Búsqueda de SNPs y CNVs en *Leishmania donovani*

Nombre estudiante: Esther Camacho Cano
Plan de Estudios del Estudiante: Máster de Bioinformática y Bioestadística
Área del trabajo final: Área ad-hoc

Consultora: Ivette Olivares Castiñeira
Supervisor científico externo: Jose María Requena
Nombre Profesor responsable de la asignatura: Carles Ventura Royo

Fecha Entrega: 24/05/2017

FICHA DEL TRABAJO FINAL

Título del trabajo:	Búsqueda de SNPs y CNVs en Leishmania donovani
Nombre del autor:	Esther Camacho Cano
Nombre de la consultora:	Ivette Olivares Castiñeira
Nombre del PRA:	Carles Ventura Royo
Fecha de entrega (mm/aaaa):	05/2017
Titulación:	Máster de Bioinformática y Bioestadística
Área del Trabajo Final:	Trabajo de Fin de Máster
Idioma del trabajo:	Español
Palabras clave	CNVs, SNPs, Leishmania

Resumen del Trabajo (máximo 250 palabras):

En este TFM se han estudiado cuatro líneas resistentes a fármacos contra la leishmaniasis (anfotericina, miltefosina, paromomicina y antimoniales) de la cepa HU3 de *Leishmania donovani* junto con una línea parental. Esta cepa causa leishmaniasis visceral, la forma más grave de leishmaniasis.

Para evaluar alteraciones a nivel del genoma, se ha llevado a cabo una búsqueda de CNVs a nivel genómico con la que se han estimado los cambios de somía entre líneas resistentes y línea WT. También, se ha realizado una búsqueda de CNVs a nivel de regiones cromosómicas mediante el análisis, por ventanas, de las coberturas de las líneas resistentes frente a la línea parental, identificando 377 CNVs.

Asimismo, se ha realizado una búsqueda de SNPs en el genoma de todas las líneas. Para ello, se ha empleado un programa especializado en búsqueda de SNPs, llamado *VarScan*. La distribución de los SNPs obtenidos en las líneas resistentes muestra zonas del genoma con mayor acumulación de SNPs. Además, se han analizado de forma más detallada los SNPs localizados en ORFs, llegándose a identificar 8 SNPs que implican la creación de tripletes de parada, 1175 que implican cambio amino-acídico y 970 cambios sinónimos.

Finalmente, se ha realizado una búsqueda bibliográfica para plantear hipótesis que asocien las alteraciones observadas en las líneas resistentes con el fenotipo de resistencias. Estas hipótesis, una vez analizadas de forma experimental, podrían ser útiles en el desarrollo de métodos de diagnóstico de resistencia y en el diseño de estrategias para el tratamiento de la leishmaniasis.

Abstract (in English, 250 words or less):

In this TFM, four drug resistant lines, together with the parental line of *Leishmania donovani* HU3 strain, have been studied. The lines were resistant to amphotericin, miltefosine, paromomycin or antimonials. *L. donovani* causes visceral leishmaniasis, which is the most severe form of leishmaniasis.

In order to evaluate genomic alterations, a CNVs search along the genomes was carried out; several somy changes between resistant lines and the parental line have been evidenced. Also, CNVs searches have been done at the chromosomal level by comparing the reads coverage, within discrete chromosomal windows, in the resistant lines regarding the parental one. As a result, 377 CNVs were identified.

In addition, a search for SNPs in the genome of all lines has been performed. For this purpose, a specialized SNP calling program, named *VarScan*, has been used. The distribution of the SNPs obtained in the resistant lines showed zones of the genome with greater accumulation of SNPs. Furthermore, SNPs located within the ORFs have been analyzed in more detail, identifying 8 SNPs involving the creation of stop triplets, 1175 involving amino-acid substitutions and 970 were synonymous changes.

Finally, a bibliographical study was made to find hypotheses that might associate the alterations observed in the resistant lines with the resistance phenotype. These hypotheses, once analyzed experimentally, could be useful in the development of resistance diagnosis methods and in the design of strategies for the treatment of leishmaniasis.

Índice

Lista de figuras

1. Introducción

1.1 Contexto y justificación del Trabajo

Se define como leishmaniasis al conjunto de enfermedades causadas por los protistas del género *Leishmania*. Muchas especies de este género son patógenas en mamíferos, incluido el ser humano. Últimamente, debido a la globalización económica y al incremento de los viajes, su alcance a personas en países desarrollados ha incrementado (1). La leishmaniasis presenta diferentes formas clínicas de la enfermedad (leishmaniasis cutánea, mucosa y visceral). Estas enfermedades infecciosas son endémicas a nivel mundial en regiones tropicales y subtropicales, incluyendo la cuenca Mediterránea (2). Estimaciones recientes indican que hay aproximadamente 12 millones de pacientes y un aumento global de hasta 2 millones de pacientes por año (ver http://www.who.int/leishmaniasis/burden/magnitude/burden_magnitude/en/).

La forma más grave es la leishmaniasis visceral que es mortal, sin el tratamiento adecuado. Las especies de *Leishmania* que causan esta forma son *L. infantum* y *L.donovani* (también referida como *L. chagasi* en sudamérica). Se manifiesta con una gran inflamación en las vísceras (en el bazo y el hígado) (3).

Desafortunadamente, aún no hay una vacuna efectiva contra este parásito (4) y los fármacos disponibles para su tratamiento son limitados en número. El control de estas enfermedades se basa en quimioterapia y en el control del vector (insectos hematófagos de la familia de los flebotominos). Además, el pequeño número de fármacos disponibles para su tratamiento, junto al emergente problema de la aparición de parásitos resistentes (5), hace que la situación sea altamente preocupante. En este contexto, el estudio a nivel genómico de los parásitos resistentes será de gran ayuda para determinar las causas subyacentes de las resistencias. Este conocimiento será de utilidad para detectar casos clínicos que presenten resistencias a los fármacos y también para diseñar nuevos fármacos con mayor eficiencia, lo que sin duda permitirá alcanzar un mejor control de la enfermedad.

Los estudios genómicos en estos parásitos resultan particularmente complicados dada la gran plasticidad del genoma de *Leishmania*. El objetivo principal de este TFM se ha centrado en la búsqueda de SNPs (del inglés *Single Nucleotide Polymorphisms*) y CNVs (*Copy Number Variations*) asociadas a la adquisición de resistencias a fármacos. En particular, la identificación de SNPs constituye una herramienta diagnóstica poderosa para diferenciar rápidamente cepas sensibles y resistentes a fármacos (6).

Por otro lado, las CNVs pueden implicar a cromosomas enteros, o afectar a genes o a regiones cromosomales. En *Leishmania*, con alta frecuencia, se producen amplificaciones de regiones cromosomales, dando lugar a DNA extracromosomales que pueden ser circulares o lineales (7).

Los antimoniales (Glucantime y Pentostam, Sb^V) fueron los primeros fármacos usados para tratar la leishmaniasis hace casi 80 años y continúan siendo la primera opción en muchas partes del mundo. Sin embargo, se han descrito parásitos resistentes en muchas regiones endémicas (8) (9). Más recientemente, se han empezado a utilizar otros fármacos para el tratamiento de la leishmaniasis visceral como por ejemplo la anfotericina B formulada en liposomas (de ahora en adelante referida como anfotericina en el TFM), la miltefosina oral que es el único fármaco que se administra por vía oral (de ahora en adelante referida como miltefosina en el TFM) y una formulación de la aminosidina (de ahora en adelante referida como paromomicina en el TFM) ((8), (10), (11)). No obstante, se han descrito resistencias de parásitos a todos ellos a nivel experimental ((12), (13), (14), (15), (16) y a amfotericina también a nivel clínico (17) y (18).

Para el desarrollo de este trabajo, se han utilizado los datos de secuenciación masiva de DNA de *L. donovani* (cepa HU3) parental (o WT) y 4 líneas resistentes a los fármacos más utilizados en clínica: anfotericina, miltefosina, paromomicina y antimoniales. Tras el desarrollo del TFM se espera encontrar alguna causa que explique las resistencias en las líneas en los SNPs y CNVs identificados en dichas líneas.

1.2 Objetivos del Trabajo

1.2.1. Objetivos generales

A continuación se detallan los objetivos generales del trabajo:

- 1. Identificación de SNPs en las cepas mencionadas de *L. donovani* (WT, resistente a anfotericina, a paromomicina, a antimoniales y a miltefosina) a nivel de todo el genoma.

- 2. Identificación de CNVs en las cepas mencionadas de *L. donovani* (WT, resistente a anfotericina, a paromomicina, a antimoniales y a miltefosina) a nivel de todo el genoma.

- 3. Evaluación de las alteraciones encontradas y búsqueda de la relación con los fenotipos de resistencias a los fármacos mencionados.

1.2.2 Objetivos específicos:

Los objetivos específicos, siguen la misma numeración que los objetivos generales y son los siguientes:

- 1.1. Desarrollar el protocolo o protocolos para obtener una lista de SNPs a partir de datos de secuenciación masiva de DNA genómico de *L. donovani* (cepa WT y líneas resistentes).

- 1.2. Localizar los SNP situados en genes, de acuerdo con la anotación en el genoma de referencia.

- 2.1. Desarrollar el protocolo o protocolos para obtener CNVs a partir de secuencias de ADN genómico de *L. donovani*.

- 2.2. Analizar las CNVs encontradas y determinar las alteraciones genómicas asociadas.

- 3.1. Analizar las alteraciones encontradas (tanto de SNPs como de CNVs) entre las cepas resistentes y evaluar su posible implicación biológica en los fenotipos de resistencia.

1.3 Enfoque y método seguido

Se ha escogido el genoma de *L. infantum* como genoma de referencia para el trabajo en lugar del genoma de referencia para *L. donovani,* ya que el primero está mejor anotado y permite obtener mayores porcentajes de alineamiento de las lecturas de secuenciación. Hay que tener en cuenta que ambas especies están filogenéticamente muy próximas y son prácticamente indistinguibles a nivel genómico.

En cuanto a la búsqueda de SNPs, esta consiste en identificar aquellas posiciones nucleotídicas que son diferentes con respecto al genoma de referencia. Para este fin, se ha utilizado un programa de acceso libre, *VarScan* (19), probando diversos parámetros de búsqueda. Se ha escogido *VarScan* porque permite trabajar con las 5 cepas a la vez, en un archivo conjunto, cosa que facilita el análisis. Con el resultado del programa, se ha realizado una tabla de SNPs con las frecuencias por base en cada cepa.

Además, se ha diseñado un *script* para seleccionar los SNPs que se encuentran en regiones codificantes, para lo que se ha utilizado la anotación de coordenadas disponible en la base de datos *TritrypDB*. La selección de SNPs, podría realizarse de manera manual, pero sería una actividad tediosa y llevaría mucho tiempo. Por este motivo, se opta por realizar un *script* que automatice al

máximo la selección. La función biológica, se ha obtenido mediante la anotación del genoma de *L. infantum*.

Por otro lado, se han evaluado los SNPs comunes en las distintas líneas resistentes mediante diagramas de *Venn*. Los diagramas de *Venn* son muy visuales y sencillos de interpretar, por ello se han usado como recurso complementario a las listas de SNPs obtenidas.

Para la búsqueda de CNVs se han realizado varios gráficos de cobertura mediante un *script* que determina los valores de cobertura en regiones o ventanas a lo largo de los distintos cromosomas. Con esta finalidad, bien se escribirá un *script* o se utilizará un *software* de acceso libre que realice esta tarea de forma adecuada. La decisión final se tomará en función del resultado utilizando diferentes tipos de normalizaciones.

Los gráficos resultantes se han utilizado para comparar la cobertura en la cepa WT frente a la cobertura de cada cepa resistente para cada cromosoma. Habrá, por tanto, 4 series de gráficos de este tipo (una serie para cada cepa resistente) que contendrán 36 gráficos cada uno (uno para cada uno de los 36 cromosomas de cada cepa). Esta estrategia visual, permite una identificación fácil y rápida de las CNVs, motivo por el cual se ha utilizado para este fin, en lugar de una estrategia no visual. De este modo, se ha obtenido una lista de CNVs y se ha establecido si implicaban ganancia o pérdida de material genético.

