Informatika Ingeniaritzako Gradua

Grado en Ingeniería Informática

Konputazio Zientziak

Departamento de Ciencias de la Computación

Gradu Amaierako Lana

Trabajo de Fin de Grado

Búsqueda de patrones en árboles tumorales en cáncer de mama

Laura Prieto Garmendia

Zuzendaritza
Dirección

Borja Calvo Molinos

Konputazio Zientziak eta Adimen Artifiziala Saila

Departamento de Ciencia de la Computación e Inteligencia Artificial

Maitena Tellaetxe Abete

Intelligent System Group (ISG) - Biodonostia

Universidad del País Vasco Euskal Herriko Unibertsitatea

INFORMATIKA FAKULTATEA
FACULTAD DE INFORMÁTICA

eman ta zabal zazu

Universidad del País Vasco
Euskal Herriko Unibertsitatea

INFORMATIKA FAKULTATEA
FACULTAD DE INFORMÁTICA

Trabajo de Fin de Grado

Grado en Ingeniería Informática

Computación

Búsqueda de patrones en árboles tumorales en cáncer de mama

Laura Prieto Garmendia

Dirección
Borja Calvo Molinos
Maitena Tellaetxe Abete

25 de junio de 2022

Agradecimientos

Después de seis intensos meses, hoy es el día. Escribo este apartado de agradecimientos por la realización de mi trabajo final de carrera. Ha sido un periodo de aprendizaje intenso, no solo en el campo de las ciencias, sino también a nivel personal.

En especial quiero agradecer a mis tutores, Borja y Maitena, por su inestimable ayuda. Me habéis ayudado cuando lo necesitaba y gracias a vosotros realizar el trabajo ha sido mucho más factible. Agradezco esta oportunidad por haber trabajado con personas tan competentes y profesionales. Muchas gracias.

También me gustaría agradecer a mi padre y a mi madre por sus sabios consejos, comprensión y paciencia. Siempre estáis a mi lado.

Resumen

Las mutaciones somáticas son las responsables de originar tumores. Un tumor es el resultado de un proceso evolutivo en el que las mutaciones somáticas se acumulan y conducen al crecimiento descontrolado de las células. Para estudiar los tumores es imprescindible analizar las mutaciones que tienen en su ADN.

En este proyecto se han analizado múltiples muestras de diferentes individuos para identificar patrones de evolución comunes. Para ello, se han preprocesado las muestras de tres pacientes de cáncer de mama de secuenciación de ADN de tumor y se ha obtenido la lista de mutaciones presentes en el tumor, junto con su frecuencia. A continuación se han obtenido varios árboles tumorales evolutivos para cada paciente utilizando un algoritmo de búsqueda local existente. Por último, concluimos el proyecto con el análisis de dichos árboles desarrollando un algoritmo para hallar el Maximum Spanning Tree que aplicamos para hallar el árbol consenso a partir de los árboles obtenidos para cada paciente.

Los resultados obtenidos no han mostrado patrones claros. Esto es probablemente debido a los datos en sí mismos que, al proceder de muestras conservadas en parafina pueden contener muchos artefactos, además de combinar muestras tanto de tumores primarios como de sus metástasis.

Índice de contenidos

Índice de figuras

Índice de tablas

CAPÍTULO 1

Introducción

Este proyecto se enmarca dentro del área de las ciencias de la computación, en concreto, en la Bioinformática.

Los tumores malignos tienen características fenotípicas y moleculares muy diversas tanto a nivel intertumoral como intratumoral. La heterogeneidad intertumoral se refiere a las diferencias encontradas entre los tumores de distintas pacientes. La heterogeneidad intratumoral, en cambio, se refiere a la existencia de poblaciones celulares tumorales distintas dentro de un mismo tumor [1]. En este proyecto se analizará la heterogeneidad intratumoral e intertumoral en tres pacientes de cáncer de mama.

La investigación para comprender y caracterizar la heterogeneidad puede permitir una mejor comprensión de las causas y la progresión de la enfermedad. A su vez, tiene el potencial de guiar la creación de estrategias de tratamiento más refinadas que incorporen el conocimiento de la heterogeneidad para obtener un mejor resultado.

1.1. Objetivos principales

El objetivo principal del trabajo es analizar la heterogeneidad intratumoral e intertumoral en cáncer de mama a partir de múltiples muestras de diferentes individuos para identificar patrones de evolución comunes.

Para ello, es necesario procesar y preparar datos en crudo obtenidos de biopsias de cáncer de mama. Una vez procesados y preparados los datos, se aplicará un algoritmo de búsqueda local existente para resolver el problema de deconvolución y evolución clonal a los datos procesados. Finalmente, se realizará un análisis de dichos árboles que incluye la comparación entre árboles de una misma paciente pero también entre pacientes.

1.2. Motivación

El cáncer de mama ya es el cáncer más diagnosticado en el mundo, superando por primera vez al cáncer de pulmón, según datos publicados por el Centro de Investigación del Cáncer (IARC) en 2021 [2]. Según los últimos datos recogidos por el Sistema Europeo de Información sobre el Cáncer (ECIS), en 2020 se diagnosticaron en España un total de

34.088 nuevos casos de cáncer de mama. Este tipo de tumor es el más frecuente entre las mujeres en España, por delante del colorrectal, útero, pulmón y ovario. Cerca del 30 % de las mujeres diagnosticadas de cáncer tienen su origen en la mama.

Es importante señalar que el 30 % de las pacientes que reciben un diagnóstico precoz experimentarán una recaída metastásica, y que según el informe 'Cancer Data 2021' de la Sociedad Española de Oncología Médica (SEOM), los tumores de mama siguen siendo la principal causa de muerte por cáncer en mujeres españolas, con el 5.5 % de todas las muertes por cáncer. Pese a ello, gracias a la investigación y a los tratamientos aplicados, la supervivencia ha aumentado de manera significativa en los últimos años [2]. Aún así, es necesario seguir investigando en este área para que la mejora y el bienestar social incremente.

1.3. Biología de un tumor

Para estudiar los tumores es imprescindible analizar las mutaciones que tienen en su ADN. El ADN humano es una cadena de aproximadamente 3300 millones de longitud cuya unidad son los nucleótidos. Hay cuatro tipos de nucleótidos: adenina (A), guanina (G), citosina (C) y timina (T). Cuando ocurre un cambio de un nucleótido a otro en una posición, se dice que ha ocurrido una mutación. En todos los seres vivos ocurren mutaciones [1] constatemente. La mayoría de las mutaciones son inocuas, pero algunas pueden conducir a la aparición de tumores. En este caso, la célula en la que ha ocurrido la mutación crece más rápido, no muere y es capaz de vivir con menos recursos. Cabe destacar sin embargo que, normalmente, un tumor no se desarrolla por una sola mutación, sino que, lo normal es que una célula tumoral tenga muchas mutaciones somáticas.

Por lo tanto, a la hora de analizar las mutaciones que hay en un tumor, hay que tener en cuenta que algunas de ellas serán germinales, mientras que otras serán somáticas.

Tal y como hemos dicho, las células tumorales tienen un crecimiento anormal, y además de eso, tienen la capacidad de metastizar. Esto quiere decir que las celulas pueden migrar desde el lugar originario (lo que se conoce como tumor primario) a cualquier otro órgano o tejido y crear allí un nuevo tumor. Esto se denomina como metástasis.

1.3.1. Evolución clonal

1.3.1.1. Evolución clonal en el cáncer

Como ya hemos adelantado, una sola mutación no es suficiente para que se desarrolle un tumor, por lo que las células tumorales suelen contener más de una mutación en su ADN. La teoría clonal del cáncer de Nowell se basa en la teoría de la evolución de Darwin para describir un modelo que intenta explicar este proceso [3]. Según esta teoría, un tumor se originaría a partir de una única célula que en algún momento adquirió una mutación que le proporcionaría una ventaja de crecimiento. Los descendientes de esta célula fundadora adquirirían mutaciones adicionales de manera aleatoria, al igual que cualquier otra célula.

[1] Hay dos tipos de mutaciones, las mutaciones germinales y las somáticas. Las mutaciones germinales son aquellas que se heredan y, por tanto, están presentes en todas y cada una de las células. En cambio, las somáticas son las que aparecen con el paso del tiempo y, por tanto, solo están en algunas células (descendientes de la primera que sufrió esa mutación). Las mutaciones que dan lugar a tumores son principalmente las mutaciones somáticas.

Las fuerzas de selección y las condiciones como la existencia o no de oxígeno, las condiciones físicas del entorno o incluso el propio tratamiento actuarían favoreciendo la supervivencia o la expansión de aquellas poblaciones celulares o clones que contendrían mutaciones beneficiosas para el crecimiento del tumor.

Llamamos mutaciones clonales a aquellas mutaciones que surgen en la célula fundadora o al principio de la expansión y mutaciones subclonales a las que se producen en una etapa posterior. Como resultado de este proceso evolutivo, en lugar de masas homogéneas, los tumores se presentan como masas heterogéneas compuestas por clones que contienen diferentes conjuntos de mutaciones y que configuran sus características como su tasa de crecimiento, respuesta al tratamiento o la capacidad de metástasis. Esta característica de los tumores se conoce como heterogeneidad intratumoral (ITH) y su estudio es esencial, no sólo para comprender mejor el desarrollo del cáncer, sino también de cara a la práctica clínica para ayudar a diseñar terapias adaptadas a las particularidades de los diferentes clones presentes en el tumor [4].

1.3.1.2. Filogenias tumorales

Dado que, como hemos visto, los tumores se desarrollan siguiendo un proceso evolutivo, podemos representarlos mediante árboles filogenéticos. La historia evolutiva de un tumor describe el orden de aparición de estas mutaciones. Un árbol filogenético tumoral es un árbol compuesto por vértices o nodos que representan clones del tumor y aristas que representan las relaciones ancestrales entre los clones. Su raíz suele ser el clon que originó el tumor. Cada uno de los nodos representa un clon que ha surgido en algún momento de la evolución del tumor. Algunos de estos clones pueden ya no estar presentes en el tumor, mientras que los otros son potencialmente identificables [5].

1.4. Secuenciación de ADN

Para poder analizar las mutaciones de un tumor, es necesario secuenciar su ADN. Para ello, en primer lugar se extraen una o varias muestras del mismo, en el laboratorio se extrae el ADN de las (millones de) células que hay en cada biopsia y se procesa para la secuenciación. Se realizan varios pasos, pero los interesantes para este proyecto son los siguentes:

- **Fragmentación del ADN.** En primer lugar se fragmenta la cadena de ADN en trozos más pequeños y manejables.
- **Selección las zonas a secuenciar y copia masiva de esos fragmentos.** En los datos que se han empleado en este proyecto no se ha secuenciado todo el genoma sino solo 2.352 regiones del mismo. La selección de esas regiones del genoma se hace mediante unas secuencias llamadas adaptadores. Los adaptadores son pequeños fragmentos de ADN que son complementarios a las regiones que delimitan la región de interés. Físicamente lo que se hace es incorporar millones de dichos adaptadores al recipiente en el que está el ADN fragmentado. Una vez que estos adaptadores están unidos al ADN, una molécula identifica que ha ocurrido esta unión y hace una copia de cada zona flanqueada por los adaptadores. Este proceso de incorporación de adaptadores y copia de la región se repite durante varios ciclos. De esta manera, se

obtienen múltiples copias de las regiones que queremos secuenciar. Estas secuencias se llaman amplicones.

- **Filtración de los amplicones.** En el recipiente se encuentran los amplicones y el ADN original. Se hace un filtrado por tamaño y se seleccionan solo los amplicones (los amplicones suelen ser más pequeños que el ADN original).
- **Secuenciación.** Finalmente se secuencian los amplicones mediante instrumentos de secuenciación de ADN.

Al secuenciar el ADN se obtienen millones de lecturas que contienen la secuencia de distintos fragmentos de ADN de todas las células que había en la biopsia tomada. En este punto comienza la parte del análisis de datos de secuenciación.

En la secuenciación se pueden producir problemas como que algunos adaptadores se unan a zonas equivocadas o que en algunos fragmentos que queramos secuenciar no se unan los adaptadores y, por lo tanto, no se secuencie ese fragmento. Es por ello por lo que los datos de lecturas no se corresponden a los valores reales sino que son estimadores. Este es un punto a tener en cuenta a la hora de tratar con estos datos.

1.5. Herramientas computacionales para el análisis de datos de secuenciación

Existen varias formas para el análisis de datos de secuenciación de ADN pero en este proyecto solo se analizan de una sola manera. Para ello, necesitamos varias herramientas:

- **FastQC [6]:** FastQC es una herramienta de análisis de control de calidad diseñada para detectar posibles problemas en conjuntos de datos de secuenciación. Su objetivo es proporcionar una manera sencilla de verificar la calidad de los datos de secuencias crudas procedentes de técnicas de secuenciación. El programa realiza un conjunto modular de análisis en uno o más archivos de secuencia cruda en formato FASTQ o BAM. Características:
  - Importa datos de archivos BAM, SAM o FASTQ.
  - Ofrece una visión rápida que destaca posibles áreas problemáticas.
  - Gráficos y tablas de resumen para una rápida evaluación de los datos.
  - Resultados de exportación a HTML.
  - Se puede utilizar sin conexión.
- **Trimmomatic [7]:** Trimmomatic realiza una variedad de tareas de recorte. Es decir, después de hacer el análisis de calidad, hay que eliminar las posibles secuencias y adaptadores de las lecturas de baja calidad. La selección de pasos de recorte y sus parámetros asociados se suministran por línea de comandos:
  - *ILLUMINACLIP:* Cortar el adaptador y otras secuencias específicas de la lectura.
  - *SLIDINGWINDOW:* Realizar un recorte, cortando las secuencias que no cumplen los criterios de calidad.

- *LEADING:* Cortar las bases del inicio de una lectura que están por debajo de un umbral de calidad.
- *TRAILING:* Cortar las bases del final de una lectura que están por debajo de un umbral de calidad.
- *CROP:* Cortar la lectura a una longitud especificada.
- *HEADCROP:* Recortar el número especificado de bases desde el inicio de la lectura.
- *MINLEN:* Descartar la lectura si está por debajo de una longitud especificada.
- *TOPHRED33:* Convertir puntuaciones de calidad en Phred-33 (un tipo de codificación para la calidad de las lecturas).
- *TOPHRED64:* Convertir puntuaciones de calidad en Phred-64 (otra codificación de calidad).

Funciona con ficheros FASTQ. Para los datos terminados, se especifican una entrada y un archivo de salida, además de los pasos de procesamiento. El primer fichero de la pareja devuelve las lecturas 'limpias', es decir, lecturas a las que ya se han hecho el trimming. El segundo fichero devuelve las lecturas que han quedado 'sueltas' tras el trimming.

- **BWA [8]:** BWA es un paquete de software para mapear lecturas cortas contra un genoma de referencia, como el genoma humano. En este proyecto se utiliza la versión del genoma humano hg19. Este paquete consta de tres algoritmos: BWA-backtrack, BWA-SW y BWA-MEM. En este caso hemos utilizado BWA-MEM; se recomienda generalmente ya que es rápido y preciso.
- **Picard [9]:** Picard es un conjunto de herramientas de línea de comando para manipular datos y formatos de secuenciación como SAM, BAM, CRAM y VCF . En este proyecto, picard se ha empleado para preparar los datos alineados (ficheros BAM) para ser tratados por Mutect2.
- **Mutect2 [10]:** Mutect2 es una herramienta para identificar mutaciones en muestras. Contiene modulos especializados para que la estimación de las frecuencias de las mutaciones sea más exacta.
- **SnpEff [11]:** Es un conjunto de herramientas de anotación de mutaciones genéticas y predicción de efectos funcionales. Anota y predice los efectos de las variantes genéticas en los genes y las proteínas. En este proyecto se ha empleado para anotar mutaciones germinales en la lista de mutaciones proporcionada por Mutect2.
- **BCFtools [12]:** BCFtools es un conjunto de utilidades que manipulan los ficheros VCF. La mayoría de los comandos aceptan archivos VCF, VCF bgzipped y BCF.
- **IGV [13]:** Integrative Genomics Viewer es un visualizador y una plataforma de integración de datos donde los datos pueden visualizarse en el contexto de otra información. Principalmente se utiliza para cargar ficheros BAM y ver dónde (en el genoma) han mapeado las lecturas genómicas.
- **RStudio:** Es un entorno de desarrollo que permite desarrollar código en lenguajes de programación como R, y que está orientado al procesamiento de datos masivos, estadísticas, etc.

1.6. Herramientas computacionales para el estudio de la ITH

En este apartado se hace un repaso de los diferentes aspectos relacionados con el análisis de la heterogeneidad intratumoral a partir de datos de secuenciación de biopsias tumorales. En primer lugar se ha definido el problema de la deconvolución y evolución clonal y a continuación se ha hablado sobre las herramientas para el análisis de los árboles.

1.6.1. Problema de la deconvolución clonal y reconstrucción de la evolución

Como se ha explicado anteriormente, para hallar el árbol filogenético de un tumor, se extraen una o varias muestras, se secuencian mediante un procedimiento de secuenciación de ADN, y, finalmente, se identifican sus mutaciones o variantes y la proporción en la que se encuentran en cada muestra. Cada muestra tiene una combinación de diferentes clones, por lo que el valor de la frecuencia de cada mutación en cada muestra es el resultado de la suma de las proporciones de los clones en los que se encuentra dicha mutación en la muestra. Por ello, para hallar el árbol evolutivo a partir de este tipo de datos, es necesario resolver el problema de la deconvolución clonal, es decir, la identificación de los clones presentes en las muestras junto con sus proporciones [5].

