant_weak"),
                        labels=c("Strong", "Weak")) +
  labs(title = "Frequency of Neoantigens",
     subtitle = "Frequency of Strong and weak neoantigens in dataset",
     x = "Binding Level", y="Frequency")
```

```
## Warning: Removed 24 rows containing missing values (geom_point).
```



Se procede a la realización de una columna para visualizar la procedencia de las muestra:

```
mut<-datos_sup[c("id", "neoant_strong", "neoant_weak")]
mut$Freq<-(mut$neoant_strong+mut$neoant_weak)
mut<-as.data.frame(mut[c("id", "Freq")])
mut[,"type"] <- "fproject"
```

Se procede a depurar los datos:

```
m_t <- merge(tcia, mut[c("id")], by.x="id2",
             by.y="id", all.x=FALSE, all.y=FALSE) #51


m_m <- merge(mut, tcia[c("id2")], by.x="id",
             by.y="id2", all.x=FALSE, all.y=FALSE) #51



colnames(m_t)<-c("id", "Freq", "type")

pruebb<-rbind.fill(m_m,m_t)
pruebb<-pruebb[with(pruebb, order(id)), ]
```

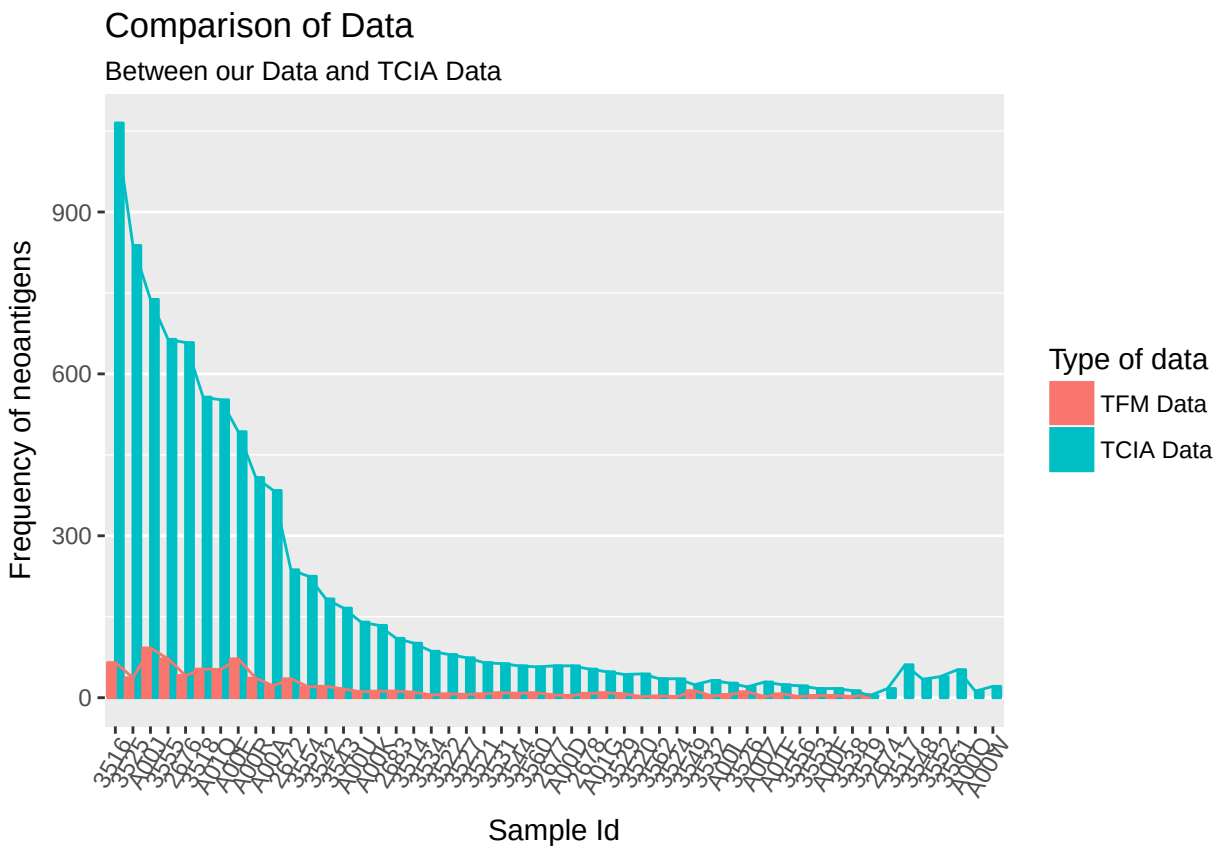
Se realiza la gráfica de péptidos:

```
ggplot(pruebb, aes(x=reorder(id, -Freq), y=Freq, group=type, fill=type, colour=type)) +
  geom_bar(stat="identity", position = "dodge") +
  geom_line(mapping=aes(colour=type)) +
  scale_fill_discrete(name  ="Type of data",
```

```
                      breaks=c("fproject", "tcia"),
                      labels=c("TFM Data", "TCIA Data")) +
  scale_colour_discrete(name  ="Type of data",
                        breaks=c("fproject", "tcia"),
                        labels=c("TFM Data", "TCIA Data")) +
  theme(axis.text.x = element_text(angle=60, hjust=.9),
        panel.grid.major.x = element_blank()) +
  labs(title = "Comparison of Data",
       subtitle = "Between our Data and TCIA Data",
       x = "Sample Id", y="Frequency of neoantigens")
```

```
## Warning: Removed 7 rows containing missing values (geom_bar).
```

```
## Warning: Removed 7 rows containing missing values (geom_path).
```



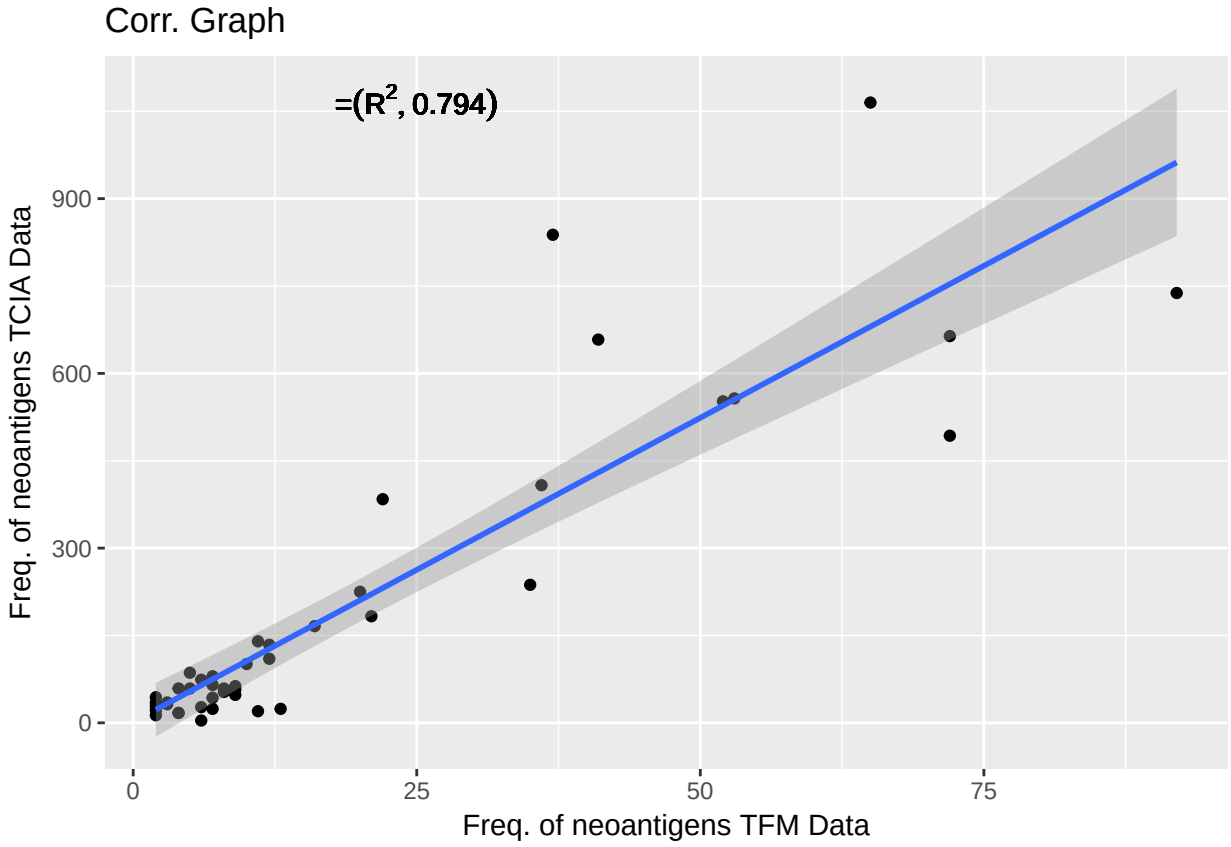
Se realiza la gráfica de correlación:

```
total <- merge(mut, tcia, by.x="id", by.y="id2", all.x=FALSE, all.y=FALSE)

colnames(total) <- c("Barcode", "Freq_fproj", "type.x", "Freq_tcia",
                     "type.y")

model <- lm(Freq_tcia ~ Freq_fproj, data=total)

x <- coef(model)
intercept <- signif(x[1], 3)
terms <- paste(signif(x[-1], 3), names(x[-1]), sep="*", collapse= " + ")
e1 <- paste(intercept, terms, collapse = " + ")
```

```r
e2 <- paste("R^2 = ", round(summary(model)$r.squared, 3))


ggplot(total, aes(x=Freq_fproj, y=Freq_tcia)) +
  geom_point() +
  geom_smooth(method=lm) +
  geom_text(aes(x=25, y=max(total$Freq_tcia),
                label =e2), parse = TRUE, position="identity") +
labs(title = "Corr. Graph",
     x = "Freq. of neoantigens TFM Data", y="Freq. of neoantigens TCIA Data")
```

```
## Warning: Removed 7 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 7 rows containing missing values (geom_point).
```



Análisis de supervivencia:

Se procede a depurar los datos con el fin de poder realizar las gráficas de supervivencia.

```r
datos_sup$st  <- NA
datos_sup$st[datos_sup$OS_STATUS=="LIVING"]<-0
datos_sup$st[datos_sup$OS_STATUS=="DECEASED"]<-1

datos_sup$neo_all <- datos_sup$neoant_strong+datos_sup$neoant_weak
datos_sup <- datos_sup[!datos_sup$OS_MONTHS=="[Not Available]",]
datos_sup$OS_MONTHS<-as.numeric(datos_sup$OS_MONTHS)
```

Generación gráficas de supervivéncia:

Neoantígenos totales:

En primer lugar, se realizará el análisis de supervivencia teniendo en cuenta el total de neoantígenos presentes.

```
summary(datos_sup$neo_all);
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    2.00    6.00    9.00   31.81   14.00 1059.00      21
```

Tal y como se puede observar, la mediana de la muestra es de 8, es el valor que se hará servir para la realización del análisis de supervivencia.

```
datos_sup$all_st <- NA;
datos_sup$all_st[datos_sup$neo_all>=8]<-"high"
datos_sup$all_st[datos_sup$neo_all<8]<-"low"


fit_all<-survfit(Surv(OS_MONTHS, st)~all_st, data=datos_sup)