L. donovani (al igual que otras especies del género *Leishmania*) puede alterar su somía en algunos cromosomas como mecanismo de adaptación al medio, por lo que lo común es que las cepas no sean completamente diploides. Por este motivo, se ha establecido una estimación de la somía de las cepas de estudio mediante la realización de gráficos de barras de la cobertura por cromosomas. Esta metodología permite comparar la somía de la cepa WT con la de las líneas resistentes.

Para realizar el TFM, se utilizará un ordenador disponible en el CBMSO con unas características concretas (procesador 2x2,66 Hexa (X5650), 96GB de RAM, HD 4TB (7200RPM)). No se prevén costes económicos extras asociados al producto.

1.4 Planificación del Trabajo

1.4.1. Tareas

Las tareas que se han realizado para llevar a cabo el trabajo son:

1.1.1. Diseño de la estrategia de búsqueda de SNPs.

1.1.2. Uso de la herramienta de *software* libre *VarScan* para obtener una lista de SNPs en todo en genoma.

1.2.1. Realización de un script para la identificación de los genes donde hay un SNP en las regiones codificantes.

2.1.1. Diseño de la estrategia para encontrar las CNVs entre la cepa WT y las cepas resistentes.

2.1.2. Uso de herramientas de software libre o realización de un script para encontrar las CNVs. De aquí, se obtendrá un gráfico que represente la cobertura por ventanas.

2.2.1. A partir de los gráficos, determinar si las CNVs encontradas implican ganancia o pérdida de genes.

3.1.1. Evaluación de la posible implicación biológica en la adquisición de resistencias, en base a las funciones de las proteínas codificadas por los genes donde hay SNPs.

3.1.2. Evaluación de la implicación biológica en la adquisición de resistencias de acuerdo a las funciones de los genes donde hay CNVs. Esto solo será posible en algunos de los casos.

3.1.3. Extracción de conclusiones sobre posibles implicaciones en las resistencias a fármacos.

1.4.2. Calendario

A continuación se muestra el calendario de planificación del TFM:

TFM- Trabajo de Fin de Máster **Tareas**	**Inicio**	**Entrega**	**Calificación**	**Duración**
PEC1- Plan de Trabajo	**01/03/17**	**15/03/17**	**22/03/17**	**11 días**
Realización del Plan de Trabajo	01/03/17	15/03/17		11 días
PEC2- Desarrollo del Trabajo- Fase 1	**16/03/17**	**05/04/17**	**19/04/17**	**15 días**
1.1.1. Diseño de la estrategia de búsqueda de SNPs	16/03/17	21/03/17		4 días
1.1.2. Uso de herramientas de software libre para obtener una lista de SNPs en todo el genoma	21/03/17	27/03/17		5 días
2.1.1. Diseño de la estrategia para encontrar las CNVs entre la cepa WT y las cepas resistentes	27/03/17	30/03/17		4 días
2.1.2. Uso de herramientas de software libre o realización de un script para encontrar las CNVs	30/03/17	05/04/17		5 días
PEC3- Desarrollo del Trabajo- Fase 2	**06/04/17**	**10/05/17**	**24/05/17**	**25 días**

1.2.1. Realización de un script para obtener la anotación de los genes donde hay un SNP en las regiones codificantes	06/04/17	10/04/17		3 días
1.2.2. Realización de una tabla con los cambios a nivel proteico que implican los SNPs en ORFs	10/04/17	15/04/17		4 días
2.2.1. A partir de los gráficos, determinar si las CNVs encontradas implican ganancia o pérdida de expresión génica.	15/04/17	20/04/17		4 días
2.2.2. Estudio de somías para determinar las CNVs a nivel cromosomal.	20/04/17	22/04/17		2 días
3.1.1. Evaluación de la posible implicación biológica en la adquisición de resistencias de las funciones de los genes donde hay SNPs	22/04/17	27/04/17		5 días
3.1.2. Evaluación de la implicación biológica en la adquisición de resistencias de las funciones de los genes donde hay CNVs. Esto solo será posible en algunos de los casos.	27/04/17	04/05/17		6 días
3.1.3. Extracción de conclusiones sobre posibles implicaciones en las resistencias a fármacos	04/05/17	10/05/17		5 días
Memoria y presentación del Trabajo Final	**11/05/17**	**24/05/17**	**02/06/17**	**10 días**
Realización y escritura de la memoria final del TFM	11/05/17	22/05/17		8 días
Realización de una presentación del TFM	19/05/17	24/05/17		4 días
Preparación de la Defensa	**25/05/17**	**06/06/17**	**-**	**9 días**
Revisión de la memoria y de la presentación del trabajo en función de la evaluación realizada por el director	25/05/17	06/06/17		9 días
Defensa pública: Tribunal TFM	**07/06/17**	**21/06/17**	**05/07/17**	**11 días**
Realización de la Defensa pública del TFM ante el Tribunal del TFM	07/06/17	21/06/17		11 días

Tabla 1. *Planificación del TFM*

Nota. *El calendario tiene en consideración los días festivos (sábados y domingos). Esto no quiere decir que, de modo puntual, no se pueda utilizar este tiempo para acabar alguna tarea. Con tal de llegar al mínimo de 300 horas de dedicación al TFM, cada día contemplado en el calendario tendrá una dedicación mínima al TFM de 4 horas.*

En el siguiente diagrama de *Gantt* se puede ver la planificación de la distribución de las tareas en el tiempo:

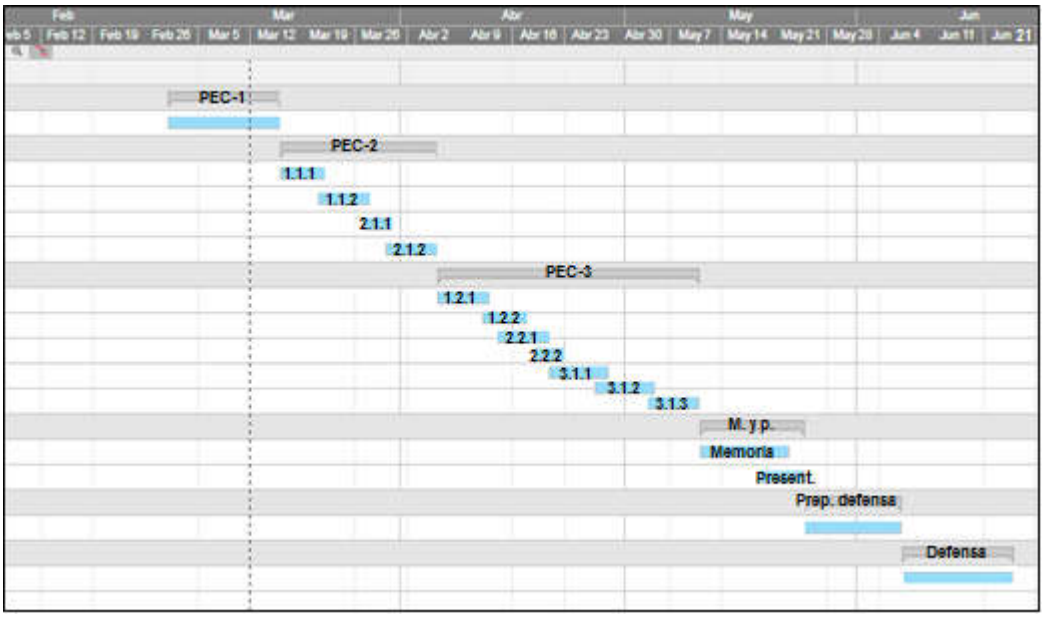

Tabla 2. *Diagrama de Gantt*

NOTA. *El tiempo para realizar la escritura de la memoria tal como viene determinado en el Plan Docente me parece muy reducido si no se ha empezado a escribir nada del trabajo realizado durante la Fase 1 y la Fase 2. Por ello, planeo ir escribiendo la memoria, en la medida que me sea posible, durante el final de la Fase 1 y, sobretodo, la Fase 2, a pesar de que la dedicación máxima sea en el periodo establecido en el Plan Docente de la asignatura.*

1.4.3. Hitos

Los hitos fijados por el Plan Docente en la realización del TFM son los siguientes:

PEC1 - Plan de trabajo	01/03/2017	15/03/2017
PEC2 - Desarrollo del trabajo - Fase 1	16/03/2017	05/04/2017
PEC3 - Desarrollo del trabajo - Fase 2	06/04/2017	10/05/2017
Memoria y presentación del trabajo final	11/05/2017	24/05/2017
Defensa pública - Tribunal TFM	07/06/2017	21/06/2017

Tabla 3. *Hitos del TFM*

Estos hitos, vienen establecidos con fechas de entrega que se han respetado.

PEC1 - Plan de trabajo: Establecer en qué consiste el trabajo de fin de máster, cuáles son los objetivos y el calendario temporal de la realización.

PEC2- Desarrollo del trabajo - Fase 1. Realización de la parte técnica del proyecto: realización de la estrategia de búsqueda de SNPs y CNVs.

PEC3- Desarrollo del trabajo - Fase 2. Determinar, dentro de los SNPs encontrados, cuáles se encuentran en regiones codificantes. Determinar si las CNVs encontradas implican ganancia o pérdida de regiones genómicas o cromosomas comparando las cepas resistentes con la cepa WT. Búsqueda de relaciones biológicas según los SNPs y CNVs encontrados.

Memoria y presentación del trabajo final. Realización y escritura de la memoria final del trabajo a partir de los protocolos de trabajo realizados y los resultados obtenidos. Realización de una presentación resumiendo los pasos seguidos en la realización del trabajo y mostrando los resultados obtenidos.

Defensa Pública- Tribunal del TFM. Defensa del TFM ante un tribunal.

1.4.4. Análisis de riesgos

Los riesgos que podían surgir en el desarrollo del TFM serían:

- Dificultad para establecer un protocolo adecuado (mala elección en el filtro de datos, normalizaciones incorrectas, etc.) para obtener SNPs y CNVs de manera precisa. Los factores limitantes eran el tiempo, la existencia de cierto porcentaje de error experimental en la secuencias generadas (ruido experimental) y las zonas cubiertas con un pequeño número de lecturas (baja cobertura).
- Dificultad para encontrar hipótesis de explicaciones biológicas para los SNPs y CNVs encontrados. Entre los factores limitantes estarían la anotación incompleta del genoma, la complejidad del genoma y la falta de tiempo para realizar posibles pruebas de refuerzo (como por ejemplo, validaciones experimentales), que darían más fuerza a las hipótesis.

1.5 Breve sumario de productos obtenidos

Tras la realización del TFM se han obtenido los siguientes productos:

- Gráficos de somías que estiman el cariotipo de las 5 líneas de estudio.

- Lista de CNVs en las 4 líneas resistentes a fármacos estudiadas comparadas con la línea parental HU3 de *L. donovani*, incluyendo la anotación funcional de los genes que están en regiones de CNVs.

- Lista de SNPs en todo el genoma con frecuencias porcentuales de las bases en las líneas resistentes a fármacos estudiadas y en la línea parental.

- Lista de mutaciones proteicas resultante de los SNPs en ORFs de las cinco líneas de estudio y la anotación funcional de la proteína alterada.