Para resolver el problema de la deconvolución clonal, se ha utilizado un algoritmo de búsqueda local existente, es decir, un algoritmo heurístico. Los métodos heurísticos exploran el espacio de soluciones del problema parcialmente en lugar de hacerlo de forma exhaustiva, con el objetivo de encontrar soluciones suficientemente buenas en un tiempo razonable, a pesar de no garantizar que las soluciones sean óptimas. Dichos algoritmos son especialmente útiles para los problemas del mundo real en los que el espacio de soluciones es demasiado amplio para ser analizado por completo [5].

1.6.2. Árboles consenso

Lo habitual es que, dado un conjunto de mutaciones de un tumor y sus frecuencias, existan múltiples árboles filogenéticos que expliquen igualmente bien cuál ha sido la evolución de dicho tumor. Así, dado un conjunto de historias evolutivas tumorales, se debe de analizar si la información de dichas historias puede combinarse para inferir una historia evolutiva mejor. Este tipo de consenso ha sido útil al aplicarse a los árboles filogenéticos tradicionales que se utilizan para mostrar las relaciones evolutivas entre diferentes especies.

Existen varios métodos para encontrar un árbol consenso.

- Uno de los métodos más sencillos es el **consenso estricto** (*strict consensus*). Crea un árbol que contiene todas las mismas agrupaciones de especies (o clones) que aparecen en todos los árboles de entrada.
- El **árbol consenso con regla de la mayoría** (*majority-rule consensus tree*) crea un árbol que contiene agrupaciones que existen en la mayoría de los árboles de entrada.
- El **consenso de Adams** observa qué especies (o clones) se agrupan a menudo en el conjunto de árboles de entrada [14].

En este proyecto el método utilizado se situaría en la categoría del árbol consenso con regla de la mayoría, ya que las aristas de los árboles obtenidos representan las aristas con mayor peso, es decir, las que existen en la mayoría de los árboles de entrada.

Para la búsqueda de árboles consenso, se ha implementado el algoritmo Maximum Spanning Tree (MST). Para ello, se ha utilizado el algoritmo Prim [15], es decir, un algoritmo voraz.

Dado un grafo conexo $G = (V, E)$, un spanning tree es un árbol que abarca todos los vértices del grafo G y es un subgrafo del grafo G, es decir, todas las aristas que incluye el spanning tree, aparecen en el grafo inicial.

Por lo tanto, el objetivo del algoritmo MST, es hallar el subconjunto de aristas que incluye todos los vértices del grafo de forma que la suma de los pesos de las aristas pueda maximizarse. Comienza con un único nodo y explora todos los nodos adyacentes con todas las aristas de conexión en cada paso. Se seleccionan las aristas con los pesos máximos que no causan ciclos en el grafo. El coste del Maximum Spanning Tree es la suma de los pesos de todas las aristas del árbol.

CAPÍTULO 2

Gestión del trabajo

En este capítulo se analiza la gestión del trabajo a nivel de proyecto. El objetivo principal del trabajo es analizar la heterogeneidad intratumoral e intertumoral en cáncer de mama a partir de múltiples muestras de diferentes individuos para identificar patrones de evolución comunes.

2.1. Objetivos específicos

Este proyecto está basado en la investigación del área de las ciencias de la computación, especialmente en la bioinformática. Cuyos objetivos son los siguientes:

- Procesar y preparar datos en crudo obtenidas de biopsias de cáncer de mama, identificando las mutaciones presentes y su presencia.
- Obtener y analizar los árboles evolutivos para cada paciente.
- Desarrollar estrategias para la búsqueda de árboles consenso a partir de árboles obtenidos de diferentes individuos.
- Obtener y analizar el/los árboles consenso a partir de los resultados obtenidos.

El hito principal del proyecto es la tercera convocatoria del curso 2021/22 para poder presentar el TFG. Es decir, el 26 de junio de 2022.

2.2. Planificación

Con el diagrama de la *Estructura de la Descomposición del Trabajo* (EDT) se ha descompuesto el trabajo realizado durante el proyecto (Figura 2.1).

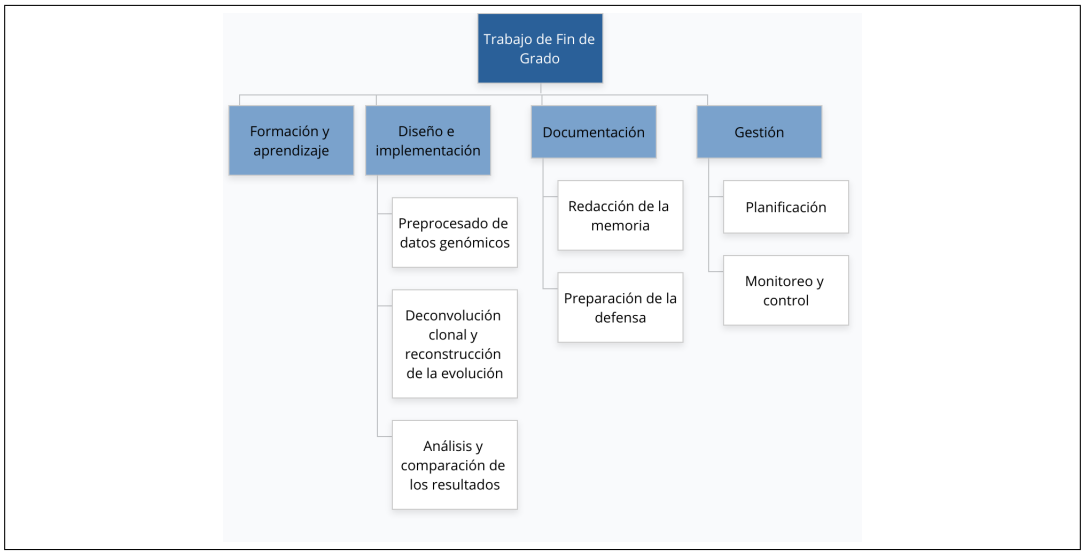


Figura 2.1: Diagrama EDT/WBS

2.2.1. Paquetes de trabajo

Los paquetes de trabajo constituyen la parte más baja del diagrama EDT (Figura 2.1) y definen las actividades y tareas principales en las cuales se divide el proyecto. En este apartado se da una explicación de cada uno de ellos.

Formación y aprendizaje

Esta sección será imprescindible para entender la base del trabajo. Se recogerán todos los conceptos básicos necesarios sobre los conceptos biológicos, preprocesado de datos genómicos y de las herramientas de análisis de ITH.

Preprocesado de datos genómicos

Este paquete de trabajo será el primer paso para poder analizar los datos. Para ello, con los datos en bruto de muestras de cáncer de mama se hará limpieza, se alinearán y se identificarán las mutaciones presentes y su frecuencia (*variant calling*).

Deconvolución clonal y reconstrucción de la evolución

En este apartado se utilizará un método heurístico para obtener de cada paciente uno o varios árboles que expliquen la composición clonal y su evolución.

Análisis y comparación de los resultados

En esta sección se recolectarán todos los resultados obtenidos a través del paquete *Deconvolución clonal y reconstrucción de la evolución*. Para ello se diseñarán algoritmos para la comparación de múltiples árboles. Se efectuará un análisis e interpretación de los resultados obtenidos.

Documentación

En este apartado se desarrollará la memoria y la defensa. A través de estos documentos se explicará todo el desarrollo del proyecto y se explicará cada paso detalladamente. Para exponer la defensa habrá un documento cuya función será mostrar lo más importante del proyecto de una manera clara y ordenada.

Planificación

En este paquete de trabajo se planifica el proyecto especificando y definiendo los objetivos del proyecto, determinando las tareas que realizar y el método de trabajo. Se especifican los tiempos, peligros y riesgos del proyecto.

Monitoreo y control

El monitoreo y control será un paquete que se prolongará durante todo el trabajo y que garantizará que el Trabajo de Fin de Grado quede terminado para las fechas indicadas. Para ello, será necesario llevar un control y efectuar reuniones de supervisión de una manera continua.

2.2.2. Planificación del tiempo

Cada paquete de trabajo se divide en varias tareas. A continuación se mostrará cada tarea y se definirá en la Tabla 2.1 cuánto tiempo estaba previsto para efectuar cada una, y en realidad cuánto tiempo se ha necesitado. Las tareas son las siguientes:

- PT1 - Formación y aprendizaje
  - T1.1 - Conceptos biológicos
  - T1.2 - Preprocesado de datos genómicos
  - T1.3 - Herramientas de análisis de ITH
- PT2 - Preprocesado de datos genómicos
  - T2.1 - Limpieza de los datos
  - T2.2 - Alineación de los datos en bruto
  - T2.3 - Variant calling
- PT3 - Deconvolución clonal y reconstrucción de la evolución
  - T3.1 - Preparación de los datos para aplicar los algoritmos
  - T3.2 - Aplicación de los algoritmos y recolección de los resultados
- PT4 - Análisis y comparación de los resultados
  - T4.1 - Diseño de medidas y algoritmos para la comparación de múltiples árboles
  - T4.2 - Diseño de algoritmos para la agregación de árboles
  - T4.3 - Implementación del algoritmo Maximum Spanning Tree
  - T4.4 - Análisis e interpretación de los resultados obtenidos en PT3
- PT5 - Documentación
  - T5.1 - Escritura de la memoria
  - T5.2 - Preparación de la defensa
- PT6 - Planificación
- PT7 - Monitoreo y control

Paquetes de trabajo / Tareas	Tiempo esperado (horas)	Tiempo real (horas)
PT1	**35**	**35**
T1.1	15	15
T1.2	10	10
T1.3	10	10
PT2	**40**	**70**
T2.1	15	25
T2.2	10	20
T2.3	15	25
PT3	**40**	**40**
T3.1	10	10
T3.2	30	30
PT4	**120**	**70**
T4.1	20	0
T4.2	20	0
T4.3	30	30
T4.4	50	40
PT5	**50**	**60**
T5.1	35	40
T5.2	15	20
PT6	**10**	**10**
PT7	**5**	**5**
TOTAL	**300**	**290**

Tabla 2.1: Tiempo invertido en cada tarea

2.2.3. Diagrama de Gantt

En el diagrama de Gantt (Figura 2.2) se definen las fechas de inicio y finalización esperadas de cada paquete de trabajo.

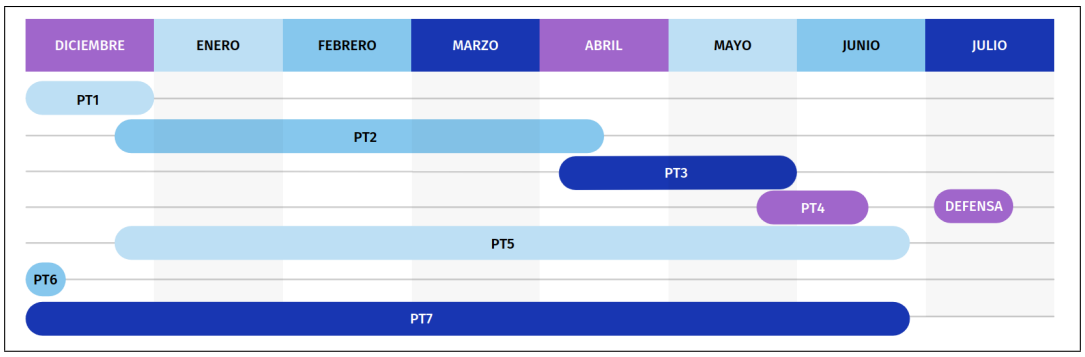


Figura 2.2: Diagrama de Gantt

2.2.4. Entregables

Los entregables son productos que dan por finalizado una o varias fases del proyecto, por lo tanto en este proyecto serán los siguientes:

- La memoria: recogerá todos los datos necesarios y dará todas las explicaciones explícitamente.

- La presentación: documento cuya función será mostrar lo más importante del proyecto de una manera clara y ordenada.

En la Tabla 2.2 se han fijado las fechas de finalización de cada entregable.

Entregable	**Fecha límite**
Memoria	26 de junio de 2022
Presentación	Del 4 al 13 de julio de 2022

Tabla 2.2: Fechas límite

2.2.5. Planificación de los riesgos

En el desarrollo del proyecto surgen riesgos o inconvenientes que pueden llegar a cambiar el rumbo del proyecto. Por eso, es imprescindible localizar los posibles riesgos y buscar soluciones para poder solucionarlos antes de que el riesgo aumente. En la Tabla 2.3 se identifica la probabilidad e impacto de cada posible riesgo.

R1 - Durante el desarrollo del proyecto pueden perderse los datos imprescindibles. Para evitar perdidas importantes se irán haciendo copias de seguridad en la nube y en un disco duro. De esta manera, en caso de que sucediese, tendría una copia y en caso de perder datos no cambiaría el rumbo del proyecto.

R2 - En proyectos de este calibre es muy común que haya grandes desviaciones respecto a la planificación principal. Este riesgo puede causar grandes retrasos en el tiempo, por lo tanto para evitarlo he previsto la finalización del proyecto con antelación. De esta manera tendría tiempo para poder terminar el trabajo.

R3 - Necesidad demasiado alta de computación. Es muy probable que la máquina local no sea capaz de gestionar los datos en brutos, ya que para el preprocesado de datos génomicos se necesitará un ordenador con gran capacidad. En caso de que ocurriese, se utilizaría un servidor con suficiente capacidad.

R4 - Resultados negativos o no concluyentes al preprocesar los datos. Es probable que al preprocesar los datos los resultados no sean triviales ya que en el cáncer de mama no se encuentran muchas mutaciones. En caso de que este riesgo ocurriese, se cambiarían los datos obtenidos a otro cáncer. No se empezaría a analizar otra vez todo el proceso, sino que se escogerían datos preprocesados, de esta manera no habría ningún inconveniente en el rumbo del proyecto.

R5 - Dificultad teórica o práctica. Comprender ciertos conceptos teóricos puede causar un riesgo ya que en este Trabajo de Fin de Grado no solo se tabaja en el ámbito de la informática, sino que también en el de la biología. No obstante, implementar los algoritmos también puede llegar a ser un riesgo ya que a la hora de programar siempre pueden haber imprevistos. Para solucionarlo, se exigiría más esfuerzo de lo previsto y se efectuarían más reuniones.

Probabilidad / **Impacto**	<= 15 %	30 %	45 %	60 %	75 %	90 % =<
Muy bajo						
Bajo			*R2*		*R3*	
Mediano	***R1***			*R5*		
Alto						*R4*
Muy alto						

Tabla 2.3: Probabilidad e impacto de los riesgos estimados

2.2.6. Herramientas de gestión

Es necesario establecer una planificación de comunicaciones. Por lo tanto, principalmente utilizaremos el correo electrónico y las reuniones para comunicarnos:

- Correo electrónico: será el canal más utilizado. Las consultas puntuales se realizarán a través de este medio, así como para las convocatorias a las reuniones.
- Reuniones: este canal será imprescindible para debatir lo realizado en el proyecto y planificar los siguientes pasos. En caso de no poder acudir presencialmente, se utilizará la plataforma WebEx.

Por otro lado, es indispensable establecer un sistema de información con el que trabajar:

- Localmente en el ordenador: se almacenarán los documentos de poco interés para que en caso de pérdida el rumbo del proyecto no cambie.
- Overleaf: la memoria se efectuará en esta plataforma, es decir, en un repositorio online cuyo riesgo de pérdida de datos es nulo.
- Memoria externa: todos los datos de gran tamaño se situarán en una memoria externa.

CAPÍTULO 3

Preprocesado de datos genómicos

En este capítulo se describen los datos usados así como el preprocesado de los mismos. En el Anexo A se incluye, a modo de ejemplo, el detalle del preprocesado de una de las muestras.

3.1. Datos

Disponemos de datos de tres pacientes de cáncer de mama. Para cada paciente, tenemos múltiples muestras obtenidas tanto del tumor primario como de las metástasis:

- Paciente 1: 9 muestras. En concreto, obtenemos una muestra del tumor primario, dos muestras de mama contralateral, cuatro muestras del ganglio positivo, una muestra del ganglio no tumoral y una de la metástasis a distancia (branquial).
- Paciente 2: 8 muestras. Concretamente, una muestra del tumor primario, cinco del ganglio positivo, una del ganglio no tumoral y una de la metástasis a distancia (ganglio cervical).
- Paciente 3: 6 muestras. Disponemos de una muestra del tumor primario, tres del ganglio positivo, una del ganglio tumoral, una del ganglio no tumoral y finalmente, una de la metástasis a distancia (ganglio supraclavicular).

Gracias a que tenemos una muestra normal, es decir, no tumoral, de cada paciente, podremos separar las mutaciones somáticas de las germinales.

En las 23 muestras se ha extraído el ADN y se ha secuenciado, es decir, se ha sacado su secuencia para identificar qué mutaciones hay.

Todas las muestras son FFPE (formalin-fixed paraffin-embedded). FFPE es una forma de conservación y preparación de muestras de biopsia en la que una muestra de tejido se conserva primero fijándola en formaldehído, también conocido como formol, para conservar las proteínas y las estructuras vitales del tejido. A continuación, se incrusta en un bloque de cera de parafina, lo que facilita el corte para su examen. El uso de esta técnica permite la conservación de muestras de tejidos a temperatura ambiente, sin la

necesidad de refrigeración, pero tiene el inconveniente de que crea artefactos a la hora de la secuenciación [16].

Ya que secuenciar todo el ADN de estas células es muy caro y se necesita ADN de muy buena calidad, sólo se han secuenciado algunas regiones. Exactamente, se han secuenciado 2.352 regiones del ADN que son importantes en el cáncer de mama porque habitualmente suelen tener mutaciones relevantes.