tiff('superv_all.tiff', units="in", width=7, height=5, res=300);
ggsurvplot( fit_all,  risk.table = TRUE, ncensor.plot=TRUE ,pval = TRUE,
            ggtheme = theme_minimal(),risk.table.y.text.col = T,
            title = "Survival curves Kaplan-Meier estimates",
            subtitle = "All neoantigens strong + weak",
            legend.labs = c("High", "Low"),
            risk.table.y.text = FALSE )
```

Neoantígenos con unión fuerte a MHC-I:

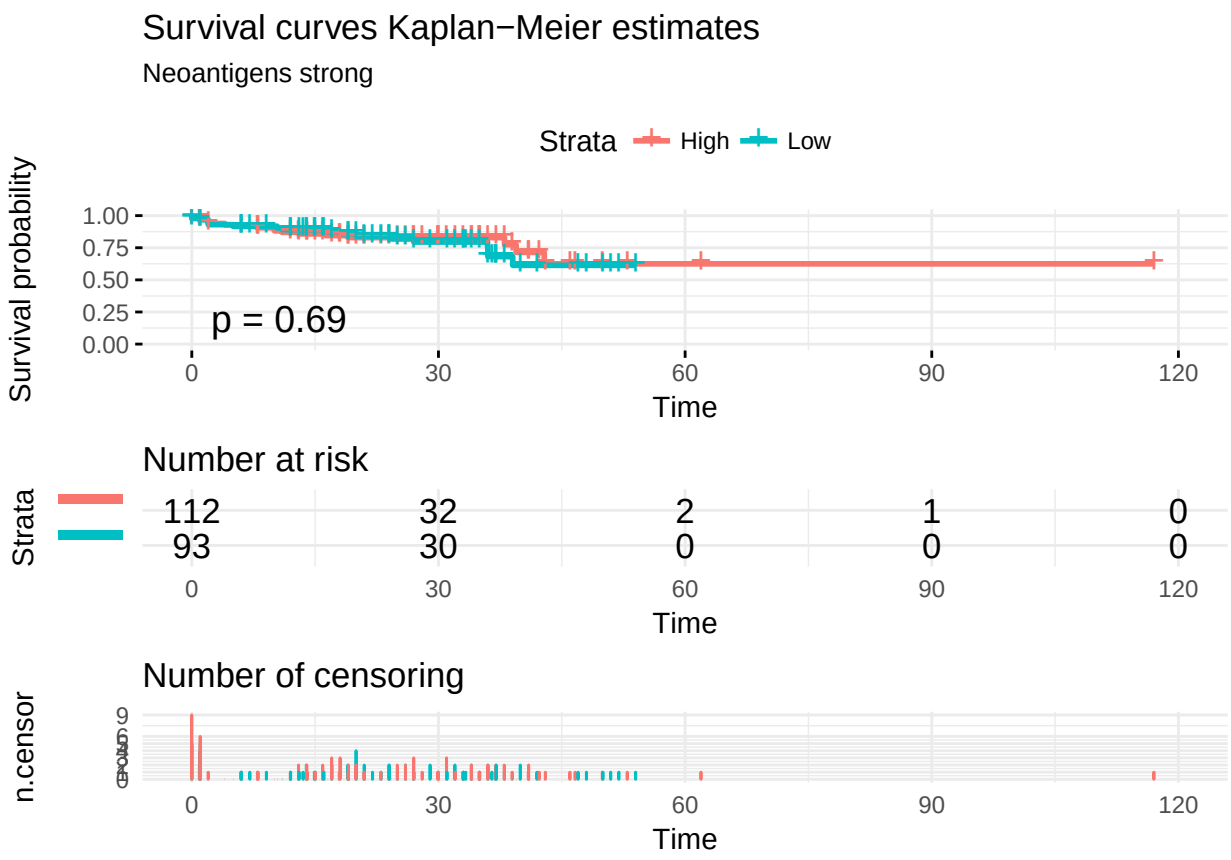
Se realiza el análisis de supervivencia teniendo en cuenta los neoantígenos con unión fuerte a MHC-I.

```
summary(datos_sup$neoant_strong);
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    1.00    2.00    3.00   10.66    5.00  372.00      17
```

Tal y como se puede observar, la mediana de la muestra es de 3, es el valor que se hará servir para la realización del análisis de supervivencia.

```
datos_sup$strong_st <- NA; # mediana
datos_sup$strong_st[datos_sup$neoant_strong>=3]<-"high"
datos_sup$strong_st[datos_sup$neoant_strong<3]<-"low"


fit_strong <-survfit(Surv(OS_MONTHS, st)~strong_st, data=datos_sup)