1.6 Breve descripción de los otros capítulos de la memoria

Capítulos del cuerpo de la memoria:

2. Materiales y métodos. En este capítulo se explica el diseño analítico que se ha llevado a cabo y la metodología empleada. Este apartado se divide en los siguientes subcapítulos:

- 2.1. Pasos iniciales del análisis. En este apartado se desarrolla la parte del análisis común en la búsqueda de CNVs y SNPs.
- 2.2. Análisis de CNVs. en este apartado se explica el diseño de la estrategia para obtener las CNVs. Este subcapítulo se divide en dos apartados:
  - 2.2.1. A nivel cromosomal. Este apartado explica el análisis realizado y la estrategia seguida para obtener las CNVs a nivel cromosomal.
  - 2.2.2. A nivel de regiones cromosomales. Este apartado explica el análisis realizado y la estrategia seguida para obtener las CNVs a nivel de regiones cromosomales.
- 2.3. Análisis de SNPs. En este subcapítulo se desarrolla el protocolo seguido para obtener los SNPs. Este subcapítulo se divide en dos apartados:
  - 2.3.1. A nivel de genoma. En este apartado se explica el protocolo seguido para obtener la lista de SNPs en todo el genoma en las líneas de estudio.
  - 2.3.2. A nivel de ORFs. En este apartado se explica la extracción de los SNPs en ORFs obtenidos de las listas de SNPs en el genoma y la obtención de la tabla de implicaciones proteicas de dichos SNPs.

3. Resultados. En este capítulo se exponen los resultados obtenidos y se realiza un análisis relacionado con sus posibles implicaciones biológicas en las líneas resistentes. Este capítulo se divide en dos subcapítulos:

- 3.1. Análisis de CNVs. En este subcapítulo se analizan los resultados del análisis de CNVs en las líneas de estudio. Se divide en dos apartados:
  - 3.1.1. A nivel de cromosomas. Se presentan y analizan los resultados obtenidos del análisis de somías de las líneas de estudio.
  - 3.1.2. A nivel de regiones cromosomales. Se presentan y analizan los resultados obtenidos en el análisis de CNVs a nivel cromosomal mediante los gráficos de cobertura enfrentada (línea resistente-línea parental).
- 3.2. Análisis de SNPs. En este subcapítulo se analizan los resultados del análisis de SNPs en las líneas de estudio. Se divide en dos apartados:
  - 3.2.1. A nivel de genoma. Se presentan y analizan los resultados obtenidos del análisis de SNPs a nivel de genoma.
  - 3.2.2. A nivel de ORFs. Se presentan y analizan los resultados obtenidos en la tabla de mutaciones obtenida a partir del análisis de SNPs en ORFs.

2. Materiales y métodos

En este apartado se explicarán los materiales y métodos utilizados para el análisis.

2.1. Pasos iniciales del análisis

Para la realización de este TFM, se han analizado 4 líneas resistentes a fármacos en la cepa resistente HU3 de *L. donovani* junto con una línea parental (Wt) en la misma cepa. Cada una de las líneas resistentes presenta resistencia a un fármaco, estos fármacos son: anfotericina, miltefosina, paromomicina y antimoniales. Estas líneas han sido secuenciadas en la plataforma Illumina HiSeq 2000. La preparación de genotecas y la secuenciación se realizó en el Centro de Análisis Genómico (CNAG, Spain; http://www.cnag.eu). De las genotecas de DNA se obtuvieron lecturas 2x100.

En primer lugar, se han realizado unos pasos comunes para la estrategia de búsqueda de SNPs y de búsqueda de CNVs a nivel de regiones cromosomales. Estos pasos corresponden a los alineamientos de las lecturas de secuenciación frente al genoma de referencia y la generación de archivos con estos alineamientos.

El genoma de referencia utilizado en este análisis ha sido el de *L. infantum* (cepa JPCM5; (20)), en su versión más reciente (versión 9, disponible en la base de datos *TriTrypDB* (21). La base de datos *TriTrypDB* es específica de kinetoplastidos forma parte del proyecto *EuPathDB*, y la curación de la anotación es realizada por el Instituto Sanger (*GeneDB*). En esta base de datos hay información sobre el genoma de muchas especies de *Leishmania* y archivos descargables con información genómica (genomas, CDCs, tránscritos anotados y archivos en formato *gtf*, entre otros).

En primer lugar, se ha realizado un alineamiento en cada una de las líneas de estudio con el programa *Bowtie2* (version 2.3.1) (22). *Bowtie2* es un alineador que funciona muy bien para alinear lecturas frente a largas secuencias de referencia. Es muy rápido y eficiente memorísticamente. Indexa el genoma con un índice FM (basado en la transformación de *Burrows-Wheeler*) para mantener un rastro de memoria pequeño.

El alineamiento fue llevado a cabo con unos parámetros determinados (*--np 0*, *--n-ceil L,0,0.02*, *--rdg 0,6*, *--rfg 0,6*, *--mp 6,2* and *--score-min L,0,-0.24*).

El parámetro *--np* indica la penalización por tener una indeterminación (N) bien en las lecturas o en la secuencia de referencia. Se decide que no haya

penalización porque el genoma de referencia tiene un número significativo de zonas con Ns.

El parámetro *--n-ceil* establece el límite máximo de número de posiciones que pueden contener caracteres ambiguos de referencia para un alineamiento válido. Este parámetro, viene determinado por una función del tipo *(f(x)=a+bx)*. El primer término, corresponde al tipo de función que en este caso es lineal (L). El segundo y el tercero corresponden al término constante y al coeficiente en dicha función, respectivamente. La variable *x*, corresponde al número de lecturas. En este caso, el término constante lo hemos dejado en 0 (tal como viene por defecto), sin embargo, el coeficiente lo hemos puesto más restrictivo (inferior al que viene por defecto) para que los alineamientos que pasen el filtro tengan menos cantidad de caracteres ambiguos.

Los parámetros *--rdg* y *--rfg* establecen la penalización de la apertura (primer número) y extensión de los *gaps* (segundo número) en la secuencia de la lectura o en la de referencia. En nuestro caso, se ha establecido penalización solo por la extensión de los *gaps*.

El parámetro *--mp* indica la penalización máxima (primer número) y mínima (segundo número) por *mismatch*.

Por último, el parámetro *--score-min* determina una función que depende de la longitud de lectura cuyo resultado establece la puntuación mínima de alineamiento para ser considerado válido. El primer término del parámetro indica el modo *end-to-end*, el segundo y tercer término son el término constante y el coeficiente respectivamente, en una función lineal del tipo *f(x)= a+bx*.

Bowtie2 da el resultado del alineamiento en formato SAM. Dado que el formato SAM genera ficheros de gran tamaño, para su posterior tratamiento se procedió a transformarlos en ficheros BAM. Para este cometido se ha utilizado la herramienta *SAMtools* (versión 1.3.1) (23).

A continuación, se ha realizado un índice Bowtie2 y se ha ordenado el archivo BAM obtenido. Ambos pasos se han realizado utilizando, de nuevo, la herramienta *SAMtools.* Estos archivos se han generado para poder visualizar, posteriormente, los alineamientos en el visualizador IGV (versión 2.3.88) (24).

SAMtools engloba un conjunto de programas para analizar datos de secuenciación masiva. Consiste en tres repositorios: *SAMtools*, *BCFtools* y *HTSlib*. En este análisis usamos el repositorio *SAMtools* que sirve para escribir, leer, editar, indexar y ver formatos SAM, BAM y CRAM.

2.2. Análisis de CNVs

2.2.1. A nivel cromosomal

En primer lugar, se ha realizado una estimación de la somía en los diferentes cromosomas de las líneas de estudio mediante la representación de las coberturas medias de cada cromosoma en gráficos de barras (25). Para realizar esta tarea, se ha realizado el alineamiento explicado anteriormente (*2.1. Pasos iniciales del análisis, p.14*), con unas ligeras modificaciones. El parámetro *--score-min* se ha modificado a *L,0,-0.6* (así el alineamiento será un poco más restrictivo).

A continuación, se eliminan las lecturas que alinean en más de un lugar del genoma (*multi reads*) con *SAMtools view* (parámetros *-Sh -f* 2 *-F* 256) y las lecturas huérfanas (lecturas no asociadas a una lectura pareada) mediante el siguiente *script* escrito en lenguaje Python (26) por el Servicio de bioinformática de la Facultad de Ciencias de la Vida, Universidad de Manchester:

```
#!/usr/bin/env python
import csv
import sys

f = csv.reader(sys.stdin, dialect="excel-tab")
of = csv.writer(sys.stdout, dialect="excel-tab")
last_read = None
for line in f :
    #take care of the header
    if(line[0][0] == "@") :
        of.writerow(line)
        continue

    if(last_read == None) :
        last_read = line
    else :
        if(last_read[0] == line[0]) :
            of.writerow(last_read)
            of.writerow(line)
            last_read = None
        else :
            last_read = line
```

Script 1. *checkpairs.py*

A continuación, se usó *SAMtools view*, de nuevo, para obtener los alineamientos en formato BAM estándar así como para contar el número de lecturas por posición(-F 0x4).

Se han obtenido, para cada cromosoma, las diferentes profundidades de cobertura y el número de bases con dichas profundidades. Para esta tarea se ha utilizado la herramienta *genomeCoverageBed* (de *BEDTools*) y el archivo BAM del paso anterior como archivo de entrada. A continuación, para cada

cobertura por cromosoma, se ha calculado la multiplicación de las profundidades de cobertura por el número de bases con dicha profundidad. Estos resultados se han sumado de manera que se ha obtenido un único número por cromosoma que se ha dividido entre el tamaño del cromosoma (*Tabla 4, p.17*). Estos valores dan la cobertura media estimada por cromosoma en cada línea de estudio.

Cromosoma	***Tamaño en pares de bases***
LinJ.01	277.749
LinJ.02	334.107
LinJ.03	382.164
LinJ.04	474.631
LinJ.05	448.718
LinJ.06	523.352
LinJ.07	591.061
LinJ.08	495.388
LinJ.09	572.109
LinJ.10	546.717
LinJ.11	575.588
LinJ.12	566.969
LinJ.13	644.947
LinJ.14	638.568
LinJ.15	616.811
LinJ.16	697.996
LinJ.17	666.535
LinJ.18	719.782
LinJ.19	742.488
LinJ.20	732.087
LinJ.21	759.492
LinJ.22	658.856
LinJ.23	774.004
LinJ.24	867.022
LinJ.25	886.206
LinJ.26	1.049.056
LinJ.27	1.043.416
LinJ.28	1.163.374
LinJ.29	1.221.192
LinJ.30	1.364.914
LinJ.31	1.468.156
LinJ.32	1.547.509
LinJ.33	1.447.318
LinJ.34	1.666.542
LinJ.35	2.066.803
LinJ.36	2.672.928
Genoma entero	32.101.728

Tabla 4. *Tamaño de los cromosomas*

Nota: *LinJ es la abreviatura para L. infantum usada en la base de datos TriTrypDB. Para referirse a los distintos cromosomas, se usa una ID del tipo LinJ.XX, siendo XX el número del cromosoma correspondiente. A lo largo de este trabajo se utilizará el mismo tipo de ID para referirse a los cromosomas de nuestra cepa de estudio.*

Esta tabla se ha obtenido con el script calculate_contigs_lengths_from_fasta_summary.py, escrito por Ramón Peiró-Pastor del Servicio de Genómica y Secuenciación Masiva del CBMSO en lenguaje Python (Script 2, p.20).

Por último, se aplica una normalización por el número de lecturas totales en las diferentes líneas.