3.2. Preprocesado de datos de secuenciación de ADN

Para ver y analizar las mutaciones de un tumor, es imprescindible limpiar y alinear los datos de secuenciación de ADN en bruto.

Existen dos formatos principales para almacenar los datos de secuenciación: FASTA y FASTQ. El formato FASTA se utiliza para representar únicamente las secuencias. En este proyecto, sin embargo, únicamente se utiliza el formato FASTQ. Este formato es una variante del formato FASTA que permite asociar una medida de calidad a cada base de la secuencia base, es decir, se representa la información de la secuencia junto con valores de calidad.

Por cada muestra, tenemos dos ficheros con el formato FASTQ. En el primero de los ficheros están las lecturas realizadas desde un extremo de cada fragmento de ADN, mientras que el segundo almacena las lecturas realizadas desde el otro extremo del fragmento.

El primer paso es hacer un análisis de calidad de las lecturas. Para ello se ha utilizado el programa FASTQC. Mediante FASTQC podemos realizar algunas comprobaciones de control de calidad de los datos de secuencias sin procesar [17].

A continuación, hay que eliminar las secuencias de adaptadores y las lecturas de baja calidad. Para ello hemos empleado el programa Trimmomatic.

La alineación de secuencias consiste en ordenar dos secuencias de manera que las regiones de similitudes se alinean [13]. En este proyecto se utiliza un alineador de lecturas cortas llamado bwa. Los alineadores de lecturas cortas son herramientas de software diseñadas para alinear un número muy grande de lecturas cortas [13]. Para el alineamiento, en este proyecto se utiliza el genoma humano de referencia hg19. Para que el programa de alineamiento funcione correctamente, se debe indexar ese genoma, es decir, se le hace un índice para luego localizar las secuencias que existen en él más rápido. A continuación se crea la versión binaria ordenada de este fichero.

Para finalizar, se procede al proceso de identificación de variantes (variant calling). En este proceso se comparan las lecturas realizadas en la muestra con la secuencia del genoma de referencia contra el que alinean, de manera que se identifican las posibles mutaciones.

Como ya hemos adelantado, existen dos tipos de mutaciones, las mutaciones germinales y las somáticas. El programa que se ha empleado para el variant calling es Mutect2. Este programa admite, para cada fichero de lecturas, otro fichero de lecturas que procede de una muestra normal (no tumoral), de manera que puede separar las mutaciones somáticas (las que solo están en la muestra tumoral) de las germinales (las que están tanto en la muestra tumoral como en la normal).

Este preprocesado se debe realizar con las 23 muestras.

3.3. Filtrado de mutaciones

El problema de la deconvolución y evolución clonal trata de hallar, a partir de una lista de frecuencias de mutaciones identificadas en una serie de muestras, cual es la composición clonal del tumor y un árbol filogenético que describa un posible proceso evolutivo que ha tenido el tumor. En esta sección se describe cómo obtener dichos datos de entrada al problema a partir de los datos preprocesados.

El fichero VCF contiene información sobre las variantes, tanto de mutaciones de un único nucleótido, como de otro tipo de variantes. En este proyecto únicamente se tiene en cuenta la información sobre las mutaciones de un único nucleótido, en concreto su frecuencia y si es germinal o no, indicando que Mutect2 elimina algunas mutaciones germinales pero no todas y que por eso es necesario hacer este paso de chequear en la base de datos.

Para localizar una mutación germinal, es necesario analizar la columna **ID** del fichero VCF. El valor predeterminado es '.', pero en caso de que la mutación sea germinal, el valor del identificador cambiará. Este valor lo asigna la base de datos SNP (dbSNP). dbSNP contiene variaciones humanas de un solo nucleótido, microsatélites e insercciones y eliminaciones a pequeña escala, junto con la frecuencia de la población, la consecuencia molecular y la información de mapeo genómico. Tal y como se ha explicado en el Capítulo 1, en este proyecto solo es necesaria la información de las mutaciones somáticas, por lo tanto únicamente nos quedaremos con las mutaciones que obtengan como identificador el signo '.'.

A continuación para cada paciente se ha calculado cuántas mutaciones se reiteran en las diferentes muestras. Para ello, en primer lugar se han recolectado todas las mutaciones en un único archivo. Una vez que tenemos la lista entera de todas las mutaciones, archivo por archivo se han contabilizado dichas mutaciones.

En la Tabla 3.1 se muestran los cálculos de cada paciente. Con la finalidad de reducir el tamaño del problema, se ha establecido como criterio que se analicen únicamente las mutaciones que se reiteren en tres muestras o más. Es decir, en la primera paciente se han analizado 172 mutaciones, en la segunda paciente 332 y en la tercera 147.

	1	2	3	4	5	6	7	8
Primera paciente	164287	2255	143	21	7	1	0	0
Segunda paciente	10985	894	230	74	25	2	1	-
Tercera paciente	16659	1015	125	17	5	-	-	-

Tabla 3.1: Número de mutaciones repetidas en las diferentes cantidades de muestras.

Con estas mutaciones, para cada paciente se ha construido la matriz F. Las filas de esta matriz representan las muestras tumorales de cada paciente. Es decir, la matriz F de la primera paciente tiene 8 filas (8 muestras de tejido tumoral), la de la segunda paciente 7 y la de la tercera paciente 5. Las columnas, en cambio, representan las mutaciones. Así, los valores de la matriz F indican la frecuencia de cada mutación en cada muestra. En la Tabla 3.2 se pueden observar algunos de los valores de las frecuencias obtenidas de la primera paciente. Por otro lado, en la Figura 3.1 se puede observar que la mayoría de los valores de las frecuencias oscilan entre los valores $(0, 0.1]$. La mayoría de los valores, en concreto 822 valores, son nulos, lo que significa que dicha mutación no se encuentra en la muestra.

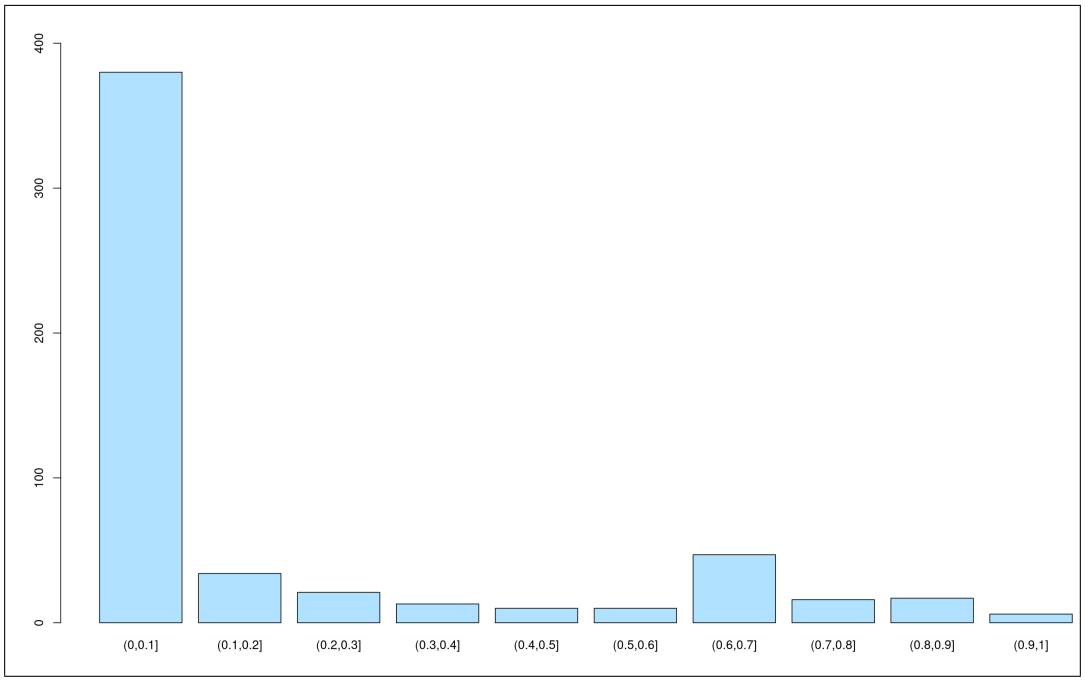


Figura 3.1: Cantidad de mutaciones para distintos rangos de frecuencias en la primera paciente

	M1	M2	M3	M4	M5	M6	M7	M8	...	M172
1	0	0	0.75	0.016	0	0	0	0	...	0
2	0	0.667	0	0	0	0	0	0	...	0.113
3	0.237	0	0	0	0.022	0.045	0.385	0	...	0
4	0	0	0	0	0	0	0	0	...	0
5	0	0.091	0.067	0	0	0.015	0.091	0.075	...	0.057
6	0	0.091	0.067	0.00156	0.039	0 0.091	0	0	...	0
7	0.006803	0	0	0.002023	0.056	0.004092	0.091	0.017	...	0
8	0.007306	0	0	0	0	0	0	0	...	0.05

Tabla 3.2: Parte de la matriz F de la primera paciente.

La matriz F es la entrada al problema de la deconvolución clonal.

CAPÍTULO 4

Análisis y comparación de los resultados

En esta sección se ha resuelto el problema de la deconvolución y evolución clonal para las tres pacientes y se han comparado los resultados.

Para ello, en primer lugar a partir de las matrices F del Capítulo 3 se han obtenido los árboles evolutivos, es decir, se ha resuelto el problema de la deconvolución y evolución clonal. Cabe recordar que el problema de la deconvolución y evolución clonal trata de hallar un árbol evolutivo tumoral a partir de unas muestras del tumor, teniendo en cuenta que cada muestra contiene una mezcla desconocida de clones en proporciones que tampoco son conocidas y que las frecuencias que se observan para una mutación en una muestra son resultado de esa mezcla, es decir, la suma de proporciones de los clones que contienen dicha mutación.

Para cada paciente o matriz F se han realizado diez ejecuciones del algoritmo heurístico que resuelve el problema, de manera que se han obtenido diez árboles para cada paciente. El objetivo final es hallar patrones evolutivos comunes, para lo que se ha implementado y aplicado el algoritmo de Maximum Spanning Tree (MST).

4.1. Generación de árboles consenso

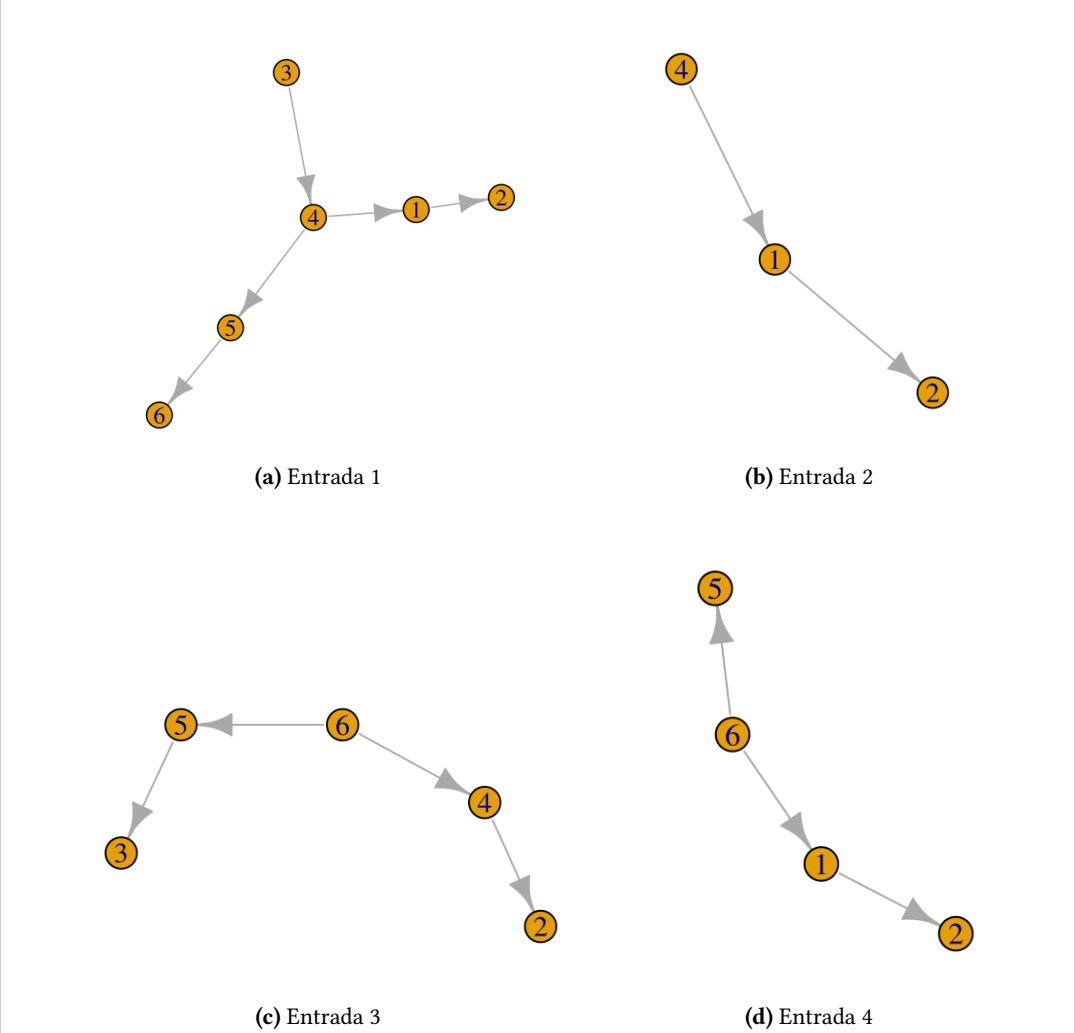
Para la generación de árboles consenso a partir de múltiples árboles se ha implementado el algoritmo MST.

Un Maximum Spanning Tree es un subconjunto del grafo G que incluye todos los atributos con el mínimo número de aristas. El objetivo de dicho algoritmo es hallar el camino que una todos los vértices con mayor peso.

Para poder resolver el problema del MST, en primer lugar, dado un conjunto de múltiples árboles, se crea un grafo dirigido ponderado que incluya todas las relaciones padre-hijo donde cada una tiene asociada como peso el número de árboles que la incluyen. Es decir, dado un conjunto $S = \{T_1, T_2, ..., T_n\}$ de árboles, se construye un grafo que represente todas las relaciones padre-hijo. El grafo $G = (V, E)$ es un grafo conexo y dirigido ponderado

con vértices y aristas directas. Cada arista se pondera con $count(u, v)$ donde $count(u, v)$ es el número de árboles $T \in S$ en los que el vértice u es padre del vértice v. Para ello, se obtienen las matrices adyacentes de cada árbol y la suma entre ellas. De esta manera, logramos la matriz del árbol con sus respectivos pesos.

En la Figura 4.1 se puede observar los mútiples árboles de entrada y en la Figura 4.2 el grafo creado a raíz de estos árboles.



(a) Entrada 1 **(b)** Entrada 2

(c) Entrada 3 **(d)** Entrada 4

Figura 4.1: Múltiples árboles de entrada al problema del MST

Una vez hemos obtenido un grafo conexo y dirigido, un spanning tree es un árbol formado con aristas del grafo y con todos los nodos. Cada spanning tree tiene un peso asociado (suma de los pesos de sus arcos), con lo que el MST es el spanning tree que tiene mayor peso.

Para hallar el MST en el grafo obtenido, se ha utilizado algoritmo Prim. El algoritmo Prim es un algoritmo que se basa en la idea de que un árbol debe tener todos sus vértices conectados. El algoritmo funciona construyendo el árbol de vértice en vértice a partir de un vértice inicial arbitrario, añadiendo la conexión con más peso posible desde el árbol a otro vértice. El resultado proporciona el MST.

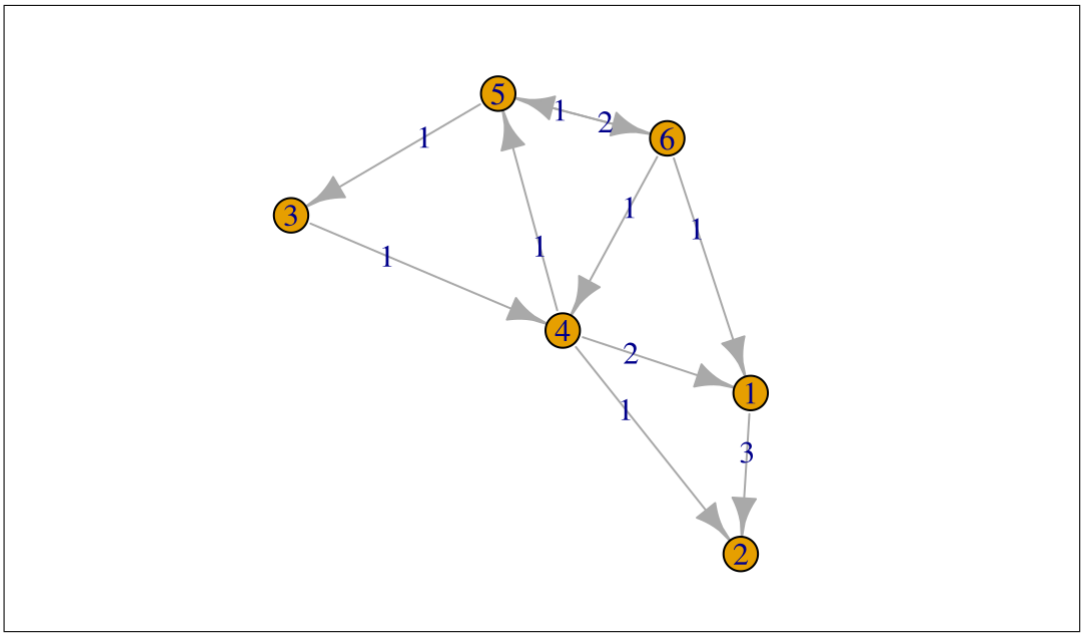


Figura 4.2: Grafo G creado a raíz de los árboles de la Figura 4.1

Durante el algoritmo es necesario controlar tanto los vértices visitados como el padre de cada vértice para evitar crear ciclos. Dicha información se guarda en un vector, cuya función es imprescindible para finalmente crear el MST.