ggsurvplot( fit_strong,  risk.table = TRUE, ncensor.plot=TRUE ,pval = TRUE,
            ggtheme = theme_minimal(),risk.table.y.text.col = T,
            title = "Survival curves Kaplan-Meier estimates",
            subtitle = "Neoantigens strong",
            legend.labs = c("High", "Low"),
            risk.table.y.text = FALSE )
```



Neoantígenos con unión débil a MHC-I:

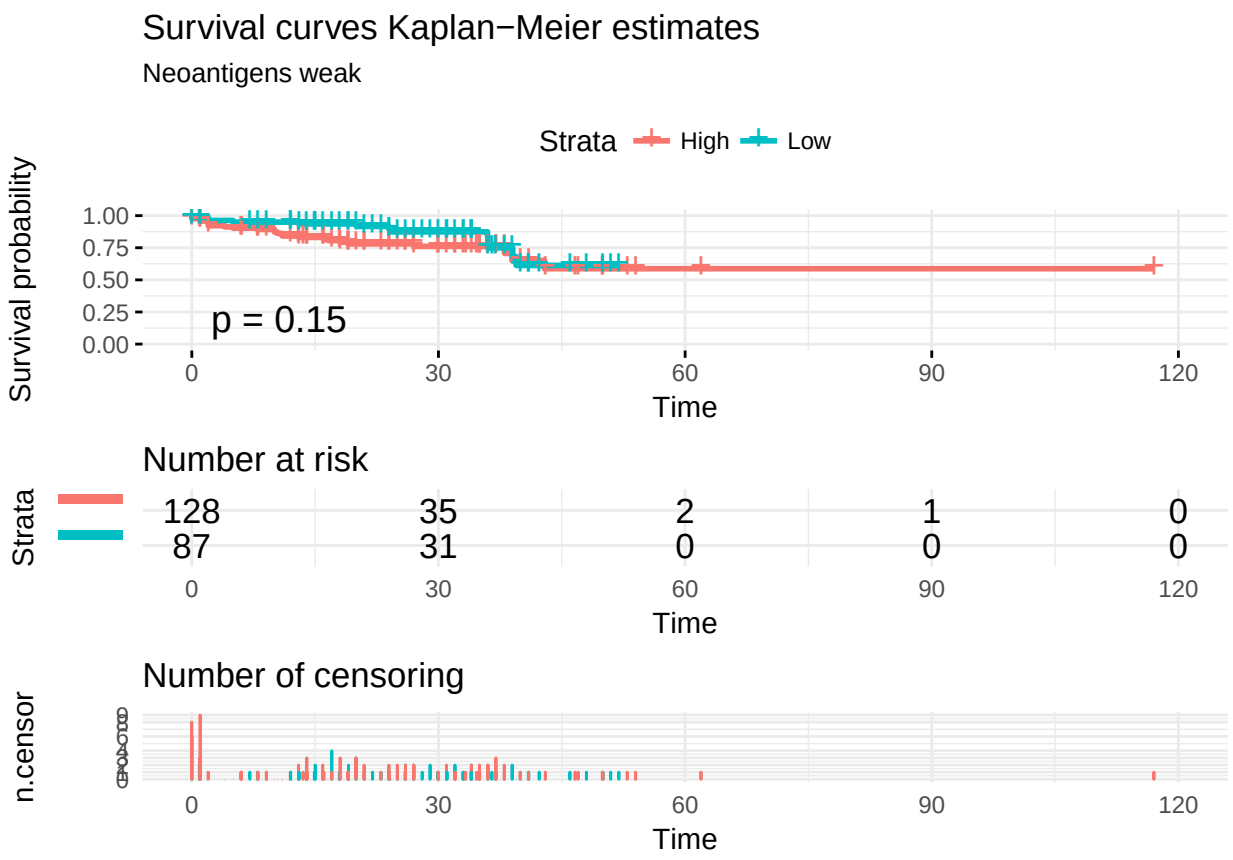
Se realiza el análisis de supervivencia teniendo en cuenta los neoantígenos con unión débil a MHC-I.

```
summary(datos_sup$neoant_weak);
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    1.00    3.00    5.00   19.82   10.00  687.00       7
```

Tal y como se puede observar, la mediana de la muestra es de 5, es el valor que se hará servir para la realización del análisis de supervivencia.

```
datos_sup$weak_st <- NA; # mediana
datos_sup$weak_st[datos_sup$neoant_weak>=5]<-"high"
datos_sup$weak_st[datos_sup$neoant_weak<5]<-"low"


fit_weak <-survfit(Surv(OS_MONTHS, st)~weak_st, data=datos_sup)


ggsurvplot( fit_weak,  risk.table = TRUE, ncensor.plot=TRUE ,pval = TRUE,
            ggtheme = theme_minimal(),risk.table.y.text.col = T,
            title = "Survival curves Kaplan-Meier estimates",
            subtitle = "Neoantigens weak",
            legend.labs = c("High", "Low"),
            risk.table.y.text = FALSE )
```




```
setwd("/Volumes/MARINA/Supervivencia/")
save(netmhc, file="netmhc.Rdata")
```