A continuación, se muestra el *script* para obtener el tamaño de los cromosomas:

```
#! /usr/bin/env python
# encoding:UTF-8

# Imports
from __future__ import division
from Bio import SeqIO
import sys

# Parameters
try:
    fasta_file = sys.argv[1]
except IndexError:
    print "\tERROR: One parameters needed !!!!"
    print "\tUsage:", sys.argv[0], "<FASTA_file>"
    print
    sys.exit(1)

print "\tReading FASTA file"
handle = open(fasta_file, "rU")
counter = {}
records_len = []
contigs = {}
for record in SeqIO.parse(handle, "fasta"):
    identity = record.id
    sequence = record.seq
    lgth = len(sequence)
    contigs[identity] = lgth
    records_len.append(lgth)
    sequence = sequence.upper()
    count_a = sequence.count('A')
    count_c = sequence.count('C')
    count_g = sequence.count('G')
    count_t = sequence.count('T')
    count_n = sequence.count('N') + sequence.count('-')
    try:
        counter['A'] = counter['A'] + count_a
    except KeyError:
        counter['A'] = count_a
    try:
        counter['C'] = counter['C'] + count_c
    except KeyError:
        counter['C'] = count_c
    try:
        counter['G'] = counter['G'] + count_g
    except KeyError:
        counter['G'] = count_g
    try:
        counter['T'] = counter['T'] + count_t
```

```
    except KeyError:
        counter['T'] = count_t
    try:
        counter['N'] = counter['N'] + count_n
    except KeyError:
        counter['N'] = count_n
    try:
        counter['bases'] = counter['bases'] + count_a + count_c
+ count_g + count_t
    except KeyError:
        counter['bases'] = count_a + count_c + count_g + count_t
    try:
        counter['num_seq'] = counter['num_seq'] + 1
    except KeyError:
        counter['num_seq'] = 1
handle.close()

print "\tCalculating ...."
sorted_len = sorted(records_len, key = int, reverse = True)
mid_total_lgth = int((counter['bases'] + counter['N']) / 2)
mean_total_lgth = int((counter['bases'] + counter['N']) /
counter['num_seq'])
median_index = int(len(sorted_len) / 2)
n50 = 0
acum_lgth = 0
for lgth in sorted_len:
    acum_lgth = acum_lgth + lgth
    if acum_lgth < mid_total_lgth:
        continue
    else:
        n50 = lgth
        break

print "\tWriting in output file"
output = fasta_file + '.lengths'
handle = open(output, "w")
string = "# Total number of contigs/scaffolds = " +
str(counter['num_seq'])
print string
string = string + "\n"
handle.write(string)
string = "# Total number of bases (ACGT)= " +
str(counter['bases'])
print string
string = string + "\n"
handle.write(string)
string = "# Total number of gaps (N or -) = " +
str(counter['N'])
print string
string = string + "\n"
handle.write(string)
string = "# Total number of A's = " + str(counter['A'])
print string
string = string + "\n"
handle.write(string)
string = "# Total number of C's = " + str(counter['C'])
print string
string = string + "\n"
handle.write(string)
string = "# Total number of G's = " + str(counter['G'])
print string
```

```
string = string + "\n"
handle.write(string)
string = "# Total number of T's = " + str(counter['T'])
print string
string = string + "\n"
handle.write(string)
string = "# Mean length = " + str(mean_total_lgth) + " bp"
print string
string = string + "\n"
handle.write(string)
string = "# Median length = " + str(sorted_len[median_index]) +
" bp"
print string
string = string + "\n"
handle.write(string)
string = "# N50 length = " + str(n50) + " bp"
print string
string = string + "\n"
handle.write(string)
string = "# Shortest length = " + str(sorted_len[-1]) + " bp"
print string
string = string + "\n"
handle.write(string)
string = "# Longest length = " + str(sorted_len[0]) + " bp"
print string
string = string + "\n"
handle.write(string)
```

Script 2 . *calculate_contigs_from_fasta_summary.py*

2.2.2. A nivel de regiones cromosomales

Para conseguir determinar la lista de CNVs en primer lugar se ha realizado el alineamiento y procesado mencionado (*2.1. Pasos iniciales del análisis, p.14*).

Se decidió realizar el análisis de CNVs mediante una estrategia en la que se enfrentaba la cobertura por ventanas de la línea Wt frente a cada una de las líneas resistente. De aquí, se obtuvo un gráfico para cada uno de los cromosomas en las diferentes líneas que presentan resistencias. Este gráfico permitió determinar rápidamente la existencia de ganancia o pérdida de material de la línea resistente respecto a la Wt.

Antes de determinar el protocolo definitivo de trabajo se realizaron pruebas con diversos programas y *scripts* preparados al efecto, testando también diferentes tipos de normalizaciones. Finalmente se optó por usar el programa *CNV-seq* (27), debido a su fácil manejo y a su adecuación a las finalidades del proyecto. El programa está escrito en Perl y es un programa predictor de CNVs a partir de enfrentar las coberturas por ventanas de dos líneas (control y muestra), utiliza como normalización la cobertura del mismo cromosoma analizado. Además de la obtención de los gráficos de cobertura enfrentada anteriormente mencionados, permite obtener una lista de CNVs según parámetros estadísticos que establece. Este método, resulta de gran interés ya que no deja

el criterio de elección de CNVs a la interpretación del analista, disminuyendo la subjetividad en el proceso.

El programa *CNV-seq*, necesita dos archivos de entrada (uno como referencia, en este caso el alineamiento de la línea Wt, y otro de muestra en este caso el alineamiento de la línea resistente). Estos archivos se generaron a partir del archivo BAM procedente del alineamiento anterior. Cada uno de estos archivos de entrada se formó con la tercera y la cuarta columna del archivo BAM. Este archivo se ha generado con la herramienta *samtools view*, opción –F4 (imprimiendo las dos columnas de interés en un nuevo archivo). Con tal de realizar gráficos para cada cromosoma (y no a nivel de genoma), se extrajeron de estos archivos los datos correspondientes a cada uno de los cromosomas en un nuevo archivo, mediante la herramienta *grep*.

A parte de los dos archivos de entrada, *CNV-seq* necesita el número de bases del cromosoma que se está evaluando. Estos datos se pueden obtener fácilmente con método mencionado anteriormente (*Script 2, p.20*).

A continuación, se ejecutó el programa *CNV-seq* por línea de comandos. De este modo, se obtuvieron 4 grupos de archivos (uno para cada línea resistente). Cada serie de archivos generada contiene 36 archivos en formato *.cnv* y 36 archivos en formato *.count*.

A continuación se muestra un ejemplo de cabecera del archivo *.cnv*:

chromosome	start	end	test	ref	position	log2	p.value	cnv	cnv.size	cnv.log2	cnv.p.value
LinJ.01	1	465	3653	5384	233	-0.16172044	0.12679496	0	NA	NA	NA
LinJ.01	233	697	3682	5456	465	-0.1694778	0.11590635	0	NA	NA	NA
LinJ.01	465	929	105	182	697	-0.3956729	0.00317545	0	NA	NA	NA
LinJ.01	697	1161	110	166	929	-0.1958035	0.08403264	0	NA	NA	NA
LinJ.01	929	1393	113	154	1161	-0.04873136	0.36442142	0	NA	NA	NA
LinJ.01	1161	1625	117	156	1393	-0.01716128	0.45133472	0	NA	NA	NA
LinJ.01	1393	1857	116	169	1625	-0.14502222	0.15265661	0	NA	NA	NA
LinJ.01	1625	2089	138	201	1857	-0.14465102	0.15326966	0	NA	NA	NA
LinJ.01	1857	2321	138	203	2089	-0.15893524	0.13087699	0	NA	NA	NA
LinJ.01	2089	2553	139	185	2321	-0.01456417	0.45866533	0	NA	NA	NA
LinJ.01	2321	2785	118	152	2553	0.03259175	0.40796689	0	NA	NA	NA
LinJ.01	2553	3017	85	131	2785	-0.22615585	0.05615387	0	NA	NA	NA
LinJ.01	2785	3249	108	154	3017	-0.11402282	0.2096047	0	NA	NA	NA
LinJ.01	3017	3481	130	163	3249	0.07151588	0.30449196	0	NA	NA	NA
LinJ.01	3249	3713	146	172	3481	0.16143602	0.12358961	0	NA	NA	NA
LinJ.01	3481	3945	148	161	3713	0.27641271	0.02372156	0	NA	NA	NA
LinJ.01	3713	4177	120	146	3945	0.11494226	0.20520432	0	NA	NA	NA
LinJ.01	3945	4409	95	152	4177	-0.28019569	0.0251587	0	NA	NA	NA
LinJ.01	4177	4641	97	170	4409	-0.41160188	0.00230204	0	NA	NA	NA

Tabla 5. *Fragmento del archivo .cnv para el cromosoma 1 en la línia resistente a anfotericina generado por CNV-seq*

Nota. *El archivo .cnv muestra en cada línea información relativa a cada una de las ventanas testadas por el programa. La primera columna hace referencia al cromosoma que se está testando, la segunda a la coordenada de inicio de la ventana, la tercera a la coordenada final. En la cuarta columna está la cobertura media en la muestra problema (en este caso cada una de las líneas resistentes) y en la quinta la de la muestra de referencia (en este caso la WT). En la sexta columna está la posición media entre las coordenadas inicial y final. En la séptima columna está el log2 de la ratio, que indica ganancia (valor positivo) o pérdida (valor negativo) de material genético de la línea resistente respecto a la WT. En la octava columna está el p-valor que contra más pequeño es, más significativa es la diferencia entre la línea*

WT y la resistente. La novena columna indica si hay o no CNVs en esa ventana. En caso de que haya, calcula su tamaño, su log2 ratio y su p-valor en las columnas 10, 11 y 12.

En el archivo *.count*, cada línea tiene información sobre una ventana. Las columnas que tiene corresponden a las 5 primeras columnas de su correspondiente archivo *.cnv*.

Los archivos en formato *.cnv* se han utilizado para obtener los gráficos de cobertura. Abriendo el archivo en R y con la librería *cnv*, se obtuvo un resumen de los estadísticos del archivo (*cnv.summary()*), la lista de CNVs *cnv.print()*, los gráficos (*plot.cnv()*) y se guardó el gráfico de coberturas generado (*ggsave()*).

A continuación se muestra un ejemplo de este tipo de gráficos:

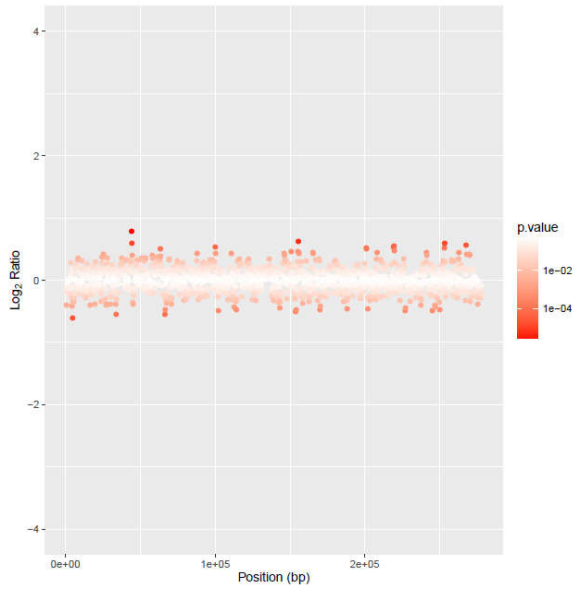

Figura 1. *Gráfico de cobertura por ventanas línea Wt frente línea resistente a anfotericina (cromosoma 1)*

El eje de ordenadas muestra el log2 ratio *(línea resistente/línea WT) y el eje de abscisas muestra la posición en el cromosoma en pares de bases. Un* log2 ratio *de* 1 *implica una ganancia de material donde en la línea resistente hay el doble de material que en la WT, por el contrario, un* log2 ratio *de* -1 *implica una pérdida de material donde en la línea resistente hay la mitad de material que en la WT. A la derecha del gráfico se indica el* p-valor *de cada punto mediante una escala de color. Contra más oscuro es el punto, más significativa es la diferencia.*

2.3. Análisis de SNPs

2.3.1. A nivel de genoma

El primer paso, propiamente dicho, del análisis de SNPs, fue crear un archivo *mpileup* conjunto con los alineamientos de las 5 líneas de estudio como archivos de entrada. Para esta tarea, también se utilizó la herramienta *samtools* con unos parámetros determinados (*-d 50000*, *-A* y *-Q 20*).

El parámetro *–d* indica el número máximo de lecturas que se leen por cada posición en el archivo de entrada. Se debe tener en cuenta que este número se dividirá por el número de archivos de entrada que haya en la orden. En este caso, como queremos obtener un *mpileup* conjunto, un parámetro *-d 50000* sería adecuado.

El parámetro *–A* determina que no se neglijan los pares de lecturas huérfanos en la búsqueda de SNPs.