La Figura 4.3 muestra el resultado de aplicar este algoritmo a los árboles de la Figura 4.1.

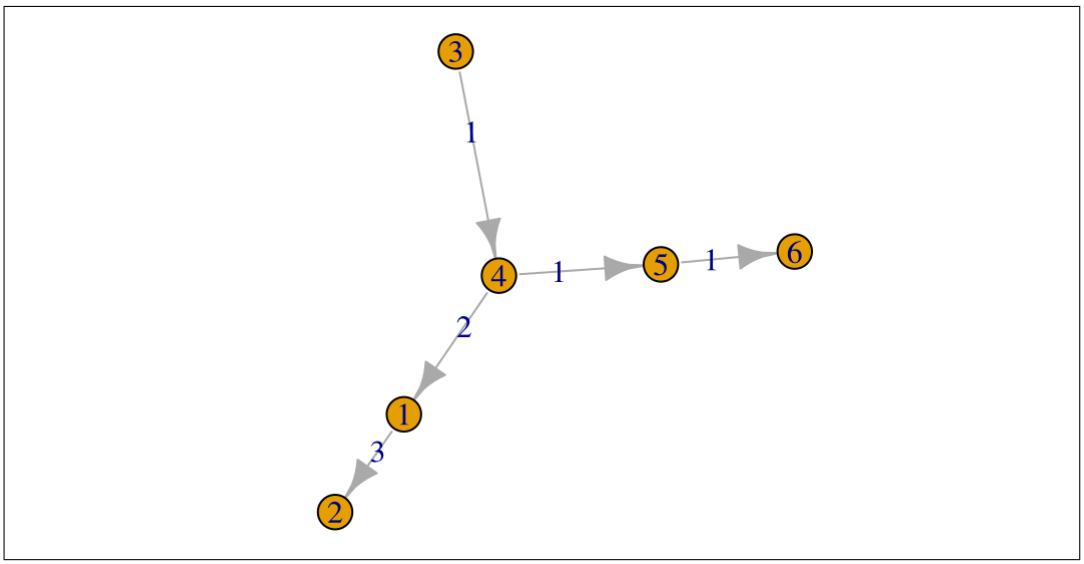


Figura 4.3: Resultado de algoritmo implementado, MST, a raíz de los árboles de la Figura 4.1

4.2. Análisis y comparación los resultados

Utilizando las matrices F y ejecutando diez veces el algoritmo heurístico a cada una de ellas, hallamos varias soluciones al problema de la deconvolución clonal. Concretamente, en cada solución obtenemos un árbol evolutivo tumoral y un error asociado.

El valor del error indica la diferencia que hay entre el árbol obtenido a través del algoritmo heurístico y el óptimo real (el árbol real), es decir, dicho valor es reflejo de su calidad (a menor error, mayor calidad). En primer lugar, hay que observar si existe algún valor atípico (*outlier*) en los árboles creados. Un valor atípico es una observación que se encuentra a una distancia inusual de los otros valores en una muestra.

Para ello en la Figura 4.4a se puede analizar que los errores de los objetos de la primera paciente oscilan entre los valores 0.06 y 0.08, lo que significa que no existe ningún árbol con valores atípicos. Es decir, todos los resultados creados para la primera paciente con el método heurístico son válidos. Los errores asociados a los árboles obtenidos de la segunda y tercera paciente se reflejan en las Figuras 4.4b y 4.4c, por lo que se puede afirmar que no hay ninguna solución posiblemente inválida en todos los árboles obtenidos.

Al comprobar que no existe ningún valor atípico, el siguiente paso es calcular la distancia entre los árboles para identificar la similitud entre ellos. Gracias a los valores de las distancias se puede deducir la similitud entre los árboles filogenéticos. La distancia que hemos empleado está basada en la cantidad de aristas independientes de los árboles que queremos comparar. Por ejemplo, imaginemos que tenemos un par de árboles llamados T_1 y T_2. La distancia entre ambos árboles puede calcularse como la suma de las aristas de T_1 que no existen en T_2 y las aristas independientes de T_2, que es igual a restar el número de aristas comunes al número total de aristas diferentes. Esta distancia se calcula tal y como se indica en la Ecuación 4.1. En la Figura 4.5 se muestra un ejemplo del cálculo de la distancia entre dos árboles filogenéticos.

$$d = |E_{T_1 \cup T_2}| - |E_{T_1 \cap T_2}| \tag{4.1}$$

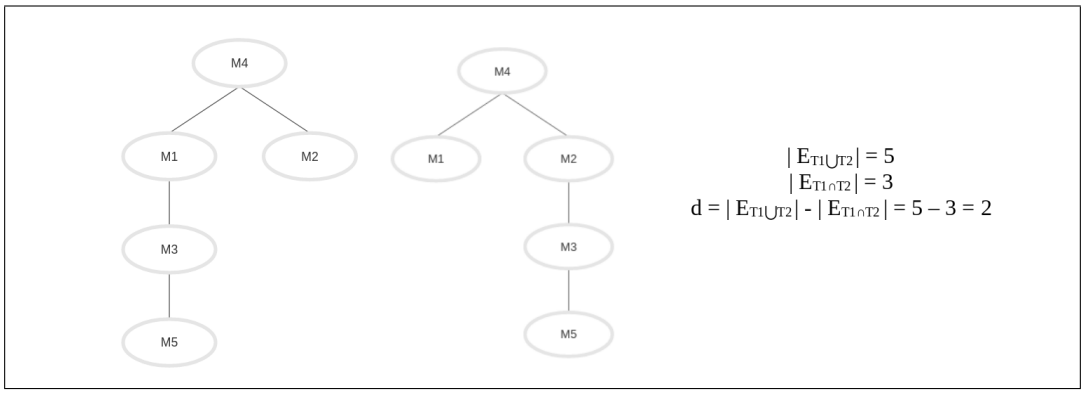


Figura 4.5: Ejemplo del cálculo de la distancia entre árboles filogenéticos

De esta manera, árboles filogenéticos iguales con las mismas relaciones ancestrales tendrán como valor de la distancia 0. Es decir, cuanto menor sea el valor de la distancia, mayor similitud habrá entre dos árboles.

En el caso de la primera paciente, las distancias se pueden observar en la Tabla 4.1. Los valores oscilan entre 298 y 322.

Las distancias obtenidas de la segunda paciente oscilan entre los valores 524 y 552. Y los valores de la tercera paciente entre 272 y 284. Cabe destacar que cada árbol tiene $N - 1$ arcos, por lo que significa que las soluciones obtenidas son muy diferentes. La conclusión general en las tres pacientes es que las distancias obtenidas en todos los casos son muy grandes, es decir, no hay mucha similitud entre los árboles obtenidos.

	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10
A1	0	320	314	312	320	314	318	304	310	320
A2	-	0	318	304	318	314	302	312	312	316
A3	-	-	0	320	318	310	320	310	314	312
A4	-	-	-	0	316	308	308	314	308	314
A5	-	-	-	-	0	322	308	318	312	318
A6	-	-	-	-	-	0	316	314	310	324
A7	-	-	-	-	-	-	0	310	298	320
A8	-	-	-	-	-	-	-	0	314	320
A9	-	-	-	-	-	-	-	-	0	312

Tabla 4.1: Distancias entre los árboles de la primera paciente. Los valores $A_1, A_2, ..., A_{10}$ representan cada árbol obtenido a raíz del algoritmo heurístico

A la hora de analizar los árboles filogenéticos existen diferentes maneras para hacer el análisis. El objetivo de este proyecto es encontrar patrones evolutivos comunes. La idea original en este proyecto consistía en hallar el MST de cada paciente y finalmente los subárboles comunes. Por desgracia, no se ha podido llevar a cabo este segundo paso ya que únicamente hay cinco mutaciones comunes a las tres pacientes, con lo que no se lograría ninguna conclusión buscando el subárbol común. En su lugar, se han coloreado esas cinco mutaciones en los tres MST. Los análisis se han hecho únicamente con estas cinco mutaciones.

En la Tabla 4.2 se pueden observar las cinco mutaciones comunes a las tres pacientes y la información que corresponde a cada una de ellas. Es decir, a qué cromosoma pertenece, en qué posición se encuentra, cuál es la base de la secuencia de referencia y cuál es la base de la variante.

CHROM	POS	REF	ALT
1	121484771	T	C
7	61968901	T	G
7	61968931	T	A
8	37551328	A	G
19	27732040	A	G

Tabla 4.2: Mutaciones comunes de las tres pacientes

Es necesario determinar qué identificador identifica cada mutación en las diferentes pacientes (Tabla 4.3). En dicha tabla se indica qué color representa cada mutación en los árboles obtenidos.

CHROM	POS	Paciente 1	Paciente 2	Paciente 3
1	121484771	10	20	9
7	61968901	58	80	45
7	61968931	59	81	46
8	37551328	86	128	61
19	27732040	159	311	142

Tabla 4.3: Identificadores de las mutaciones comunes en las tres pacientes

Con los árboles tumorales evolutivos obtenidos con el algoritmo heurístico, se ha aplicado el algoritmo MST implementado en este proyecto. En la Figura 4.6a se resume el MST obtenido de la primera paciente. En dicha figura únicamente se muestran las mutaciones comunes entre las tres pacientes (mutaciones a las que identificamos con los identificadores 10, 58, 59, 86, 159), ya que las demás variantes no son de importancia para el análisis de la heterogeneidad intratumoral. Cada nodo de color indica una diferente mutación común, en cambio, la raíz del árbol se muestra de color blanco. Los nodos con línea discontinua indican la cantidad de mutaciones que se hallan de un nodo a otro. En dicho árbol, se puede observar que para cada nodo se ramifica una rama diferente y que la mayoría de las mutaciones tienen una única mutación que les une a la raíz.

La Tabla 4.4 representa las distancias que hay entre las diferentes mutaciones. Esta distancia está basada en el número de saltos que se deben dar en el árbol para ir de un nodo a otro, es decir, en este caso se analiza la distancia entre nodos dentro de un mismo árbol. En dicha tabla, se puede analizar que la mutación con el identificador 10 se encuentra la más cercana a las demás, dicha mutación se sitúa en el cromosoma 1.

	10	58	59	86	159
10	0	3	3	3	3
58	-	0	4	4	4
59	-	-	0	4	4
86	-	-	-	0	4

Tabla 4.4: Distancias entre las mutaciones comunes a las tres pacientes en el MST de la primera paciente

El resumen del MST obtenido con los datos de la segunda paciente se muestra en la Figura 4.6b . Por otro lado, en la Tabla 4.5 se hace ver la distancia que hay entre las mutaciones. En dicha tabla se puede observar que las mutaciones más cercanas son aquellas con los identificadores 20 y 80, las cuales están situadas en los cromosomas 1 y 7.

	20	80	81	128	311
20	0	3	7	6	8
80	-	0	6	5	7
81	-	-	0	7	7
128	-	-	-	0	8

Tabla 4.5: Distancias entre las mutaciones comunes a las tres pacientes en el MST de la segunda paciente

El MST de la tercera paciente se muestra en la Figura 4.6c y en ella se puede observar

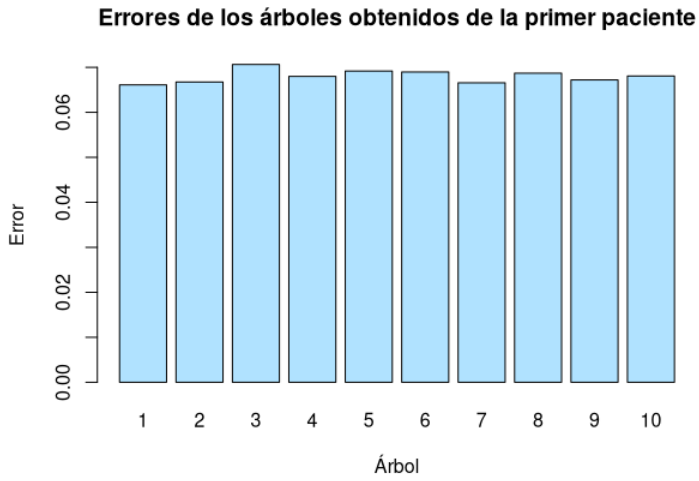
que cada mutación no se encuentra en una ramificación diferente y que existen varias mutaciones que unen las diferentes variantes comunes a la raíz. La mutación que se encuentra en el cromosoma 19 se sitúa con mayor distancia de la raíz. En la Tabla 4.6 se muestran las distancias entre las mutaciones comunes. Las mutaciones con menor distancia entre ellas son aquellas con los identificadores 9 y 46, es decir, las situadas en los cromosomas 1 y 7.

	9	45	46	61	142
9	0	13	4	13	9
45	-	0	11	8	16
46	-	-	0	11	7
61	-	-	-	0	16

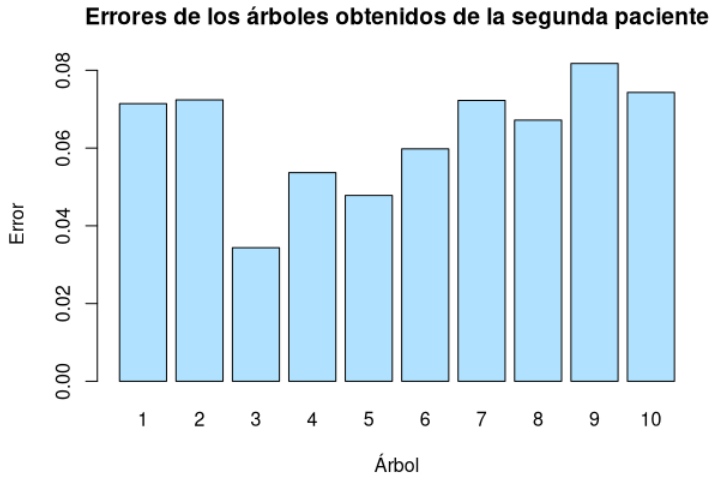
Tabla 4.6: Distancias entre las mutaciones comunes a las tres pacientes en el MST de la tercera paciente

Cabe destacar que en las tres pacientes la mutación que se sitúa en el cromosoma 19 se encuentra a mayor o igual distancia que las demás mutaciones de la raíz. Al contrario, en general, la mutación que se sitúa en el cromosoma 8 se encuentra a menor distancia de la raíz. Una mayor distancia a la raíz indica una aparición de la mutación más tardía.

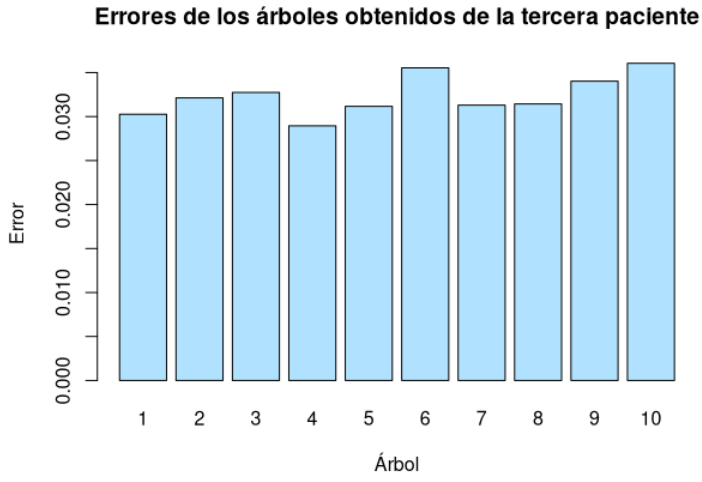
Una vez obtenidos todos los resultados, es preciso mencionar que no se ha hallado un patrón claro. El único patrón que se percibe ligeramente es el de la mutación situada en el cromosoma 8 (color naranja). Dicha mutación parece situarse separada de las demás mutaciones. En cualquier caso, no es un patrón concluyente.