El parámetro *–Q* indica la mínima calidad de la base para ser considerada.

El formato *mpileup* es un formato tabular que tiene diversas variantes, según el programa que lo genere. El que se ha obtenido en el análisis de este estudio, tiene en una primera columna el cromosoma al que hace referencia la línea, en la segunda columna la coordenada de la base, en la tercera columna la base en la referencia. Las siguientes columnas contienen información para cada línea de estudio. Estos grupos de columnas tienen una columna con el número de lecturas que cubren ese sitio en concreto, otra columna que con las bases leídas y una última con información sobre la calidad de dichas bases.

De este archivo *mpileup* conjunto, se extrajeron los SNPs con *mpileup2snp*, una herramienta del *software VarScan* (versión v2.3) (19) que permite obtener una lista con los SNPs que hay en el archivo *mpileup* conjunto. *VarScan* es un programa codificado en Java que se ejecuta mediante la línea de comandos. Para la búsqueda de variantes, necesita como archivo de entrada un archivo tipo *pileup*. Este archivo es el que se ha generado anteriormente.

Los parámetros que se usaron en este análisis fueron *--min-coverage 12*, *--min-reads2 3* y *--strand-filter 1.*

El parámetro *--min-coverage* establece la cobertura mínima en lecturas que debe tener una determinada posición para poder ser considerada SNP. Se establece en 12.

El parámetro *--min-reads2* indica el número mínimo de lecturas que deben encontrarse en la base alterada para establecer un sitio SNP. En este caso se estableció en 3.

El parámetro *--strand-filter 1* ignora las variantes con más de un 90% de soporte en una hebra (*forward* o *reverse*). Este parámetro, no viene por defecto, y se ha añadido en nuestro análisis para eliminar las variantes con un desajuste muy grande entre lecturas *forward* y *reverse*.

En este punto, se realizó un filtro adicional para descartar los SNPs respaldados con 3 o menos lecturas para cada línea. Así, se obtuvieron 5 archivos distintos, cada uno con los SNPs que hay en cada cepa.

Este tipo de archivo, resultado de *VarScan* contiene diferentes columnas. La primera columna corresponde al cromosoma, la segunda a la posición según el genoma de referencia, la tercera a la base según el genoma de referencia, la cuarta a la base variante para dicha posición. En la quinta columna viene una serie de información sobre las variantes detectadas separadas por dos puntos (':'): la variante consenso en formato IUPAC, la profundidad de cobertura, el número de lecturas que son iguales a la referencia, las que son igual a la variante, la frecuencia de la variante alélica y el p-valor de las lecturas observadas frente a las esperadas según la referencia. La sexta columna contiene la información necesaria para evaluar el sesgo de cadena usando todas las lecturas, separada por dos puntos (':'): las lecturas que apoyan la base de referencia en la cadena *forward*, las lecturas que apoyan la base de referencia en la cadena *reverse*, las lecturas que apoyan la base variante en la cadena *forward* y las lecturas que apoyan la base variante en la cadena *reverse*. La séptima columna corresponde al número de muestras con la misma base que la referencia. La octava columna corresponde al número de muestras marcadas como variante heterocigota. La novena columna corresponde a las marcadas como variantes homocigóticas. La décima columna hace referencia a

las no cubiertas. La undécima columna contiene información de las variantes encontradas por cada muestra introducida en el archivo *mpileup* de entrada. Las diferentes muestras estan separadas por un espacio, cada una de ellas tiene seis parámetros separados por dos puntos (':'). Estos parámetros son: genotipo consenso en formato *IUPAC*, profundidad de lectura total, el número de lecturas que son iguales a la referencia, las que son igual a la variante, la frecuencia de la variante alélica y el p-valor de las lecturas observadas frente a las esperadas según la referencia.

Tal como se acaba de ver, la posibilidad de obtener un archivo conjunto con los SNPs en todas las líneas nos permite obtener información de dicha posición, no solo en las líneas que presentan la variante sino en todas las incluidas en el estudio. Esto es de suma utilidad para el análisis posterior. Por este motivo, se escogió este *software* en lugar de otros *softwares* de obtención de SNPs.

Para obtener una lista de SNPs para cada una de las líneas de estudio (WT, resistente a anfotericina, resistente a paromomicina, resistente a miltefosina y resistente a antimoniales), se seleccionan por cada línea tan solo aquellas variantes cuya frecuencia en la base alterada era mayor (en valor absoluto) a la desviación estándar de la frecuencia de las 5 líneas de estudio (con parámetro sigma 1). Así, finalmente, se obtienen 5 listas de SNPs, una para cada línea de estudio.

Con el programa *bam-readcount* se realizó una tabla con las frecuencias de cada base para cada SNP tanto para la línea que lo contiene como para el resto de líneas. El programa *bam-readcount* es un programa capaz de contar lecturas de secuencia de DNA en archivos BAM generando métricas en posiciones nucleotídicas.

Para poder usar el programa *bam-readcount*, se preparó un archivo de entrada en formato tabular con la pertinente información (cromosoma, coordenada de inicio, coordenada final). Lógicamente, al evaluarse posiciones puntuales, las coordenadas de inicio y final serán las mismas en cada posición y en cada variante. Dicho archivo se crea para cada variante de cada línea de estudio y después se concatena para poder obtener una sola lista al final del proceso.

Además de este archivo, para ejecutar *bam-readcount* se tuvo que especificar el genoma de referencia. Se ha establecido la calidad mínima de la lectura en 15 (parámetro –*b*).

Del archivo generado por *bam-readcount*, se puede obtener el número de lecturas por base y, a partir de aquí, el porcentaje de cada base en cada SNP y para cada una de las líneas de parásitos. Estos tratamientos de los datos se han realizado en *Calc* de *LibreOffice*. Por último, se añaden las cabeceras pertinentes a esta tabla para separar los SNPs de cada línea y, así, se obtiene finalmente la lista de SNPs en el genoma.

Tras obtener esta lista, nos planteamos realizar gráficos de distribución de la mayor diferencia de frecuencia de bases como modo de medir cuan diferente a la línea parental es cada SNP. Se ha generado un gráfico por cada cromosoma y para cada línea resistente.

Así pues se recupera el archivo de frecuencias de bases en los SNPs de todo el genoma y se realiza este cálculo. El cálculo consiste en realizar la diferencia porcentual de la frecuencia de las bases de la línea WT respecto a cada una de las líneas resistentes. De cada SNP se seleccionó aquel de los cuatro valores que era el máximo en valor absoluto. El cálculo se puede resumir en la siguiente fórmula:

$$Diferencia\ porcentual\ máxima = Max(Abs(WT_{BASE} - R_{BASE}))$$

Dónde:

WT_{BASE}= frecuencia porcentual en la base de la línea WT

R_{BASE}= frecuencia porcentual en la base de la línea resistente

Este calculo se realiza para cada una de las bases presentes en esa posición SNP, aunque finalmente solo se asigna a dicho SNP el valor máximo de los cuatro. En la *Tabla 6, p.27*, se se muestra un ejemplo de este cálculo.

Línea	*WT*				*Resistente a anfotericina*				*Diferencia porcentual máxima (WT-Resistente a anfotericina)*				
Base	**A**	**C**	**G**	**T**	**A**	**C**	**G**	**T**	**A**	**C**	**G**	**T**	**Máxima**
SNP	60%	0%	40%	0%	30%	0%	40%	20%	30%	0%	0%	20%	30%

Tabla 6. *Ejemplo de cálculo de la diferencia percentual máxima para un SNP teórico en la línea resistente a anfotericina*

Realizando estos cálculos se obtuvo una diferencia porcentual máxima para cada SNP de las líneas resistentes.

2.3.2. A nivel de ORFs

A partir de la lista de SNPs obtenida con el programa *VarScan* y el archivo en formato *.gtf* con la anotación de los genes en el genoma de referencia de la base de datos TriTrypDB se procedió a generar una lista de los SNPs que se localizaban en ORFs para cada línea de estudio. Para este objetivo se preparó un *script* en Python (*Script 3, p.27*).

En primer lugar, se creó un diccionario con la ID del cromosoma donde se encuentra el SNP como llave, y una tupla con la coordenada donde se encuentra el SNP en dicho cromosoma, la base en la referencia y la nueva base (base alterada).

```
from collections import defaultdict
snp = open("SNPs_donovani_varscan.csv", "r")
rows = (row.strip().split() for row in snp)
row1 = zip(*rows)
casos=zip(row1[1],row1[2],row1[3],row1[4],)
lista_casos=zip(row1[0],casos)
d1 = defaultdict(list)
for k, v in lista_casos:
    d1[k].append(v)
dict_snp = dict((k, tuple(v)) for k, v in d1.iteritems())
```

Script 3. *Diccionario de SNPs*

A continuación se modificó el archivo *.gff* obtenido de la base de datos *TriTrypDB* para seleccionar tan solo las entradas correspondientes a los genes codificantes de proteínas. En un nuevo archivo (‘*Linfantum_mod.gff*’) se seleccionó la primera, la cuarta, la quinta y la novena columna del archivo *.gff* que solo contenía las entradas correspondientes a genes. Estas columnas

correspondían a la ID del gen, a la coordenada de inicio, a la coordenada final y a la anotación funcional de la proteína codificada en el gen. Con esta información, se creó un segundo diccionario, mediante el *Script 4, p.28*.

```
from collections import defaultdict
genes = open("Linfantum_mod.gff ", "r")
rows = (row.strip().split() for row in genes)
row1 = zip(*rows)
casos=zip(row1[1],row1[2],row1[3])
lista_casos=zip(row1[0],casos)
d1 = defaultdict(list)
for k, v in lista_casos:
    d1[k].append(v)

dict_genes = dict((k, tuple(v)) for k, v in d1.iteritems())
```

Script 4. *Diccionario con la información de la anotación funcional de los genes*

Por último, se identificaron los SNPs situados en ORFs mediante el siguiente *Script 5, p.28.*

```
with open('variantes_ORFs_AMPSwt.txt','w') as results:
            for chr in dict_snp.keys():
                  for snp in dict_snp[chr]:
                        pos=int(snp[0])
                        for gen in dict_genes[chr]:
                              ref=""
                              x=int(gen[0])
                              y=int(gen[1])
                              if x<=pos<=y:
                                    ref=chr,snp,gen
                                    results.write(str(ref))
                                    results.write("\n")
                                    break
```

Script 5*. Obtención de los SNPs en ORFs*

Haciendo los cambios pertinentes con las herramientas *grep* y *sed* en la terminal de linux, se obtuvieron las listas de SNPs en ORFs para cada línea de estudio. De este modo, además de la lista de SNPs en ORFs, se anotó la posible función predicha para la proteína codificada en los genes con SNPs (*Tabla 7, p.29*).