(a) Paciente 1

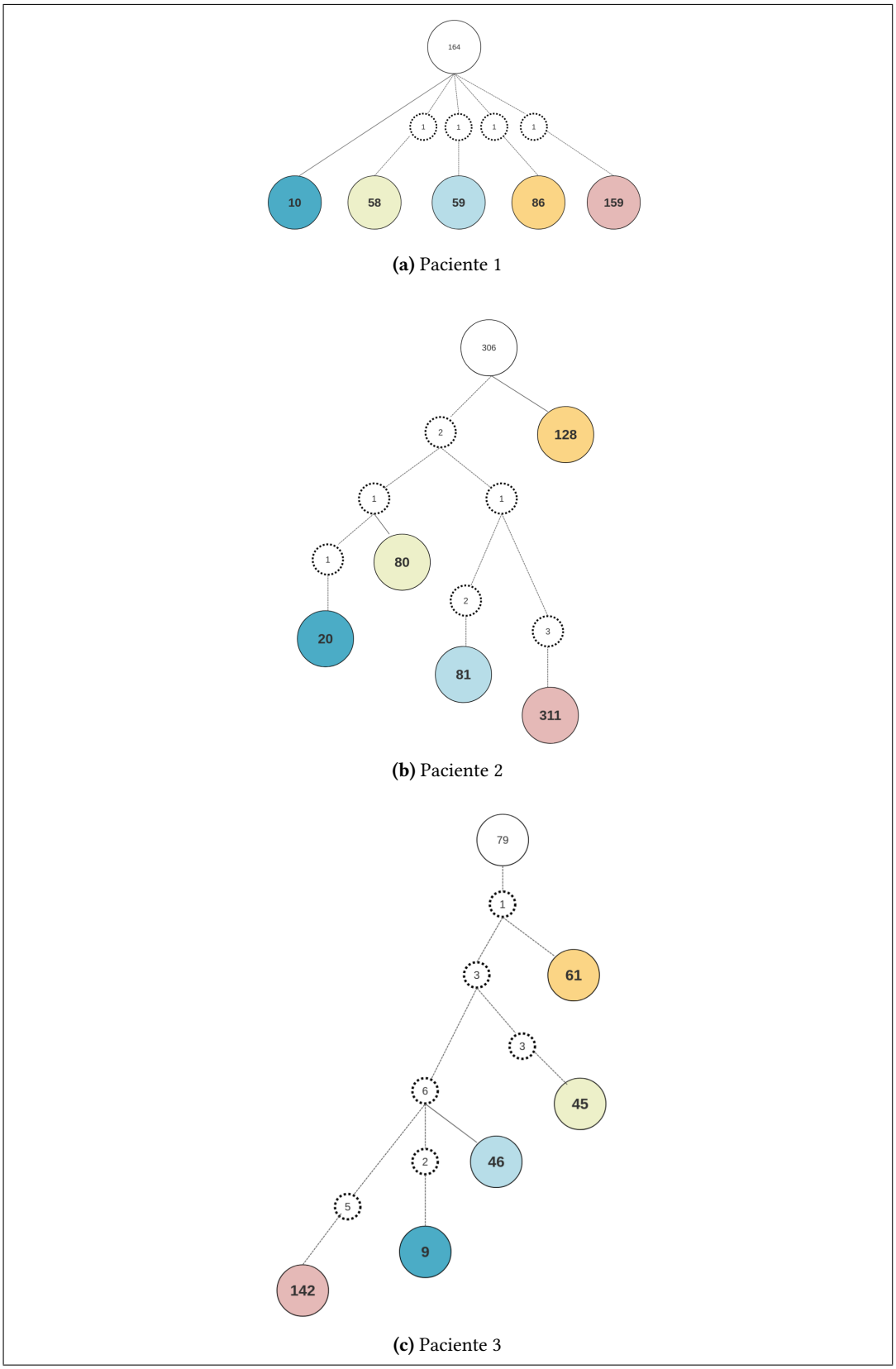


(b) Paciente 2



(c) Paciente 3

Figura 4.4: Error de los diez árboles obtenidos de cada paciente



(a) Paciente 1

(b) Paciente 2

(c) Paciente 3

Figura 4.6: MST de cada una de las pacientes en los que únicamente se han ilustrado las cinco mutaciones comunes a las tres pacientes además del nodo raíz. Los nodos con línea discontínua indican la cantidad de mutaciones que se hallan de un nodo a otro

CAPÍTULO 5

Conclusiones y trabajo futuro

En este capítulo se presentan las conclusiones obtenidas del trabajo así como las líneas futuras de mejora.

5.1. Conclusiones generales

En este Trabajo de Fin de Grado se planteaba como objetivo analizar la heterogeneidad intratumoral e intertumoral en cáncer de mama a partir de múltiples muestras de diferentes individuos para identificar patrones de evolución. Debido al tiempo limitado y a la limitación de datos, tanto en cantidad como en calidad, el análisis ha sido más limitado de lo esperado como se ha especificado más adelante.

Para analizar la heterogeneidad intratumoral en primer lugar ha sido necesario preprocesar las muestras de secuenciación de ADN de los tumores. En ese proceso se ha obtenido la lista de mutaciones presentes en cada tumor junto con su frecuencia. Tras realizar una serie de filtrados para reducir el número de mutaciones a utilizar en el estudio, se ha utilizado un algoritmo de búsqueda local existente para resolver el problema de la deconvolución clonal y así hallar varios árboles tumorales evolutivos para cada paciente. Por último, se ha implementado y utilizado un algoritmo para obtener el árbol consenso de todas las soluciones obtenidas para cada paciente.

Los resultados obtenidos no son, en gran medida, los esperados. En primer lugar, porque el objetivo último era la búsqueda de patrones evolutivos comunes entre las tres pacientes, pero el número de mutaciones comunes ha resultado ser muy reducido (en concreto, cinco). Una de las posibles explicaciones es que muchas de las mutaciones identificadas no sean realmente mutaciones, sino artefactos introducidos por la conservación en parafina. Otra posible explicación es que efectivamente la evolución de los tumores de las tres pacientes no tengan demasiados elementos en común.

Otro de los motivos por los que los resultados no son los esperados es la gran heterogeneidad de las diferentes soluciones obtenidas por el método heurístico, siendo todas ellas de similar calidad. Esa heterogeneidad puede ser debida a diferentes motivos. Uno de ellos es el hecho de haber considerado conjuntamente muestras del tumor primario y de las metástasis. En cualquier caso, dada una matriz de frecuencias de mutaciones pueden

existir multitud de árboles evolutivos que expliquen igualmente bien los datos, hecho que también puede explicar la heterogeneidad observada.

5.2. Trabajo futuro

Una de las líneas para tratar de mejorar este análisis es añadir una etapa en el preprocesado para tratar explícitamente el problema de los artefactos introducidos por la parafina, haciendo uso de herramientas para ello [18].

Otra de las posibles mejoras estaría en el propio algoritmo para resolver el problema de la deconvolución y evolución clonal. Recientes estudios apuntan a que el uso de secuencias más grandes o, incluso, el uso de datos de secuenciación de tipo *single cell* pueden ayudar a reducir el número de posibles soluciones al problema [19].

Una vez obtenemos las mutaciones de las pacientes, y junto a ello, las tres matrices, tal y como se ha hecho en este proyecto, se obtendrían varios árboles tumorales evolutivos para cada paciente utilizando un algoritmo de búsqueda local existente. Sería interesante analizar para cada solución en qué parte del árbol se concentra el error, ya que dicho error representa la calidad de cada uno de ellos. Con dichos árboles existen varios métodos para elaborar el análisis.

Se podrían analizar los árboles evolutivos obteniendo los árboles consenso. Es decir, podríamos obtener un árbol consenso para cada paciente, y a continuación, hallar un único árbol. De esta manera, se podría analizar si la evolución de dicho tumor puede combinarse para inferir una historia evolutiva mejor. Para ello, se podría utilizar la función *combine_trees* implementada en el paquete GeRnika. Dicho árbol recolectaría todas las relaciones ancestrales entre los clones de los árboles. El árbol de consenso recogería cuatro tipos de relaciones distintas entre nodos: las relaciones ancestrales comunes presentes en todos los árboles, es decir, las que componen los subárboles comunes, las relaciones evolutivas de la primera paciente que no existen en las demás, las relaciones ancestrales independientes de la segunda paciente y las de la tercera.

En lugar de analizar la heterogeneidad intertumoral con los árboles consenso, se podrían utilizar los árboles comunes, es decir, en primer lugar se hallaría un árbol común para cada paciente y finalmente un único árbol para las tres. En los casos en que exista un elevado número de árboles que expliquen los datos aún siendo muy diferentes entre sí, esta aproximación no será adecuada. El mejor ejemplo es lo que ocurre en el caso de la tercera paciente, donde hemos observado que no existe ningún árbol común.

Otro método diferente para realizar el análisis sería aplicando el algoritmo implementado en este proyecto, es decir, el algoritmo que halla el MST de múltiples árboles. Como en los métodos anteriores, sería necesario hallar un MST para cada paciente, y a continuación un único árbol. De esta manera, se podría analizar qué relaciones entre mutaciones se reiteran con mayor frecuencia.

Con el método para obtener los árboles implementado se define un clon por mutación. Esto hace que, en presencia de muchas mutaciones, los árboles sean excesivamente grandes y complejos. Algunos métodos existentes crean clusters de mutaciones para reducir el problema. Una línea interesante a futuro sería desarrollar aproximaciones para simplificar los árboles, planteando, por ejemplo, clusterizar mutaciones (clones) una vez obtenidos los árboles.

Por otro lado, sería interesante hacer un análisis biológico con los datos obtenidos.

El análisis biológico es un enfoque científico que combina herramientas analíticas y contenido biológico en un solo lugar, para que los investigadores puedan obtener una compresión fundamentalmente más profunda y amplia de las relaciones y procesos biológicos que se vinculan con las observaciones.

El análisis biológico puede convertir los resultados del análisis de datos básicos en resultados útiles para la investigación, de manera que los investigadores puedan usar sus hallazgos para tomar decisiones informadas, generar hipótesis estructuradas, diseñar experimentos de seguimiento y proporcionar evidencia biológica.

5.3. Reflexión personal

A nivel personal, el desarrollo de este proyecto ha sido muy satisfactorio. Por un lado, estoy satisfecha de haber podido realizar el trabajo con mis directores, ya que debido a su experiencia en este tipo de proyectos, han podido ayudarme en todo momento.

Pese a que el objetivo principal del trabajo no se haya efectuado al completo debido al tiempo limitado y a la limitación de datos, tratar un tema tan interesante como los tumores ha sido una gran motivación a lo largo del proyecto. Por otro lado, he podido tomar conciencia del trabajo que hay en la investigación de este área y lo importante que es para poder avanzar y mejorar el bienestar social.

Cabe destacar que al inicio del proyecto mi conocimiento en el área de la bioinformática era muy limitado. El desarrollo de este proyecto me ha ayudado a adquirir una visión general e interés para en un futuro poder seguir investigando en temas relacionados.

APÉNDICE A

Preprocesado de datos genómicos

A.1. Limpieza de los datos

Para poder ver y analizar las mutaciones de un tumor, es necesario limpiar y alinear los datos en bruto.

Existen dos formatos principales para almacenar estos datos en bruto: FASTA y FASTQ.

El formato FASTA es el *workhorse* de la bionformática. Se utiliza para representar la información de secuencias biológicas. El formato es sencillo. Cada registro consta de dos líneas: una con la cabecera y la otra con la secuencia. Más en concreto:

- Un símbolo '>' en la línea de la cabecera FASTA indica el inicio de un registro fasta.
- El símbolo '>' puede ir seguido de una cadena de letras llamada ID de la secuencia.
- La línea de cabecera puede contener una cantidad arbitraria de texto (incluyendo espacios) en la misma línea.
- Las líneas siguientes contienen la secuencia.

La secuencia se representa con un alfabeto que corresponde a una entidad biológica. Por ejemplo, el alfabeto estándar para los nucleótidos es ATGC. Un alfabeto ampliado puede contener también una N para indicar una base que podría ser cualquiera de ATGC. Es decir, la N significa que no se ha conseguido secuenciar ese nucleótido [13].

El formato FASTQ es una variante del formato FASTA que permite asociar una medida de calidad o fiabilidad a cada base de la secuencia, es decir, su estructura es como la de un formato FASTA pero con dos líneas adicionales que indican estos valores de calidad. Así el formato FASTQ consta de cuatro secciones [13]:

- Una cabecera similar a la de FASTA, pero en lugar del símbolo '>' se utiliza el símbolo '@' acompañado con un ID y un texto opcional.
- La segunda sección contiene la secuencia (normalmente en una sola línea).

- La tercera sección comienza con el símbolo '+' y opcionalmente puede ir seguida del mismo identificador de secuencia y la cabecera de la primera sección.
- La última línea codifica los valores de calidad de la secuencia de la segunda sección, por lo que tiene su misma longitud.

En la Figura A.1 se muestra un ejemplo de una entrada de un fichero FASTQ.

```
@SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65
```

Figura A.1: Ejemplo del formato FASTQ de Wikipedia

En este proyecto, los datos estaban en formato FASTQ.

Al secuenciar un fragmento de ADN, en general se secuencia dos veces: la primera empezando por un extremo, cuyo contenido se almacena en un primer fichero al que llamaremos R1 y una segunda vez empezando por el otro extremo, cuyo contenido se almacena en un segundo fichero al que llamaremos R2.

Como anteriormente hemos explicado, el formato FASTQ contiene datos de calidad para cada nucleótido, lo que nos permite hacer un análisis de calidad de las lecturas. Para ello se ha utilizado la herramienta FastQC.

```
file=R1.fastq.fq.gz
fastqc $file
```

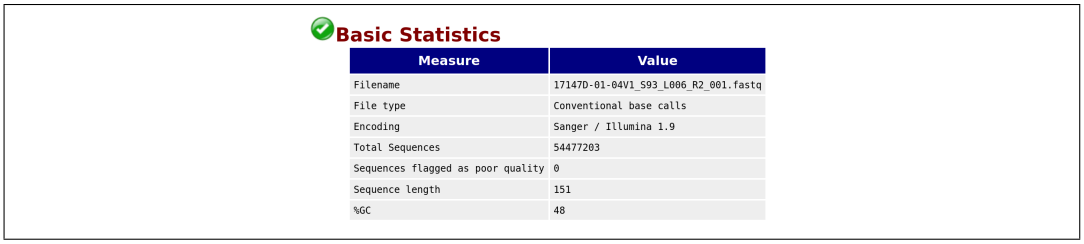
De aquí obtenemos un fichero html con una serie de métricas sobre la muestra (uno para las lecturas en R1 y otro para las lecturas en R2).

El objetivo de FASTQC es proporcionar una forma sencilla de realizar algunas comprobaciones de control de calidad de los datos de secuenciación sin procesar. Proporciona un conjunto modular de análisis que facilita una rápida impresión de si los datos tienen algún problema del que deberíamos ser conscientes antes de realizar cualquier otro análisis. Las principales funciones de FastQC son [17]:

- Importación de datos desde archivos BAM, SAM o FASTQ.
- Proporcionar una visión general rápida para indicar en qué áreas puede haber problemas.
- Gráficos y tablas de resumen para evaluar rápidamente los datos.
- Exportación de los resultados a un informe permanente basado en HTML.
- Funcionamiento sin conexión para permitir la generación automática de informes sin necesidad de ejecutar la aplicación interactiva.

Cada fichero de salida se divide en varias secciones, y en cada una de ellas se muestra una característica o aspecto concreto de las lecturas. A continuación se analizan algunos de estos apartados para las lecturas R2 de una de las muestras analizadas.

- **Basic Statistics.** Las estadísticas básicas indican qué archivo ha sido preprocesado, qué tipo de archivo es, cuántas secuencias se han procesado, cuál es la longitud de la secuencia y cuál es el porcentaje de GC de las lecturas. En la Figura A.2 se puede observar las estadísticas de una muestra en concreto.



Basic Statistics

Measure	Value
Filename	17147D-01-04V1_S93_L006_R2_001.fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	54477203
Sequences flagged as poor quality	0
Sequence length	151
%GC	48

Figura A.2: FastQC. Basic Statistics de la muestra

- **Per base sequence quality.** La calidad de la secuencia se muestra a través de un gráfico. En el eje X se indica la posición de los nucleótidos en la lectura (que son todas de longitud 150) y el eje Y muestra la distribución de los valores de calidad para los nucleótidos en cada posición.

  La caja amarilla indica el rango intercuartil. La línea roja dentro de las cajas indica el valor de la mediana, es decir, el percentil 50. La línea azul indica la puntuación de calidad media para cada posición. Se suele considerar la lectura de buena calidad por encima de 20.

  Por lo tanto, podemos apreciar que la muestra de la Figura A.3 es de buena calidad.

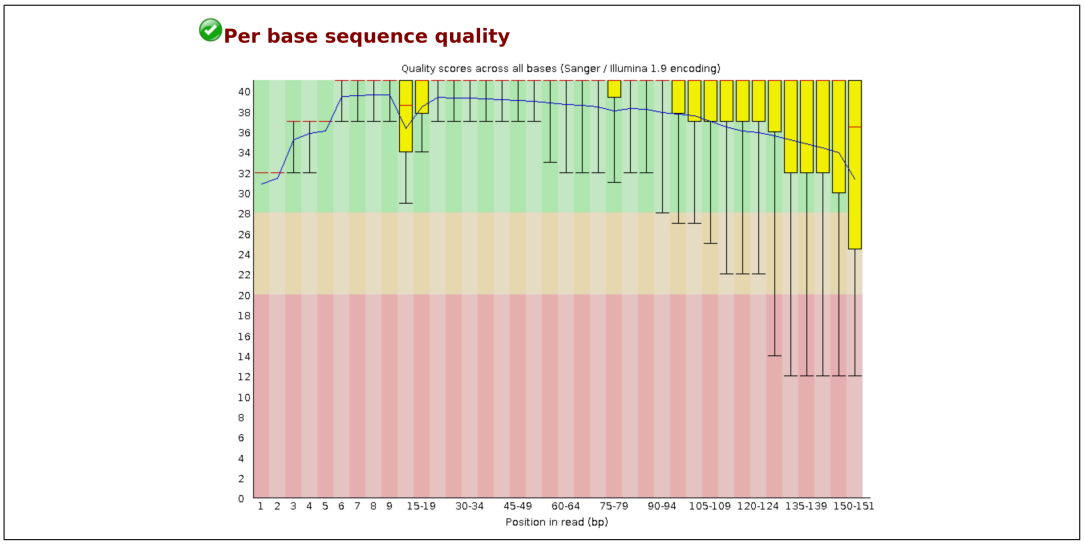


Figura A.3: FastQC. Per base sequence quality de la muestra

- **Per sequence quality scores.** Esta métrica calcula la puntuación media de todas las bases en cada secuencia y muestra su distribución. En la Figura A.4 se puede apreciar que la muestra es de buena calidad. La mayoría de las secuencias tienen una calidad muy alta (cercana a 40).

- **Per base sequence content.** Esta métrica muestra la distribución media del contenido en cada posición de las lecturas. En el contenido de la secuencia por base se

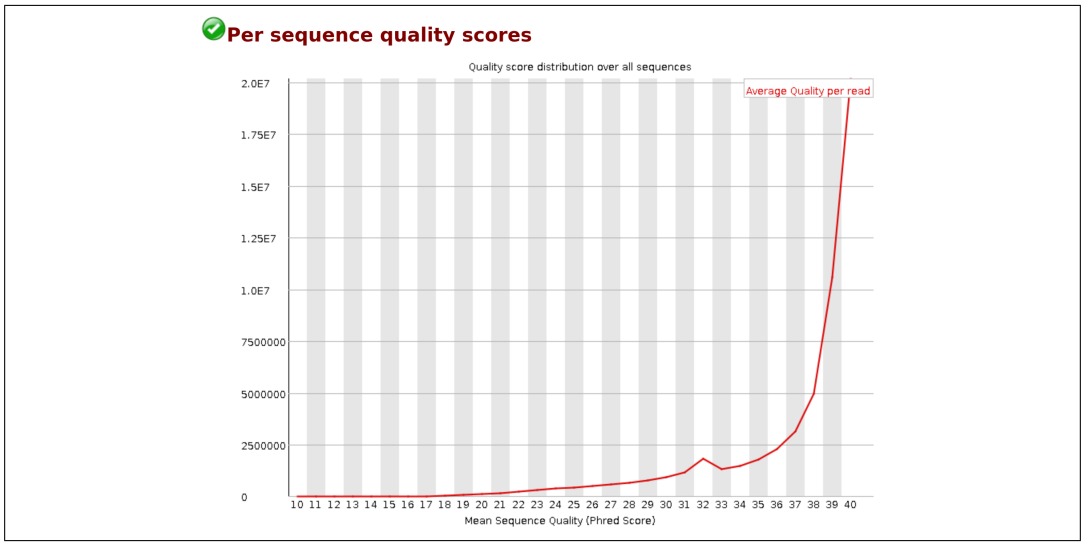


Figura A.4: FastQC. Per sequence quality scores de la muestra

debería apreciar una distribución uniforme de las cuatro bases que no cambia con la posición de la base, es decir líneas paralelas. Esto es debido a que el contenido de nucleótidos debería ser uniforme a lo largo de las lecturas. En este caso se puede apreciar que las líneas no son uniformes, sino que existe cierta alteración en torno a las posiciones 15-19 y también hacia el final de las lecturas, por lo tanto no es lo optimo aunque no se espera que este tipo de artefactos dificulten en gran medida la tarea de análisis posterior. En la Figura A.5 se muestra dicha métrica.

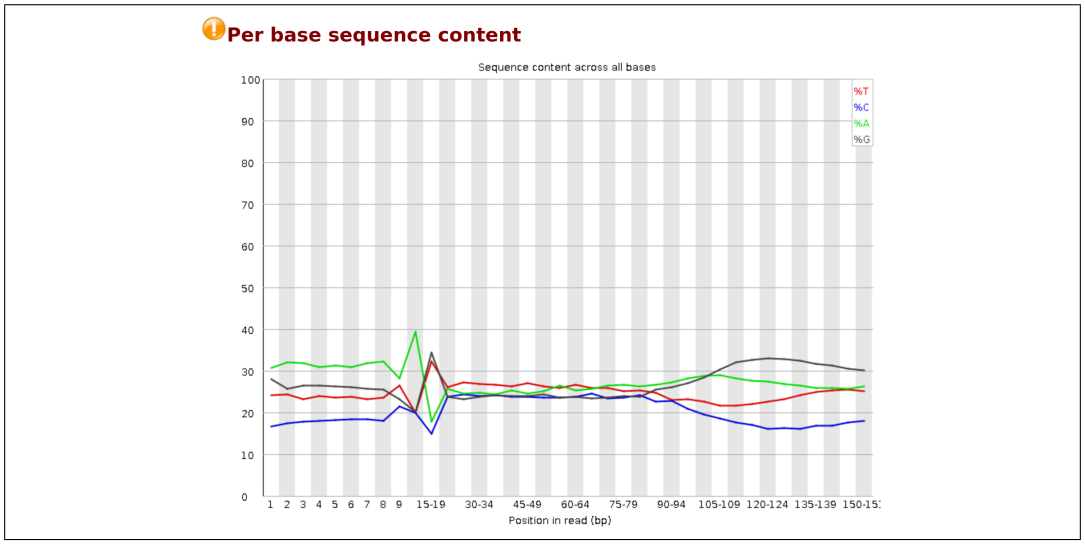


Figura A.5: FastQC. Per base sequence content de la muestra

- **Per sequence GC content.** En el contenido de GC de la secuencia se pueden observar dos líneas. La línea roja indica la distribución de nuestra muestra. La línea azul en cambio, indica una distribución teóricamente idílica. En la Figura A.6, se puede observar que las dos líneas tienen la misma media y desviación estándar, lo que indica una buena calidad de los datos.

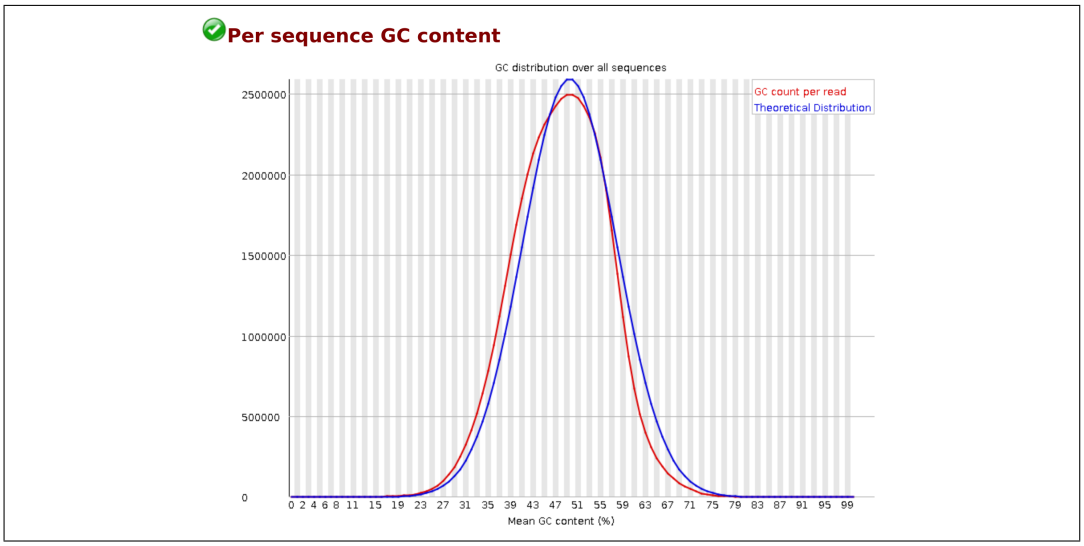


Figura A.6: FastQC. Per sequence GC content de la muestra

- **Per base N content.** En este apartado se indica si existen bases no identificadas en las lecturas. No deberían ser muchas. En la Figura A.7 no hay ninguna, por lo que es lo adecuado.

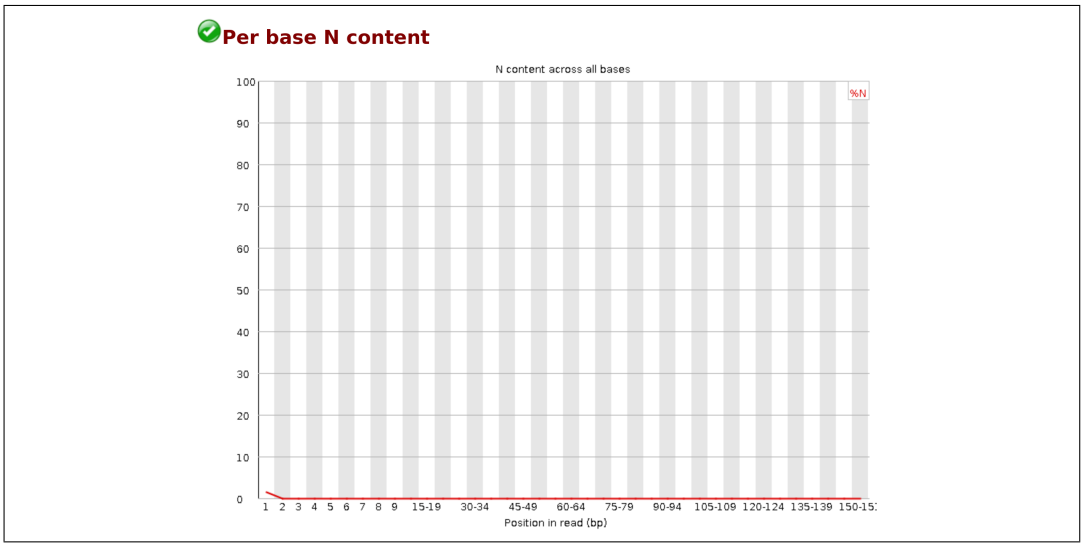


Figura A.7: FastQC. Per base N content de la muestra

- **Sequence Length Distribution.** La distribución de la longitud es un gráfico que indica la distribución de la longitud de las lecturas. En la Figura A.8, todas las lecturas son de longitud 151.

- **Sequence Duplication Levels.** Los niveles de duplicación nos indican cómo de únicas son las secuencias. Idealmente las secuencias sólo se deben leer una vez. Alrededor de un 25 % de las lecturas son únicas, y de las restantes, la mayoría tiene un nivel de duplicación de entre 2 y 10 veces. Niveles de duplicación más allá de estos son poco comunes en estos datos. Dichos datos se pueden observar en la Figura A.9

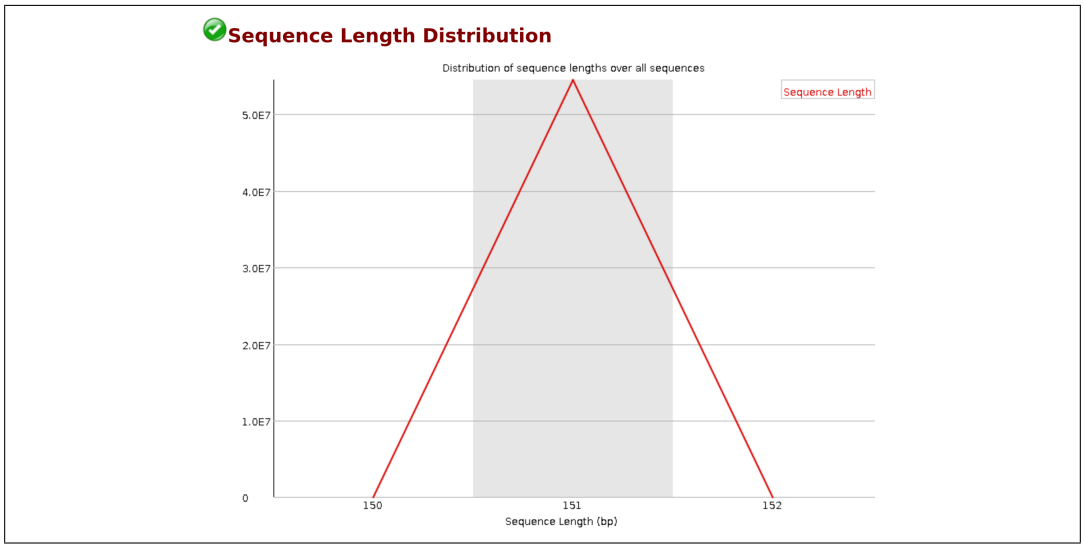


Figura A.8: FastQC. Sequence Length Distribution de la muestra

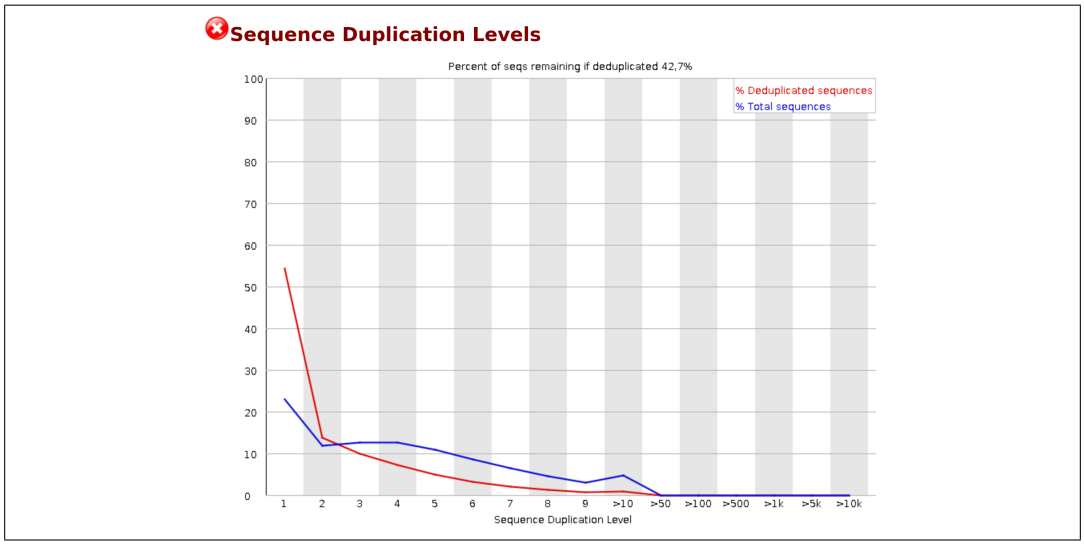


Figura A.9: FastQC. Sequence Duplication Levels de la muestra

- **Overrepresented sequences.** En el apartado de secuencias sobrerrepresentadas se buscan secuencias que están sobrerrepresentadas en el conjunto. En la Figura A.10 se puede observar que no hay ninguna secuencia sobrerrepresentada, que es lo esperado.

Figura A.10: FastQC. Overrepresented sequences de la muestra

Después de hacer el análisis de calidad hay que eliminar las secuencias de adaptadores y las lecturas de baja calidad. Para ello se ha utilizado el programa Trimmomatic. Tal y como se ha explicado en la Sección 1.5, al programa le damos como entrada los dos ficheros

FASTQ de una muestra (R1 y R2) y nos devuelve 4 (2 parejas) ficheros. El código para hacer la limpieza de datos es la siguiente:

```
R1=myreads_R1.fq.gz
R2=myreads_R2.fq.gz
sample="samplename"
java -jar ./../Trimmomatic/trimmomatic/pkg/trimmomatic/opt/Trimmomatic/
    ↪ trimmomatic-0.39.jar PE -threads 6 $R1 $R2 $sample"_trimmed_R1.fastq"
    ↪ $sample"_unpaired_R1.fastq" $sample"_trimmed_R2.fastq" $sample"
    ↪ _unpaired_R2.fastq" ILLUMINACLIP:./../Trimmomatic/trimmomatic/pkg/
    ↪ trimmomatic/opt/Trimmomatic/adapters/TruSeq3-PE-2.fa:2:30:10:10:TRUE
    ↪ SLIDINGWINDOW:4:20 LEADING:3 TRAILING:3 MINLEN:36
```

Una vez hecha la limpieza de los datos, hay que hacer el alineamiento de las lecturas a un genoma de referencia.

A.2. Alineación de los datos

La alineación de secuencias es un concepto esencial para la bioinformática.

La alineación de secuencias (también llamada alineación por pares) consiste en ordenar dos secuencias de manera que las regiones de similitudes se alineen [13].

En este proyecto se ha usado un alineador de lecturas cortas. Los alineadores de lecturas cortas son herramientas de software, diseñadas para alinear un número muy grande de lecturas cortas (miles de millones) contra una secuencia de referencia [13].

Para ello es necesario tener los genomas de referencia. Los genomas de referencia son secuencias de genoma completo, los hay de casi todas las especies. La secuencia no procede de un único humano, sino de múltiples individuos. En este proyecto se ha utilizado el genoma de referencia *hg19*.

Para realizar el alineamiento, en primer lugar, es necesario indexar ese genoma. Al indexar un genoma se le hace un índice para luego saber más rápido dónde buscar una secuencia en él.

```
./bwa index ./../../Genoma/hg19.fa
```

Una vez indexado el genoma se procede al alineamiento, en este caso hemos utilizado el programa de alineamiento bwa-mem (explicado en la Sección 1.5). De esta manera obtenemos las lecturas alineadas:

```
REF=./../../Genoma/hg19.fa
./bwa mem $REF "./../../Datuak/Garbiak/mysamplename_trimmed_R1.fastq" "./../../
    ↪ Datuak/Garbiak/mysamplename_trimmed_R2.fastq" > mysamplename.sam
```

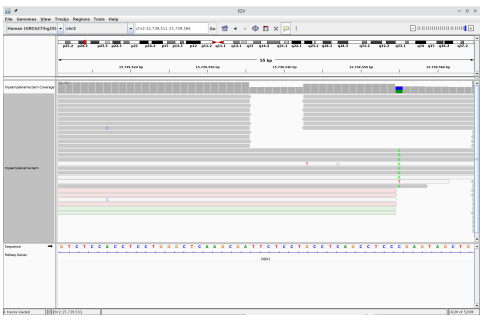
Este proceso es muy costoso por lo que en mi caso al ordenador le cuesta mucho tiempo ejecutarlo, exactamente 16 horas y 51 minutos. El archivo de salida es un archivo de texto plano (SAM) y suele tener un gran tamaño, en este concreto 33.7GiB. Por ello, lo habitual es trabajar con su versión binaria ordenada, que tiene un menor tamaño:

```
samtools sort mysamplename.sam -o ./Ordenatuta/mysamplename.bam
```

Una vez hecha la alineación de los datos, podemos visualizarlos mediante el programa IGV (explicado en la Sección 1.5).