LinJ.15	168503	G	A	R:36:27:9:25%:1.1061E-3	164717	174782	ID=LinJ.15.0490;Name=LinJ.15.0490;description=hypothetical+protein+%28pseudogene%29;size=10066
LinJ.15	168504	A	G	R:36:27:9:25%:1.1061E-3	164717	174782	ID=LinJ.15.0490;Name=LinJ.15.0490;description=hypothetical+protein+%28pseudogene%29;size=10066
LinJ.15	168512	A	G	R:41:32:9:21.95%:1.1955E-3	164717	174782	ID=LinJ.15.0490;Name=LinJ.15.0490;description=hypothetical+protein+%28pseudogene%29;size=10066
LinJ.15	168656	A	G	G:87:12:75:86.21%:6.3912E-37	164717	174782	ID=LinJ.15.0490;Name=LinJ.15.0490;description=hypothetical+protein+%28pseudogene%29;size=10066
LinJ.15	176002	G	A	G:169:143:26:15.38%:5.2608E-9	175384	183346	ID=LinJ.15.0500;Name=LinJ.15.0500;description=hypothetical+protein;size=7963
LinJ.15	248683	G	A	R:57:20:37:64.91%:9.5081E-16	246560	249427	ID=LinJ.15.0660;Name=LinJ.15.0660;description=hypothetical+protein%2C+unknown+function;size=2868
LinJ.15	473356	G	A	A:245:43:202:82.45%:3.0013E-95	465421	473517	ID=LinJ.15.1180;Name=LinJ.15.1180;description=protein+kinase%2C+putative;size=8097
LinJ.15	531939	C	T	C:213:175:38:17.84%:5.9398E-13	531779	536146	ID=LinJ.15.1340;Name=LinJ.15.1340;description=Eukaryotic+translation+initiation+factor+4+gamma+type+
LinJ.14	43434	A	G	G:76:0:73:96.05%:1.7006E-43	42485	43567	ID=LinJ.14.0160;Name=LinJ.14.0160;description=hypothetical+protein%2C+conserved;size=1083
LinJ.14	50655	T	G	G:117:21:96:82.05%:2.3711E-45	50020	51534	ID=LinJ.14.0180;Name=LinJ.14.0180;description=carboxypeptidase%2C+putative%2Cmetallo-peptidase%2
LinJ.14	97820	G	A	A:116:3:113:97.41%:7.5819E-64	97621	99756	ID=LinJ.14.0330;Name=LinJ.14.0330;description=hypothetical+protein%2C+unknown+function;size=2136
LinJ.14	120097	C	T	Y:81:52:29:35.8%:8.6928E-11	118383	127118	ID=LinJ.14.0370;Name=LinJ.14.0370;description=hypothetical+protein%2C+conserved;size=8736
LinJ.14	120104	C	T	Y:94:54:40:42.55%:4.4668E-15	118383	127118	ID=LinJ.14.0370;Name=LinJ.14.0370;description=hypothetical+protein%2C+conserved;size=8736
LinJ.14	120126	A	G	R:92:45:47:51.09%:2.3391E-18	118383	127118	ID=LinJ.14.0370;Name=LinJ.14.0370;description=hypothetical+protein%2C+conserved;size=8736
LinJ.14	120146	T	C	Y:76:42:34:44.74%:4.794E-13	118383	127118	ID=LinJ.14.0370;Name=LinJ.14.0370;description=hypothetical+protein%2C+conserved;size=8736
LinJ.14	368817	C	T	Y:199:116:83:41.71%:1.885E-30	368088	370061	ID=LinJ.14.0950;Name=LinJ.14.0950;description=hypothetical+protein%2C+conserved;size=1974
LinJ.14	477307	C	T	Y:28:20:8:28.57%:2.188E-3	477285	487127	ID=LinJ.14.1180;Name=LinJ.14.1180;description=kinesin+K39%2C+putative;size=9843
LinJ.14	477786	A	G	R:251:166:85:33.86%:4.7105E-30	477285	487127	ID=LinJ.14.1180;Name=LinJ.14.1180;description=kinesin+K39%2C+putative;size=9843
LinJ.14	478417	A	T	W:68:27:41:60.29%:6.4299E-17	477285	487127	ID=LinJ.14.1180;Name=LinJ.14.1180;description=kinesin+K39%2C+putative;size=9843
LinJ.14	478485	T	A	W:234:126:108:46.15%:2.6364E-40	477285	487127	ID=LinJ.14.1180;Name=LinJ.14.1180;description=kinesin+K39%2C+putative;size=9843
LinJ.14	478561	C	G	C:512:421:91:17.77%:4.9906E-30	477285	487127	ID=LinJ.14.1180;Name=LinJ.14.1180;description=kinesin+K39%2C+putative;size=9843

Tabla 7. *Fragmento de la lista de SNPs en ORFs con la anotación funcional*

Nota. *En la primera columna de esta tabla aparece el cromosoma donde está el SNP, en la segunda la coordenada donde hay el SNP en el cromosoma, en la tercera la base alterada, en la cuarta columna viene una serie de información sobre las variantes detectadas separadas por dos puntos (':'): la variante consenso en formato IUPAC, la profundidad de cobertura, el número de lecturas que son iguales a la referencia, las que son igual a la variante, la frecuencia de la variante alélica y el p-valor de las lecturas observadas frente a las esperadas según la referencia, en la quinta y la sexta aparecen la coordenada inicial y final del gen, y en la séptima columna aparece la anotación funcional del gen.*

Con ayuda del comando *grep –f* y comparando con las listas de los SNPs de las líneas a nivel de genoma, se obtiene como resultado final una lista de SNPs asociadas a ORFs para cada línea.

Identificación de los SNPs situados en ORFs que conducen a una mutación en la secuencia aminoacídica de la proteína codificada en dicha ORF

Los pasos seguidos para obtener esta información fueron:

A - Obtención de la secuencia de los genes que tienen algún SNP.

B - Mutar estos genes en la posición donde está el SNP.

C - Traducir los genes (anotados y mutados) a proteínas.

D - Comparar las proteínas traducidas de los genes anotados con las proteínas mutadas.

E - Anotar el tipo de mutación (cambio o no de aminoácido) y la anotación funcional del gen para cada SNP.

A. Obtención de la secuencia de los genes que tienen algún SNP

En primer lugar se obtuvo la secuencia de los genes anotados, descargando el archivo *TriTrypDB-9.0_LinfantumJPCM5_AnnotatedTranscripts.fasta* de la base de datos *TriTrypDB*. A continuación, se obtuvo en un archivo los genes que tienen algún SNP, utilizando la herramienta *faidx* de *SAMtools* y dándole como archivos de entrada el archivo *fasta* con las secuencias que se acaban de mencionar y un archivo con una lista de las ID de los genes.

Para realizar esta tarea, se ha utilizado el archivo *AnnotatedTranscripts* en lugar del *AnnotatedCDSs* porque el primero contiene los RNAs y hay algún SNP que afecta a RNAs.

Se han realizado dos modificaciones al archivo *multifasta* obtenido. La primera consistió en que cada secuencia (sin tener en cuenta las cabeceras) estuviese en una sola línea. Para ello se usó el siguiente comando haciendo uso de la herramienta *awk*:

```
awk '/^>/ {printf("\n%s\n",$0);next; } { printf("%s",$0);}
END {printf("\n");}' < archivo_1.fasta >  archivo_2.fasta
```

Donde:

archivo_1.fasta es el archivo de entrada, contiene las secuencias con un número fijo de caracteres por línea.

archivo_2.fasta es el archivo de salida, contiene los carácteres de cada secuencia en una sola línea

Se separó el archivo *multifasta* de salida a *singlefasta* con el siguiente comando escrito en *bash*:

```
while read line
do
    if [[ ${line:0:1} == '>' ]]
    then
        outfile=${line#>}.fa
        echo $line > $outfile
    else
        echo $line >> $outfile
    fi
done < archivo_2.fasta
```

B. Mutación de los genes en la posición donde está el SNP

Para mutar estos genes, interesa saber si están en la hebra positiva o negativa. Para ello, se usó grep para obtener las cabeceras de las secuencias en el archivo de la base de datos *TriTrypDB* que se ha utilizado anteriormente (*TriTrypDB-9.0_LinfantumJPCM5_AnnotatedTranscripts.fasta*). El *output* resultante se volcó en un archivo que se procesó en *Calc* para obtener una columna con el sentido de la hebra para cada gen donde hay algún SNP.

Con esta información y con el archivo generado anteriormente con los SNPs que hay en ORFs, se generó una nueva tabla en formato CSV con la información distribuida por columnas, tal como se muestra en la *Tabla 8, p. 31*(*tabla_provisional.csv*).

Cromosoma	Base alterada	Sentido de hebra	Posición SNP cromosoma	Coordenada inicial	Coordenada final
$1	$2	$3	$4	$5	$6

Tabla 8. *Esquema de las columnas de la tabla y de la distribución por columnas que se usará con la herramienta gawk*

En la segunda fila se indica el nombre de la columna para que sea más fácil de seguir el procedimiento empleado para obtener la posición en el gen.

La instrucción que se utilizó fue:

```
cat tabla_provisional.csv |gawk '{if ($3 == "-") {print
$1,$2,$3, $4, $5, $6,$7=$6-$4}}' > variantes_neg.csv
cat tabla_provisional.csv |gawk '{if ($3 == "+") {print
$1,$2, $3, $4, $5, $6,$7=$4-$5}}' > variantes_pos.csv
```

Se hizo esta distinción entre hebras porque la anotación de las coordenadas inicial y final de los genes no tenía en cuenta la hebra, por tanto, la obtención de la posición del gen cambiaría dependiendo de si la hebra estuviese en sentido positivo o negativo.

Para los genes que estaban en la hebra positiva, este cálculo fue:

Posición del SNP en el cromosoma – Coordenada inicial del gen

Para los genes que estaban en la hebra negativa, este cálculo fue:

Coordenada final del gen - Posición del SNP en el cromosoma

En la lista de la hebra negativa, se tuvo que cambiar la segunda columna (base alterada) a la base complementaria (que era la que nos interesaba obtener).

Estos cambios se realizaron porque las coordenadas de los cromosomas en el genoma de referencia están referidos a la hebra positiva. Además, la base alterada en la hebra negativa también se anotó respecto a la hebra positiva, por ello a la hora de anotar la mutación en el gen, se tuvo que obtener la base complementaria.

El esquema mostrado en la figura 2 (*Figura 2, p. 32*) permite entender mejor el proceso seguido.

Figura 2. *Esquema de la orientación de los genes según el tipo de cadena*

En rojo se enmarca las coordenadas tal como están anotadas en el genoma de referencia. Las cajas verdes corresponderían a genes. Tal como se puede ver en el esquema el gen verde está en la cadena negativa, sin embargo, su anotación es la correspondiente a sus coordenadas en la cadena positiva.

Este procedimiento se realizó mediante la herramienta *tr* en el archivo *variantes_neg.csv* y se obtuvo un nuevo archivo llamado *variantes_neg_modificado.csv*.

A continuación, se concatenaron ambos archivos (*variantes_neg_modificado.csv* y variantes_pos.csv) en un nuevo archivo (*variantes_definitivo.csv*) y se ordenaron (*variantes_definitivo_sorted.csv*).

A partir de este último archivo, se creó un nuevo archivo con tan solo las columnas de interés para los pasos posteriores: ID del gen, posición mutada en el gen y base alterada (*variantes_definitivo_para_dict.csv*).

En la tabla 9 (*Tabla 9, p.32*) se muestra el tipo de información incluido en dicho archivo.

LinJ.01.0020	1165	T
LinJ.01.0020	1124	A
LinJ.01.0170	1066	C
LinJ.01.0170	922	A
LinJ.01.0170	899	A
LinJ.01.0170	602	C
LinJ.01.0180	1295	C
LinJ.01.0180	785	T
LinJ.01.0180	1940	G
LinJ.01.0410	1815	T
LinJ.02.0100	9344	T
LinJ.02.0100	5183	T
LinJ.02.0130	4139	G
LinJ.02.0150	1401	G
LinJ.02.0670	110	A
LinJ.02.0670	1316	A
LinJ.02.0670	1444	T
LinJ.02.0670	1910	C
LinJ.02.0670	2268	T
LinJ.02.0670	2748	A

Tabla 9. *Fragmento del archivo variantes_definitivo_para_dict.csv*

A continuación se preparó un *script* en Python para obtener la secuencia de los genes con SNPs. La primera parte del *script* consistió en la creación de un diccionario con los datos obtenidos en el archivo anterior (Script 6).

```
from collections import defaultdict
SNPs = open("variantes_definitivo_para_dict.csv", "r")
```

```
rows = (row.strip().split() for row in SNPs)
row1 = zip(*rows)
snp=zip(row1[1],row1[2])
lista_snps=zip(row1[0],snp)
d1 = defaultdict(list)
for k, v in lista_snps:
    d1[k].append(v)
dict_genes_snps = dict((k, tuple(v)) for k, v in
d1.iteritems())
```

***Script 6**. Obtención de los genes mutados*

En segundo lugar, se preparó un *script* que va a incluir en la secuencia el cambio definido en el SNP, generando un nuevo archivo con las secuencias mutadas de los correspondientes genes. Para la identificación de los genes mutados se utilizó la siguiente nomenclatura:

'ID del gen' + _pos_+ 'posición mutada en el gen' + _mut.fa

De este modo, se consigue para cada SNP el correspondiente gen mutado. Si hay algún gen con más de un SNP se obtienen tantos genes mutados como SNPs tenga. Además, el nombre de cada uno de estos archivos, al contener la posición mutada, permitirá diferenciarlos.