Para poder visualizar los datos en IGV son necesarios los ficheros que hemos creado anteriormente (BAM) y elegir el genoma contra el que se han alineado los datos, en este caso, el genoma humano hg19. Así se puede ver qué bases difieren del genoma de referencia.

(a) Todas las bases



(b) Bases que varían respecto a la referencia

Figura A.11: Programa IGV

En las subfiguras de la Figura A.11 se pueden apreciar tres secciones:

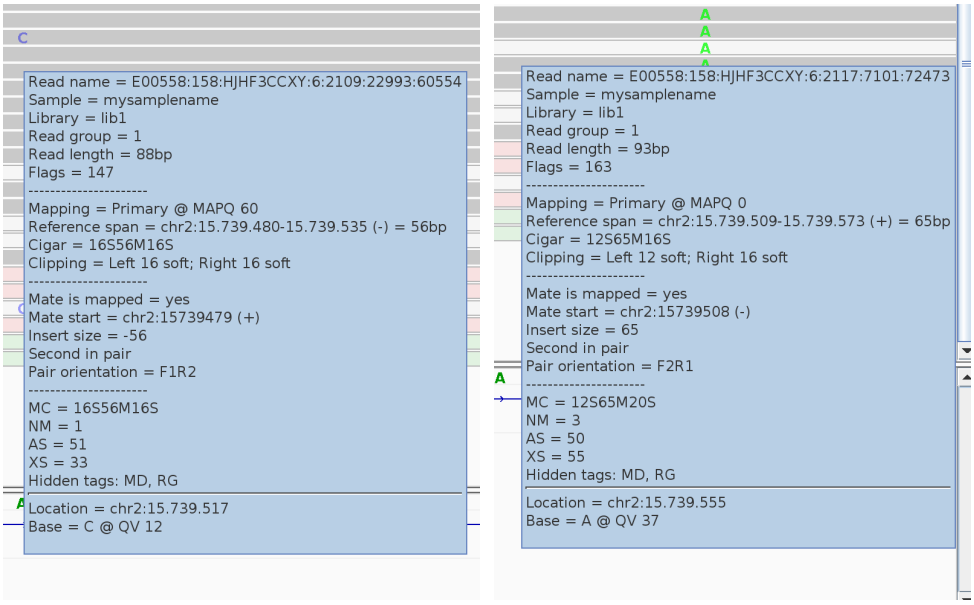
- En la sección de arriba están los cromosomas.
- En el apartado principal se observan las lecturas alineadas a la región indicada en la sección superior. En la Figura A.11a podemos observar todas las bases de cada alineamiento, que difieren de la referencia o no. En cambio en la Figura A.11b, sólo se indican las bases que difieren de la referencia.
- Por último, en la última sección se puede apreciar una secuencia compuesta de diferentes bases (A, G, C o T). Esta secuencia es el genoma de referencia hg19. Es decir, las variantes que se observan en la segunda sección son bases que no concuerdan con esta secuencia.

Tal y como se puede apreciar en la Figura A.11b algunas bases están sombreadas y otras no. Las bases con colores más vivos son de mayor calidad y tienen más probabilidad de ser coincidencias reales. Mientras que las bases con colores más tenues son de menor calidad y tienen menos probabilidad de ser reales.

En la Figura A.12 podemos observar ciertos valores para dos posiciones genómicas distintas. En la subfigura a, la puntuación de calidad es 12, mientras que la base de la subfigura b tiene una puntuación de calidad de 37.

Volviendo a la Figura A.11, se puede observar la diferencia de colores de las lecturas. Esto se debe a que no todas las lecturas tienen la misma calidad de alineamiento. Por ejemplo, la lectura de la Figura A.12a tiene una calidad de alineamiento de 60 (zona grisácea), mientras que la lectura de la Figura A.12b tiene una calidad de alineamiento de 0. IGV indica todas las lecturas que tienen calidad de alineamiento de cero como huecas o blancas.

Existen múltiples anotaciones en el programa IGV. En la Figura A.13 se pueden observar dos anotaciones diferentes. El número 8 morado significa la inserción de 8 bases, en este



(a) Calidad de la sección gris y base sombreada **(b)** Calidad de la sección blanca y base no sombreada

Figura A.12: Programa IGV. Calidad.

caso de las bases GTGTGTGA. Las bases eliminadas en cambio, se indican con un guión y el número de la cantidad de bases eliminadas. En este caso se pueden observar seis bases eliminadas.

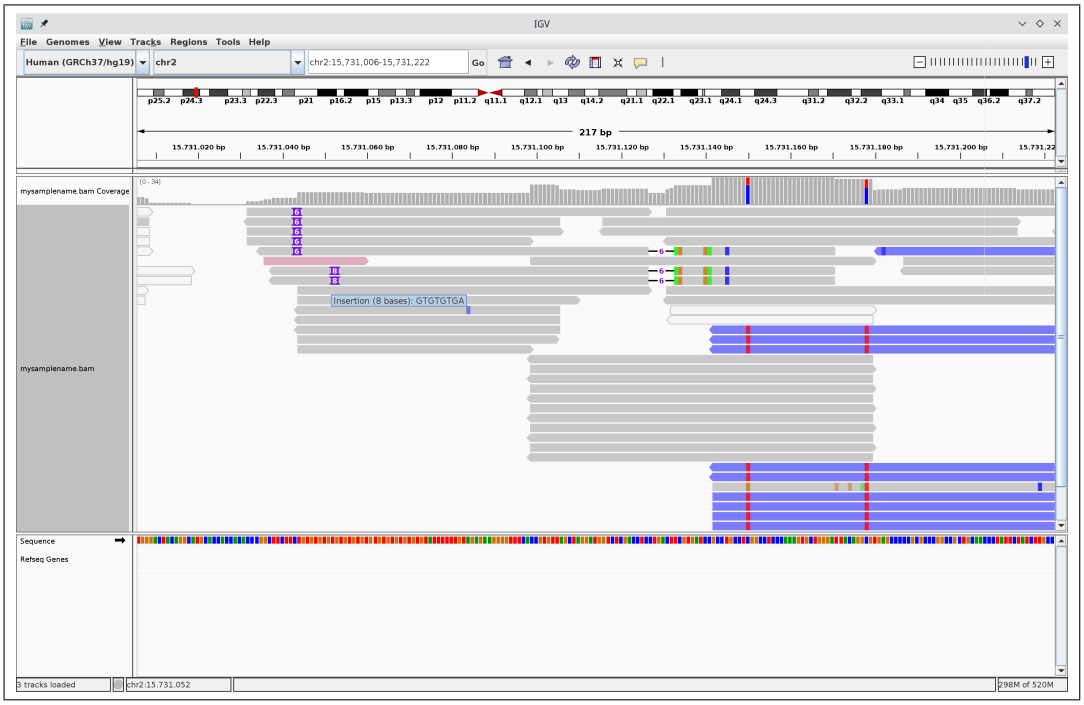


Figura A.13: Programa IGV. Inserción y eliminación de bases.

A.3. Variant calling

La identificación de variantes (Variant calling) es el proceso de identificar y catalogar las diferencias entre las bases de lecturas de secuenciación y el genoma de referencia. Las variantes suelen determinarse a partir de los archivos BAM [13].

En este proyecto, para identificar las mutaciones que tenemos en nuestras muestras hemos utilizado el programa Mutect2 de GATK (explicado en la Sección 1.5). Para ello, primero es necesario completar las anotaciones de los ficheros BAM para que puedan ser usados por Mutect2 con picard.

```
java –jar picard.jar AddOrReplaceReadGroups I="./../Datuak/Garbiak/Ordenatuta/
    ↪ mysamplename.bam" O="./../Datuak/Garbiak/Ordenatuta/mysamplename_rg.
    ↪ bam" RGLB=lib1 RGPL=illumina RGPU=unit1 RGSM=mysamplename
mv ./../Datuak/Garbiak/Ordenatuta/mysamplename_rg.bam ./../Datuak/Garbiak/
    ↪ Ordenatuta/mysamplename.bam
```

También debemos indexar el archivo BAM para que Mutect2 pueda usarlo.

```
samtools index mysamplename.bam
```

Al analizar el ADN de un tumor y sacar la lista de las mutaciones, hay que diferenciar las mutaciones somáticas de las germinales. Para ello, utilizando Mutect2 se ha hecho un variant calling proporcionando al programa una muestra de tejido normal.

```
java –jar ./GATK/gatk–4.2.4.1/gatk–package–4.2.4.1–local.jar Mutect2 –R ./Genoma/
    ↪ hg19.fa –I ./Datuak/Garbiak/Ordenatuta/mysamplename.bam –tumor
    ↪ mysamplename –I ./Datuak/Garbiak/Ordenatuta/normalsample.bam –normal
    ↪ normalsample –O mysamplename.vcf
```

El fichero de salida es un fichero de texto plano VCF. El formato VCF es un formato genérico para almacenar datos de mutaciones de ADN tales como SNP, inserciones, deleciones y variantes estructurales, junto con anotaciones [20].

El fichero se divide en dos partes principales: la parte del comentario que comienza con '#' y la parte principal sin '#'. Cada línea de la parte principal representa la información de una variante:

- **CHROM:** Cromosoma.
- **POS:** La posición genómica donde se encuentra la variante.
- **ID:** ID de la variante. El valor predeterminado es '.'.
- **REF:** Base de la secuencia de referencia.
- **ALT:** Base de la variante.
- **QUAL:** La calidad de la variante. Cuanto mayor sea el valor, mayor será la probabilidad de que sea una variante real.
- **FILTER:** Datos para un filtrado adicional.
- **INFO:** Información sobre la variante.

- **FORMAT:** Información adicional sobre la variante.
- **SAMPLES:** Los valores de los campos indicados en la columna FORMAT.

Para finalizar con el variant calling, se debe consultar una base de datos con mutaciones germinales para poder eliminarlas de nuestra lista de mutaciones tumorales. Para ello, se ha utilizado la base de datos dbSNP y para anotar las variantes, se ha utilizado el programa snpEff. Gracias a este programa se ha modificado la columna ID en caso de que la variante esté en dbSNP y sea, por lo tanto, germinal.

```
dbsnp=/home/laurap/Downloads/common_all_20180423.vcf
java -jar /home/laurap/Downloads/snpEff/SnpSift.jar annotate $dbsnp mysamplename.
    ↪ vcf > mysamplename_annotated.vcf
```

Eliminamos del fichero las mutaciones germinales ya que no son relevantes para el análisis de un tumor.

APÉNDICE B

Resultados de la segunda y tercera paciente

En este apartado se recolectan los resultados obtenidos a raíz de la deconvolución clonal de las tres pacientes.

B.1. Primera paciente

Con los árboles tumorales evolutivos obtenidos con ayuda de un algoritmo heurístico, se ha aplicado el algoritmo implementado en este proyecto (MST). Al aplicar el algoritmo, el peso máximo del grafo de la primera paciente es 411. En la Tabla B.1 se muestran los valores obtenidos a raíz de aplicar el algoritmo creado MST. En dicha tabla se hacen ver las aristas que crean el MST, en este caso, las aristas que unen las mutaciones 164-71 y 164-72 tienen un peso de 10, esto significa que estas relaciones aparecen en los diez árboles de entrada, por lo que dichas relaciones formarían el árbol común. Las filas de color representan las mutaciones comunes entre las tres pacientes.

B.2. Segunda paciente

En la Tabla B.2 se muestran los valores obtenidos de la segunda paciente al conseguir la matriz F y en la Figura B.1 se puede observar que la mayoría de los valores de las frecuencias oscilan entre los valores 0 y 0.1.

El siguiente paso es calcular la distancia entre los árboles. Dichas distancias se pueden observar en la Tabla B.3.

El MST obtenido con los datos de la segunda paciente se muestra en la Tabla B.4, y en la Figura 4.6b el resumen del árbol creado. Tal y como se puede observar en dicha tabla, el árbol común estaría formado por 29 aristas unidas por el mismo nodo, una de dichas aristas sería la 306-128, es decir, la mutación común entre las tres pacientes. El peso total del MST es de 817. Por otro lado, en la Tabla 4.5 se muestra la distancia que hay entre las mutaciones. En dicha tabla se puede observar que las mutaciones más cercanas son las que tienen los indicadores 20 y 80, es decir, entre las mutaciones situadas en los cromosomas 1 y 7.

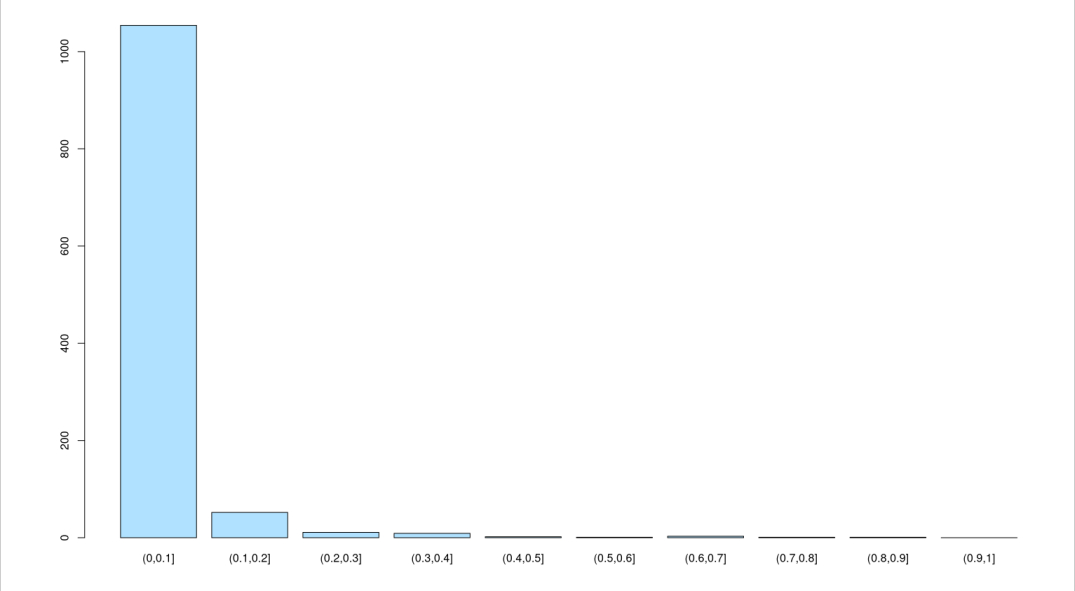


Figura B.1: Cantidad de mutaciones para distintos rangos de frecuancia en la segunda paciente

Aristas		Peso	Aristas		Peso	Aristas		Peso	Aristas		Peso
Desde	A		Desde	A		Desde	A		Desde	A	
2	1	1	51	84	1	71	167	1	19	250	1
11	2	1	185	85	1	25	168	2	1	251	1
11	3	4	6	86	1	9	169	1	21	252	1
6	4	1	121	87	1	122	170	1	142	253	1
32	5	1	332	88	2	24	171	1	83	254	1
58	6	2	3	89	1	37	172	1	75	255	1
20	7	1	108	90	1	306	173	9	306	256	9
3	8	1	52	91	1	29	174	1	252	257	2
58	9	1	306	92	9	202	175	2	306	258	9
53	10	1	25	93	1	306	176	10	306	259	10
125	11	1	6	94	1	26	177	1	45	260	1
11	12	1	306	95	10	279	178	2	29	261	1
51	13	1	6	96	1	270	179	2	8	262	1
45	14	1	39	97	2	47	180	1	244	263	4
1	15	1	102	98	1	88	181	2	306	264	10
41	16	1	34	99	1	35	182	1	33	265	1
21	17	1	7	100	1	82	183	1	306	266	10
26	18	1	46	101	1	15	184	1	53	267	1
8	19	1	99	102	1	306	185	7	306	268	7
31	20	1	306	103	10	66	186	1	16	269	1
29	21	1	14	104	3	32	187	1	51	270	1
46	22	1	50	105	1	23	188	1	306	271	8
10	23	1	34	106	1	20	189	1	306	272	9
6	24	1	74	107	1	306	190	10	51	273	1

(Continúa en la página siguiente)

Tabla B.4: Las aristas con sus respectivos pesos al obtener el MST de la segunda paciente

(Viene de la página anterior)

Aristas		Peso	Aristas		Peso	Aristas		Peso	Aristas		Peso
Desde	A		Desde	A		Desde	A		Desde	A	
3	25	1	43	108	1	80	191	1	1	274	1
10	26	1	45	109	1	109	192	1	68	275	1
5	27	1	60	110	1	51	193	1	1	276	1
46	28	1	9	111	1	76	194	1	306	277	9
45	29	1	306	112	10	12	195	1	43	278	1
43	30	1	7	113	1	306	196	10	39	279	1
2	31	1	306	114	9	52	197	1	14	280	1
11	32	1	25	115	1	35	198	1	306	281	10
19	33	1	306	116	10	75	199	1	306	282	9
1	34	1	39	117	1	306	200	10	32	283	1
40	35	1	36	118	1	306	201	5	42	284	1
30	36	1	1	119	1	306	202	9	75	285	1
306	37	5	32	120	1	306	203	10	306	286	10
35	38	2	47	121	1	46	204	1	57	287	1
8	39	1	15	122	1	82	205	1	6	288	1
24	40	1	35	123	1	306	206	10	28	289	1
24	41	1	12	124	1	24	207	1	20	290	1
57	42	1	306	125	8	15	208	1	51	291	1
27	43	1	121	126	2	32	209	1	306	292	9
86	44	1	99	127	1	8	210	2	306	293	10
16	45	1	306	128	10	306	211	10	78	294	1
2	46	1	19	129	1	16	212	1	83	295	1
31	47	1	39	130	1	306	213	9	53	296	1
45	48	1	24	131	1	306	214	10	47	297	1
104	49	1	14	132	1	121	215	2	31	298	1
81	50	1	94	133	2	75	216	1	7	299	1
54	51	1	102	134	1	306	217	8	20	300	1
11	52	1	306	135	10	3	218	1	33	301	1
24	53	1	9	136	1	55	219	1	12	302	1
31	54	1	27	137	1	94	220	1	99	303	1
32	55	1	57	138	2	35	221	1	3	304	1
306	56	8	102	139	1	306	222	10	306	305	9
16	57	1	82	140	1	45	223	1	75	307	1
8	58	1	306	141	10	51	224	2	29	308	1
14	59	1	22	142	1	16	225	1	181	309	2
15	60	1	21	143	1	172	226	1	45	310	1
34	61	1	306	144	10	14	227	1	33	311	1
5	62	1	35	145	1	306	228	10	45	312	1
5	63	1	296	146	2	15	229	1	306	313	10
255	64	2	78	147	1	102	230	1	94	314	1
20	65	1	39	148	1	45	231	1	84	315	1
58	66	1	70	149	1	52	232	1	306	316	10

(Continúa en la página siguiente)

Tabla B.4: Las aristas con sus respectivos pesos al obtener el MST de la segunda paciente

(Viene de la página anterior)

Aristas		Peso	Aristas		Peso	Aristas		Peso	Aristas		Peso
Desde	A		Desde	A		Desde	A		Desde	A	
38	67	1	306	150	10	46	233	1	306	317	8
4	68	1	75	151	1	71	234	1	94	318	1
5	69	1	306	152	10	46	235	1	72	319	1
20	70	1	43	153	1	35	236	1	47	320	1
18	71	1	6	154	1	51	237	1	104	321	1
29	72	1	306	155	9	35	238	1	268	322	2
80	73	1	99	156	1	25	239	1	306	323	10
94	74	1	144	157	1	289	240	2	24	324	1
27	75	1	120	158	1	113	241	2	306	325	4
108	76	1	306	159	9	108	242	1	306	326	9
25	77	1	34	160	1	306	243	9	119	327	2
34	78	1	75	161	1	306	244	10	75	328	1
43	79	1	121	162	1	15	245	1	1	329	1
2	80	1	32	163	1	51	246	1	82	330	1
168	81	2	31	164	1	94	247	2	138	331	1
4	82	1	95	165	1	121	248	1	5	332	1
63	83	1	84	166	1	53	249	1			

Tabla B.4: Las aristas con sus respectivos pesos al obtener el MST de la segunda paciente

B.3. Tercera paciente

Los valores de la matriz F de la tercera paciente se muestran en la Tabla B.5. En la Figura B.2 se puede observar que la mayoría de los valores de la frecuencia oscilan entre los valores 0 y 0.1 como en las anteriores pacientes.

A continuación se ha calculado la distancia entre los árboles. Dichas distancias se pueden observar en la Tabla B.6.

El MST de la tercera paciente se muestra en la Tabla B.7. El peso total de dicho árbol es de 201. Contiene menos nodos ya que muchas se han eliminado en el filtrado. En la Figura 4.6c se puede observar que cada mutación no se encuentra en una ramificación diferente y que existen varias mutaciones que unen las diferentes variantes comunes a la raíz. La mutación que se encuentra en el cromosoma 19 se sitúa con mayor distancia de la raíz. En la Tabla 4.6 se muestran las distancias entre las mutaciones comunes. Las mutaciones con menor distancia entre ellas las identificadas como 9 y 46, es decir, las que están en los cromosomas 1 y 7.

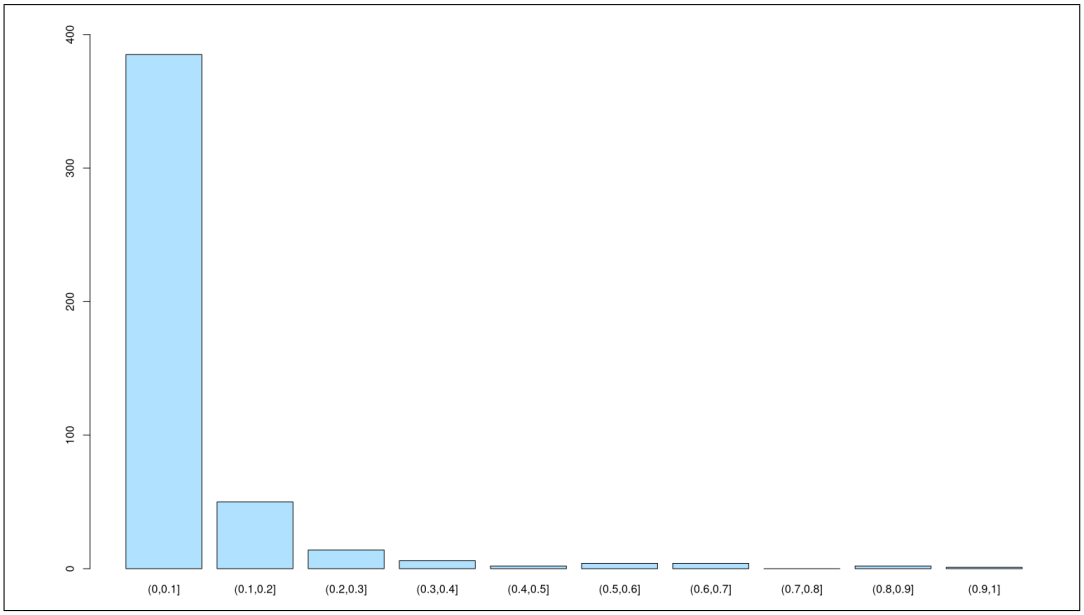


Figura B.2: Cantidad de mutaciones para distintos rangos de frecuancia en la tercera paciente

Aristas		Peso	Aristas		Peso	Aristas		Peso	Aristas		Peso
Desde	A		Desde	A		Desde	A		Desde	A	
6	1	1	73	44	3	164	87	3	164	130	5
168	2	2	105	45	2	164	88	3	151	131	2
28	3	2	88	46	3	3	89	1	164	132	2
45	4	1	45	47	1	14	90	1	39	133	1
164	5	3	127	48	2	21	91	1	62	134	3
10	6	2	36	49	3	7	92	1	143	135	2
164	7	4	8	50	1	39	93	2	164	136	2
58	8	1	136	51	2	118	94	3	131	137	2
116	9	3	25	52	1	80	95	2	2	138	1
164	10	2	59	53	2	23	96	1	43	139	1
164	11	3	58	54	2	14	97	1	10	140	2
143	12	2	144	55	2	16	98	2	122	141	2
164	13	3	22	56	1	116	99	2	66	142	2
15	14	1	3	57	1	164	100	3	87	143	2
10	15	1	172	58	2	164	101	2	164	144	5
44	16	1	171	59	2	11	102	1	164	145	4
2	17	1	164	60	2	123	103	2	2	146	1
157	18	3	26	61	2	113	104	2	18	147	1
164	19	2	9	62	1	6	105	2	128	148	3
7	20	1	74	63	2	1	106	1	101	149	2
164	21	2	63	64	3	5	107	1	19	150	1
147	22	3	30	65	2	21	108	4	21	151	2
45	23	2	164	66	3	21	109	3	171	152	3
35	24	2	21	67	2	55	110	1	168	153	2
3	25	2	11	68	3	28	111	1	164	154	9
164	26	2	95	69	2	29	112	2	6	155	2
11	27	1	164	70	7	11	113	2	31	156	3
27	28	2	164	71	10	98	114	2	59	157	2
164	29	5	164	72	10	128	115	3	44	158	2
7	30	2	169	73	2	63	116	3	21	159	1
152	31	2	164	74	9	48	117	2	164	160	8
164	32	9	101	75	2	48	118	2	141	161	2
2	33	1	164	76	2	164	119	2	16	162	1
53	34	2	164	77	9	116	120	1	39	163	2
36	35	2	34	78	1	90	121	2	158	165	2
5	36	1	164	79	3	7	121	1	29	166	1
5	37	1	59	80	2	155	123	3	142	167	2
5	38	1	19	81	1	164	124	6	164	168	5
138	39	2	164	82	3	119	125	2	164	169	4
164	40	5	53	83	2	112	126	3	164	170	4
164	41	2	19	84	1	13	127	1	164	171	2
9	42	3	7	85	1	18	128	2	164	172	7
21	43	1	29	86	1	100	129	2			

Tabla B.1: Las aristas con sus respectivos pesos al obtener el MST de la primera paciente

	M1	M2	M3	M4	M5	M6	M7	M8	...	M332
1	0	0.003147	0.022	0	0	0.015	0.01	0.013	...	0.177
2	0.029	0.012	0	0.021	0.02	0	0.115	0	...	0.036
3	0	0	0	0	0	0	0	0	...	0.117
4	0.014	0.014	0.016	0	0	0	0.008642	0	...	0.148
5	0.006817	0.034	0.015	0.025	0.029	0.015	0.023	0.039	...	0.038
6	0	0	0	0	0.015	0	0	0.032	...	0.076
7	0	0	0	0.012	0.011	0.015	0	0	...	0.074

Tabla B.2: Parte de la matriz F de la segunda paciente

	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10
A1	0	540	532	530	524	538	540	534	530	538
A2	-	0	542	540	544	546	550	542	538	544
A3	-	-	0	532	532	534	552	534	532	544
A4	-	-	-	0	530	532	542	528	526	544
A5	-	-	-	-	0	542	536	532	528	542
A6	-	-	-	-	-	0	552	534	526	534
A7	-	-	-	-	-	-	0	540	536	546
A8	-	-	-	-	-	-	-	0	528	534
A9	-	-	-	-	-	-	-	-	0	534

Tabla B.3: Distancias entre los árboles de la segunda paciente

	M1	M2	M3	M4	M5	M6	M7	M8	...	M147
1	0	0.004936	0.051	0.044	0.019	0.079	0	0.522	...	0.364
2	0.199	0	0	0.333	0.212	0	0	0	...	0.096
3	0	0.049	0.035	0.032	0.026	0.058	0.023	0.388	...	0
4	0.005802	0.00988	0.016	0	0	0.014	0.01	0.187	...	0
5	0.019	0	0	0	0	0	0.026	0	...	0.041

Tabla B.5: Parte de la matriz F de la tercera paciente

	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10
A1	0	276	276	272	282	280	282	284	286	280
A2	-	0	282	276	274	280	276	276	274	284
A3	-	-	0	286	280	280	272	274	276	278
A4	-	-	-	0	276	284	278	276	280	284
A5	-	-	-	-	0	276	282	280	278	284
A6	-	-	-	-	-	0	284	280	282	282
A7	-	-	-	-	-	-	0	280	282	282
A8	-	-	-	-	-	-	-	0	280	278
A9	-	-	-	-	-	-	-	-	0	274

Tabla B.6: Distancias entre los árboles de la tercera paciente

Aristas		Peso	Aristas		Peso	Aristas		Peso	Aristas		Peso
Desde	A		Desde	A		Desde	A		Desde	A	
79	1	1	7	38	1	147	75	2	3	112	2
37	2	2	27	39	1	5	76	1	15	113	1
67	3	2	46	40	2	5	77	1	19	114	1
14	4	1	6	41	1	16	78	1	61	115	2
25	5	2	5	42	1	7	80	1	27	116	2
16	6	1	40	43	1	1	81	2	26	117	1
3	7	1	1	44	1	8	82	1	15	118	1
6	8	1	13	45	1	16	83	1	87	119	3
28	9	1	7	46	1	15	84	1	5	120	1
1	10	2	1	47	3	129	85	2	103	121	2
59	11	2	140	48	2	3	86	1	84	122	2
23	12	1	11	49	1	5	87	1	43	123	1
14	13	1	94	50	2	31	88	1	5	124	1
120	14	3	143	51	2	2	89	1	8	125	3
3	15	1	23	52	1	62	90	1	35	126	1
2	16	1	24	53	1	7	91	1	7	127	1
16	17	2	5	54	1	130	92	1	8	128	1
81	18	2	10	55	1	1	93	1	7	129	1
22	19	1	23	56	1	101	94	2	7	130	1
28	20	2	72	57	2	53	95	1	13	131	1
11	21	1	22	58	1	112	96	2	19	132	1
26	22	1	1	59	1	8	97	1	8	133	2
24	23	1	11	60	1	139	98	2	26	134	1
70	24	2	1	61	1	59	99	2	41	135	1
44	25	3	11	62	1	85	100	2	129	136	2
23	26	1	3	63	1	44	101	2	24	137	1
96	27	2	15	64	1	13	102	1	33	138	1
129	28	2	28	65	1	8	103	1	5	139	1
77	29	2	26	66	1	27	104	1	10	140	2
138	30	2	2	67	1	19	105	1	41	141	2
23	31	1	8	68	1	53	106	1	33	142	1
25	32	2	53	69	2	124	107	2	2	143	1
12	33	1	7	70	1	19	108	1	14	144	1
5	34	1	74	71	1	67	109	2	87	145	2
7	35	1	107	72	2	13	110	1	16	146	1
3	36	1	35	73	1	33	111	1	57	147	2
139	37	2	32	74	1	3	112	2			

Tabla B.7: Las aristas con sus respectivos pesos al obtener el MST de la tercera paciente

Bibliografía

[1] Santiago Ramón y Cajal, Marta Sesé, Claudia Capdevila, Trond Aasen, Leticia De Mattos-Arruda, Salvador J. Diaz-Cano, Javier Hernández-Losa, and Josep Castellvi. Clinical implications of intratumor heterogeneity: challenges and opportunities. *Molecular Medicine*, 31:1, 2020. Ver página 1.

[2] GEICAM. El cáncer de mama en españa: situación actual, 2021. Ver páginas 1, 2.

[3] Peter C Nowell. The clonal evolution of tumor cel populations. *Science*, pages 23–28, 1976. Ver página 2.

[4] Rebecca A Burrell Nicholas McGranahan Jiri Bartek Charles Swanton. The causes and consequences of genetic heterogeneity in cancer evolution. *Nature*, pages 338–345, 2013. Ver página 3.

[5] Maitena Tellaetxe. An iterated local search algorithm for the clonal deconvolution and evolution problem. *Departamento de Ciencias de la Computación e Inteligencia Artificial*, pages 1–9. Ver páginas 3, 6.

[6] S. Andrews. *A quality control analysis tool for high throughput sequencing data*, 2020. Ver página 4.

[7] A. Bolger. *Trimmomatic: A flexible read trimming tool for Illumina NGS data.* USADELLAB, 2021. Ver página 4.

[8] Heng Li. *bwa - Burrows-Wheeler Alignment Tool.* Sanger Institute, 2013. Ver página 5.

[9] Broadinstitute. *Picard.* Broadinstitute, Boston, Estados Unidos, 2014. Ver página 5.

[10] D. Benjamin T. Sato K. Cibulskis G. Getz C. Stewart and L. Lichtenstein. Calling somatic snvs and indels with mutect2. *The Broad Institute*, page 1, 2019. Ver página 5.

[11] P. Cingolani, A. Platts, M. Coon, T. Nguyen, L. Wang, S.J. Land, X. Lu, and D.M. Ruden. A program for annotating and predicting the effects of single nucleotide polymorphisms, snpeff: Snps in the genome of drosophila melanogaster strain w1118; iso-2; iso-3. *Fly*, 6(2):80–92, 2012. Ver página 5.

[12] Heng Li, Bob Handsaker, Petr Danecek, Shane McCarthy and John Marshall. *bcftools(1) Manual Page.* Genome Research Ltd., 2022. Ver página 5.

[13] Dr. Istvan Albert. *The BIOSTAR HANDBOOK Bioinformatics data analysis guide.* Biostar Genomics, 2016. Ver páginas 5, 16, 33, 39, and 42.

[14] Layla Oesper Kiya Govek, Camden Sikes. A consensus approach to infer tumor evolutionary histories. *International Conference on Bioinformatics, Computational Biology and Health Informatics*, pages 1–10, 2018. Ver página 6.

[15] Arogundade O. T. Sobowale B. and Akinwale A. T. Prim algorithm approach to improving local access network in rural areas. *International Journal of Computer Theory and Engineering*, 3:1, 2011. Ver página 7.

[16] Sheldon Zhai. What is ffpe tissue and what are its uses, 2022. Ver página 16.

[17] Simon Andrews. *FastQC*. Babraham Bioinformatics, Cambridge, Reino Unido, 2010. Ver páginas 16, 34.

[18] Tellaetxe-Abete M., Calvo B., and Lawrie C. Ideafix: a decision tree-based method for the refinement of variants in ffpe dna sequencing data. *NAR genomics and bioinformatics*, 3(4), 2021. Ver página 30.

[19] Qi Y. and Pradhan D. & El kebir M. Implications of non-uniqueness in phylogenetic deconvolution of bulk dna samples of tumors. *Algorithms for Molecular Biology*, 14(1):1–14, 2019. Ver página 30.

[20] Petr Danecek Adam Auton Goncalo Abecasis Cornelis A. Albers Eric Banks Mark A. DePristo Robert E. Handsaker Gerton Lunter Gabor T. Marth Stephen T. Sherry Gilean McVean Richard Durbin. The variant call format and vcftools. *Bioinformatics*, 27:2156–2158, 2011. Ver página 42.