Para esta tarea se preparó el Script 7.

```
import os
for filename in os.listdir(os.getcwd()):
     fp = open(filename,'r')
     prefix = filename.split('.fa')[0]
     s=list(fp)
     line = s[1]
     n = 1
     seq=[line[i:i+n] for i in range(0, len(line), n)]
     for key in dict_genes_snps:
          if key == prefix:
               for u in dict_genes_snps[key]:
                    with open(prefix +'_pos_' +
u[0]+'_mut.fa','w') as gene_mut:
                         cseq=seq[:]
                         cseq[int(u[0])]=u[1]
                         seq2="".join(cseq)
                         gene_mut.write('>' + prefix +
'_' +str(u[0]))
                         gene_mut.write('\n')
                         gene_mut.write(str(seq2))
          else:
               continue
```

***Script 7.** Obtención de las secuencias mutadas*

***Nota**. La carpeta donde está el script solo debe contener los archivos a mutar (acabados en formato .fa).*

C. Traducción de los genes (anotados y mutados) a proteínas.

Una vez se tienen los genes mutados para todos los SNPs, se realiza un nuevo *script* en Python (Script 8)

```
from Bio.Seq import Seq
import os
for filename in os.listdir(os.getcwd()):
      seq = open(filename,'r')
      prefix = filename.split('.fa')[0]
      protein=open(prefix + '_prot.fa','w')
      seql=list(seq)
      line=seql[1]
      n = 1
      seqnt =[line[i:i+n] for i in range(0, len(line), n)]
      seqdef= seqnt[0:len(seqnt)-1]
      seq2="".join(seqdef)
      coding_dna = Seq(seq2)
      try:
            prot=coding_dna.translate(to_stop=True)
            print >>protein,seql[0], prot
      except BaseException:
            print 'Translation error in', prefix
      except IndexError:
            print 'Index out of range', prefix
```

Script 8. *Traductor de secuencias de DNA a proteina*

Primero, se traducen todos los genes (de referencia y mutados) en secuencia de aminoácidos.

D. Comparación de las proteínas traducidas de los genes anotados con las correspondientes proteínas mutadas

En esta etapa se procedió a comparar las parejas de proteínas, originales y las traducidas de los genes mutadas. Para esta labor, en primer lugar se recuperó el diccionario *dict_genes_snps()* escrito anteriormente en el trabajo.

En segundo lugar, se definió una función, que comparaba las dos secuencias proteicas. Esta función, identificaba el tipo de mutación (cambio aminoacídico, mutación de parada o mutación sinónima). Además, imprimía el resultado en una tabla junto con la ID del gen, la posición del gen donde había la mutación y el cambio de aminoácido que conllevaba en la posición proteíca mutada. Esta tarea se realizó mediante el *script 9*, también escrito en Python.

```
def sequence_compare(seq_a, seq_b):
        len1= len(seq_a)
        len2= len(seq_b)
        for pos in range (0,min(len1,len2)) :
          if seq_a[pos] != seq_b[pos]:
               if seq_b[pos]=='\n':
```

```
                              print
>>f,prefix,',',u[0],',',(str(seq_a[pos])),pos+1,'*,STO
P'
                    else:
                         print
>>f,prefix,',',u[0],',',(str(seq_a[pos])),pos+1,(str(s
eq_b[pos])),',Amino_acid_change'
     if seq_a==seq_b:
          print >>f,prefix,',',u[0],',-,Synonymous'
     else:
          pass
```

Script 9. *Función que compara dos proteínas*

Por último, se escribió un último *script* que aplicó la función anterior a todas nuestras parejas (proteína parental-proteína truncada) que estaban incluidas en el diccionario previamente creado (*dict_genes_snps()*). El resultado se imprimió en un nuevo archivo *.csv* (que se llamó *mutaciones.csv*). El *script*, escrito en *Python*, se muestra a continuación:

```
import os
f = open('mutaciones.csv', 'w')
for filename in os.listdir(os.getcwd()):
     if len(filename.split('_'))<3:
          prefix = filename.split('_')[0]
          for key in dict_genes_snps.keys():
               if key == prefix:
                    for u in dict_genes_snps[key]:
                         try:
                              WT=
open(prefix+'_prot.fa','r')
                              mut=  open(prefix  +'_pos_'
+ u[0]+'_mut_prot.fa','r')
                              sWT=list(WT)
                              smut=list(mut)
                              seq_a=sWT[1]
                              seq_b=smut[1]
                              sequence_compare(seq_a,
seq_b)
                         except IOError:
                              print
>>f,'Error_en_el_gen'+prefix+'.No_hay_codon_de_STOP.IGNORAR
'

f.close()
```

Script 10. *Script para obtener la tabla de mutaciones proteicas*

De nuevo, este proceso se realizó para todas las proteínas truncadas, sin ningún error. Se obtuvo, por tanto, 2153 líneas en la tabla de resultados (una para cada SNP). Este número era el esperado.

En la *Tabla 10, p.36* se muestra un fragmento de la tabla obtenida tras hacer algún cambio de formato con la herramienta *sed* para quitar los guiones bajos al texto en la terminal e insertar una cabecera en la herramienta *Calc*.

Gen ID	Variant nt position	Mutation	Kind of mutation
LinJ.02.0150	1401	M468V	Amino acid change
LinJ.02.0670	110	-	Synonymous
LinJ.02.0670	1316	-	Synonymous
LinJ.02.0670	1444	A482V	Amino acid change
LinJ.02.0670	1910	-	Synonymous
LinJ.02.0670	2268	P757S	Amino acid change
LinJ.02.0670	2748	A917T	Amino acid change
LinJ.02.0670	404	-	Synonymous
LinJ.02.0680	1049	-	Synonymous
LinJ.02.0680	1517	-	Synonymous
LinJ.30.3120	24	-	Synonymous
LinJ.30.3120	992	-	Synonymous
LinJ.30.3120	263	-	Synonymous
LinJ.30.3120	174	E59K	Amino acid change
LinJ.03.0210	3	S2P	Amino acid change
LinJ.30.2840	176	-	Synonymous
LinJ.03.0420	296	-	Synonymous
LinJ.03.0490	4452	T1485A	Amino acid change
LinJ.03.0490	4453	T1485I	Amino acid change
LinJ.22.0750	1498	G500D	Amino acid change
LinJ.30.2570	212	-	Synonymous
LinJ.22.0660	1490	-	Synonymous
LinJ.22.0660	1794	T599A	Amino acid change
LinJ.22.0660	2338	P780R	Amino acid change

Tabla 10. *Fragmento de la tabla de mutaciones proteicas obtenidas a partir de los SNPs en ORFs*

La primera columna hace referencia al ID del gen, la segunda a la posición en el gen donde está el SNP, la tercera a la mutación y la cuarta al tipo de mutación

E. Anotación funcional del gen para cada SNP

A esta tabla, a continuación, se le añadieron tres columnas más, una con la coordenada en el cromosoma afectado, otra con la anotación funcional del gen afectado y otra que con la línea o líneas resistentes afectadas.

Para ello, se recuperó el archivo *.txt* resultante del diagrama de *Venn* en ORFs y se realizó una copia. En dicha copia, se quitó el encabezado, en la primera columna se escribió el símbolo de la línea en mayúscula (A, M, P, S y WT, para resistencia a anfotericina, miltefosina, paromomicina, antimoniales y línea

parental, respectivamente). En los casos en que había más de una línea afectada, estas se han separado por guion. Este archivo se llamó *archivo_venn_orfs.csv*.

En la tabla 11 (*Tabla 11, p.37*) se muestran las columnas añadidas.

LinJ.32	404215	A-M-P
LinJ.34	124262	A-M-P
LinJ.29	833146	A-M-P
LinJ.02	315769	A-M-S
LinJ.31	513936	A-M-S
LinJ.29	981103	A-M-S
LinJ.29	872441	A-M-S
LinJ.11	503554	A-M-WT
LinJ.05	227345	A-P-S
LinJ.29	873113	A-P-S
LinJ.27	22541	A-P-S
LinJ.36	2495095	A-P-S
LinJ.29	760372	A-P-S
LinJ.29	498438	A-P-S
LinJ.29	764136	A-P-S
LinJ.17	511906	A-P-S
LinJ.29	757405	A-P-S
LinJ.29	764419	A-P-S
LinJ.17	630322	A-P-WT
LinJ.31	525774	A-P-WT
LinJ.17	506780	A-P-WT
LinJ.28	650689	A-P-WT
LinJ.33	1243911	A-P-WT

***Tabla 11**. Fragmento del archivo* archivo_venn_orfs.csv

En la primera columna se escribió el cromosoma, en la segunda la coordenada del cromosoma donde se localiza el SNP y en la tercera las líneas afectadas.

En este punto, se recuperó también el archivo *.csv* que contenía las variantes en ORFs en todas las líneas. Se copiaron en un nuevo archivo las columnas que interesaban para obtener esta distribución: Cromosoma, posición en cromosoma, base en referencia, base alterada, ID del gen, anotación funcional (en un archivo tabular). De la descripción, mediante el empleo de la herramienta *sed,* se quitan los caracteres 'ID=' y 'description=' con el objeto de facilitar la lectura en la tabla que se quiere conseguir al final del proceso.

A continuación, se prepararon los datos para poder comparar los dos archivos (*archivo_venn_orfs.csv* y *variantes.csv*). Se pusieron las ID del gen y las coordenadas del cromosoma donde estaba la mutación en una misma columna separadas por ':' (ej. ID:coordenada).

Se aplicó de nuevo, una instrucción de *awk* para combinar ambos archivos:

```
awk 'NR==FNR{a[$1]=$2; next} { { found=0;
for(i=1;i<=NF;i++) { if($i in a) { print $0,a[$i]; found=1;
break; } } if (!found) { print $0,"N/A"} } }'
archivo_venn_orfs.csv variantes.csv>
informacion_para_tabla.csv
```

De este modo, se creó un nuevo archivo, llamado *informacion_para_tabla.csv*. En la *Tabla 12, p.38* se muestra un fragmento de dicho archivo.

LinJ.15:168503	LinJ.15.0490	hypothetical+protein+%28pseudogene%29	WT
LinJ.15:168504	LinJ.15.0490	hypothetical+protein+%28pseudogene%29	WT
LinJ.15:168512	LinJ.15.0490	hypothetical+protein+%28pseudogene%29	WT
LinJ.15:168656	LinJ.15.0490	hypothetical+protein+%28pseudogene%29	M
LinJ.15:176002	LinJ.15.0500	hypothetical+protein	M
LinJ.15:248683	LinJ.15.0660	hypothetical+protein%2C+unknown+function	P
LinJ.15:473356	LinJ.15.1180	protein+kinase%2C+putative	A-M
LinJ.15:531939	LinJ.15.1340	Eukaryotic+translation+initiation+factor+4+gamma+type+2%2C+putative+%28eif4g2%29	A
LinJ.14:43434	LinJ.14.0160	hypothetical+protein%2C+conserved	WT
LinJ.14:50655	LinJ.14.0180	carboxypeptidase%2C+putative%2Cmetallo-peptidase%2C+Clan+MA%28E%29%2C+Family+M32	S
LinJ.14:97820	LinJ.14.0330	hypothetical+protein%2C+unknown+function	A
LinJ.14:120097	LinJ.14.0370	hypothetical+protein%2C+conserved	A
LinJ.14:120104	LinJ.14.0370	hypothetical+protein%2C+conserved	P-WT
LinJ.14:120126	LinJ.14.0370	hypothetical+protein%2C+conserved	P-S
LinJ.14:120146	LinJ.14.0370	hypothetical+protein%2C+conserved	S
LinJ.14:368817	LinJ.14.0950	hypothetical+protein%2C+conserved	P
LinJ.14:477307	LinJ.14.1180	kinesin+K39%2C+putative	A
LinJ.14:477786	LinJ.14.1180	kinesin+K39%2C+putative	S
LinJ.14:478417	LinJ.14.1180	kinesin+K39%2C+putative	A
LinJ.14:478485	LinJ.14.1180	kinesin+K39%2C+putative	M

Tabla 12. *Fragmento del archivo* informacion_para_tabla.csv

En la primera columna se encuentra el cromosoma y la coordenada en el cromosoma, separadas por ':', en la segunda el ID del gen, en la tercera la anotación funcional del gen y en la cuarta las líneas donde está el SNP.

Se realizó otro cambio para poder combinar este archivo con la tabla de mutaciones que consistía en incorporar una columna con la posición en el gen. Para ello, en un nuevo archivo (*posicion_gen.csv*), se recuperó esta información del archivo *variantes_definitivo_sorted.csv* generado anteriormente (*Tabla 13, p.38*).

LinJ.01.0020	469	T	-	6637	6168	7802	1165
LinJ.01.0020	510	A	-	6678	6168	7802	1124
LinJ.01.0170	307	C	-	42063	41756	43129	1066
LinJ.01.0170	451	A	-	42207	41756	43129	922
LinJ.01.0170	474	A	-	42230	41756	43129	899
LinJ.01.0170	771	C	-	42527	41756	43129	602
LinJ.01.0180	1017	C	-	45569	44552	46864	1295
LinJ.01.0180	1527	T	-	46079	44552	46864	785
LinJ.01.0180	372	G	-	44924	44552	46864	1940
LinJ.01.0410	1815	T	+	99015	97200	99665	1815
LinJ.02.0100	5376	T	-	39632	34256	48976	9344
LinJ.02.0100	9537	T	-	43793	34256	48976	5183
LinJ.02.0130	1773	G	-	57899	56126	62038	4139
LinJ.02.0150	812	G	-	76569	75757	77970	1401
LinJ.02.0670	110	A	+	309833	309723	312533	110
LinJ.02.0670	1316	A	+	311039	309723	312533	1316
LinJ.02.0670	1444	T	+	311167	309723	312533	1444
LinJ.02.0670	1910	C	+	311633	309723	312533	1910
LinJ.02.0670	2268	T	+	311991	309723	312533	2268

Tabla 13. *Fragmento del archivo* variantes_definitivo_sorted.csv

En la primera columna aparece el ID del gen, en la segunda el aminoácido mutado, en la tercera la base alterada, en la cuarta el sentido de la hebra, en la quinta la coordenada del cromosoma, en la sexta la coordenada de inicio del gen en el genoma, en la sextat la coordenada final del gen en el genoma, en la séptima la coordenada del gen.

A continuación se combinan este archivo y el archivo *informacion_para_tabla.csv* y se obtiene ya el archivo para poder comparar con la tabla de mutaciones inicial:

```
awk 'NR==FNR{a[$1]=$2; next} { { found=0;
for(i=1;i<=NF;i++) { if($i in a) { print $0,a[$i]; found=1;
break; } } if (!found) { print $0,"N/A"} } }'
posicion_gen.csv informacion_para_tabla.csv >
informacion_tabla_definitiva.csv
```

El resultado obtenido se ilustra en la *Tabla 14, p.39*.

LinJ.15.0490:168503	hypothetical+protein+%28pseudogene%29	WT	3786
LinJ.15.0490:168504	hypothetical+protein+%28pseudogene%29	WT	3787
LinJ.15.0490:168512	hypothetical+protein+%28pseudogene%29	WT	3795
LinJ.15.0490:168656	hypothetical+protein+%28pseudogene%29	M	3939
LinJ.15.0500:176002	hypothetical+protein	M	618
LinJ.15.0660:248683	hypothetical+protein%2C+unknown+function	P	2123
LinJ.15.1180:473356	protein+kinase%2C+putative	A-M	161
LinJ.15.1340:531939	Eukaryotic+translation+initiation+factor+4+gamma+type+2%2C+putative+%28eif4g2%29	A	4207
LinJ.14.0160:43434	hypothetical+protein%2C+conserved	WT	949
LinJ.14.0180:50655	carboxypeptidase%2C+putative%2Cmetallo-peptidase%2C+Clan+MA%28E%29%2C+Family+M32	S	635
LinJ.14.0330:97820	hypothetical+protein%2C+unknown+function	A	199
LinJ.14.0370:120097	hypothetical+protein%2C+conserved	A	1714
LinJ.14.0370:120104	hypothetical+protein%2C+conserved	P-WT	1721
LinJ.14.0370:120126	hypothetical+protein%2C+conserved	P-S	1743
LinJ.14.0370:120146	hypothetical+protein%2C+conserved	S	1763
LinJ.14.0950:368817	hypothetical+protein%2C+conserved	P	1244
LinJ.14.1180:477307	kinesin+K39%2C+putative	A	22
LinJ.14.1180:477786	kinesin+K39%2C+putative	S	501
LinJ.14.1180:478417	kinesin+K39%2C+putative	A	1132
LinJ.14.1180:478485	kinesin+K39%2C+putative	M	1200
LinJ.14.1180:478561	kinesin+K39%2C+putative	S-WT	1276
LinJ.14.1180:478603	kinesin+K39%2C+putative	A	1318
LinJ.14.1180:478607	kinesin+K39%2C+putative	S	1322
LinJ.14.1180:479009	kinesin+K39%2C+putative	S	1724

Tabla 14. *Fragmento del archivo* informacion_tabla_definitiva.csv

En la primera columna aparece el ID del gen y la coordenada en el cromosoma separadas por ':', en la segunda se encuentra la anotación de cada gen en el genoma, en la cuarta las líneas donde está el SNP y en la última la coordenada en el gen.

De este modo, se obtuvo la tabla de interés mediante una última comparación de este último archivo generado y la tabla de mutaciones inicial (después de haber realizado algún procesado en *Calc* en las columnas para que ambos archivos fueran comparables). A continuación se incluye la línea de comandos utilizada:

```
awk 'NR==FNR{a[$1]=$2; next} { { found=0;
for(i=1;i<=NF;i++) { if($i in a) { print $0,a[$i]; found=1;
```

```
break; } } if (!found) { print $0,"N/A"} } }'
tabla_mutaciones_modificada.csv
informacion_tabla_definitiva.csv> tabla_completa.csv
```

Tras realizar algún cambio de formato del archivo obtenido en *Calc* y en la terminal con *sed*, se obtiene una nueva tabla de mutaciones (*Tabla 15, p. 40*).

Lines	Gen ID	Coordinate in chromosome	Position in gene	Mutation	Kind of mutation	Gene function
A	LinJ.02.0670	310127	404	-	Synonymous	hypothetical protein, conserved
A	LinJ.02.0710	324141	1544	-	Synonymous	dipeptylcarboxypeptidase (DCP)
A	LinJ.04.0610	250494	1581	G528S	Amino acid change	hypothetical protein
A	LinJ.07.1080	487232	500	E167D	Amino acid change	hypothetical protein, conserved
A	LinJ.08.0790	334514	15	G6S	Amino acid change	amastin-like protein
A	LinJ.08.0790	334520	21	I8L	Amino acid change	amastin-like protein
A	LinJ.08.0960	410373	1092	M365V	Amino acid change	cathepsin L-like protease
A	LinJ.09.1371	505960	75	D26N	Amino acid change	hypothetical protein
A	LinJ.09.1371	505976	59	-	Synonymous	hypothetical protein
A	LinJ.09.1460	528555	59	-	Synonymous	hypothetical protein, conserved
A	LinJ.09.1470	530277	718	S240F	Amino acid change	hypothetical protein, unknown function
A	LinJ.09.rRNA3	413253	63	R22G	Amino acid change	5S(M5) ribosomal RNA
A	LinJ.10.0170	67835	142	Q48L	Amino acid change	hypothetical protein
A	LinJ.10.0520	216971	886	H296R	Amino acid change	GP63, leishmanolysin,metallo-peptidase, Clan MA(M), Family M8 (GP63-3)
A	LinJ.10.0521	220320	275	-	Synonymous	hypothetical protein, unknown function
A	LinJ.11.1240	513754	1757	-	Synonymous	ATP-binding cassette protein subfamily A, member 4, putative,ABC transporter, putative (ABCA4)

Tabla 15. *Tabla definitiva de mutaciones para cada SNP*

En la primera columna aparece la línea donde hay el SNP, en la segunda la ID del gen, en la tercera la coordenada del cromosoma donde se encuentra, en la cuarta la posición en el gen, en la quinta la mutación, en la sexta el tipo de mutación que supone y en la última la función del gen.

3. Resultados

3.1. Análisis de CNVs

3.1.1. A nivel cromosomal

Primeramente, se realizó un análisis del número de copias de cada uno de los cromosomas, mediante los datos de cobertura de lecturas obtenidas por secuenciación masiva. Esto se hizo para la línea parental (*Figura 3, p.41*), como para las líneas resistentes a anfotericina, miltefosina, paromomicina y antimoniales (*Figura 4, p.42; Figura 5, p.42; Figura 6, p.43; Figura 7, p.43*).Primeramente, se realizó un análisis del número de copias de cada uno de los cromosomas, mediante los datos de cobertura de lecturas obtenidas por secuenciación masiva. Esto se hizo para la línea parental (Fig. 3), como para las líneas resistentes a

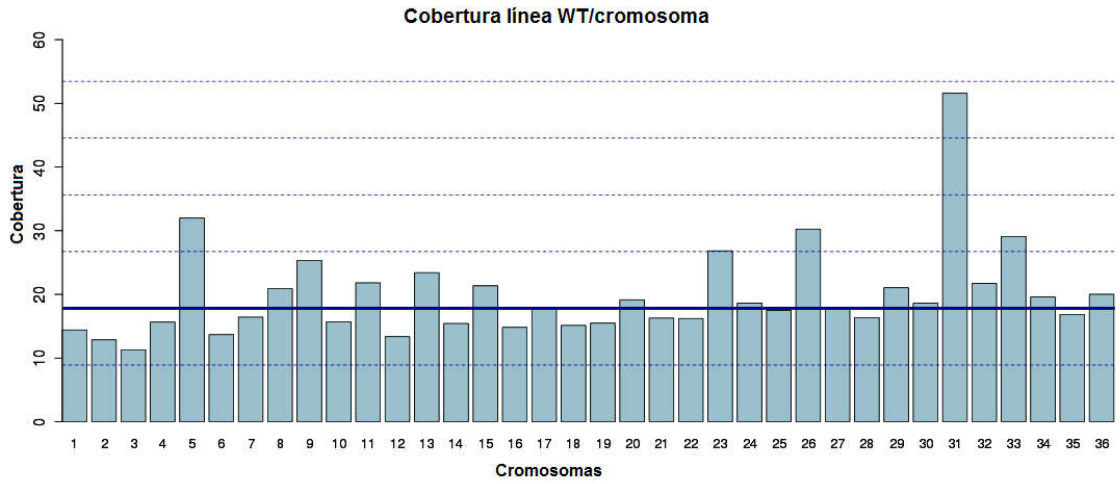

Figura 3. *Gráfico de somías para la línea Wt*

En este gráfico de barras se muestra la somía esperada para cada cromosoma de la línia parental (representados en el eje de ordenadas), según la cobertura media de cada cromosoma (eje de abcisas).

Los datos resultantes del análisis de cobertura media por cromosoma fueron representados en R con la función barplot (). La línea uniforme muestra la mediana de cobertura en cada línea, que corresponde con la disomía teórica esperada. Las líneas discontinuas corresponden con el resto de somías teóricas. Estas líneas han sido dibujadas en el gráfico mediante la función abline() de R.

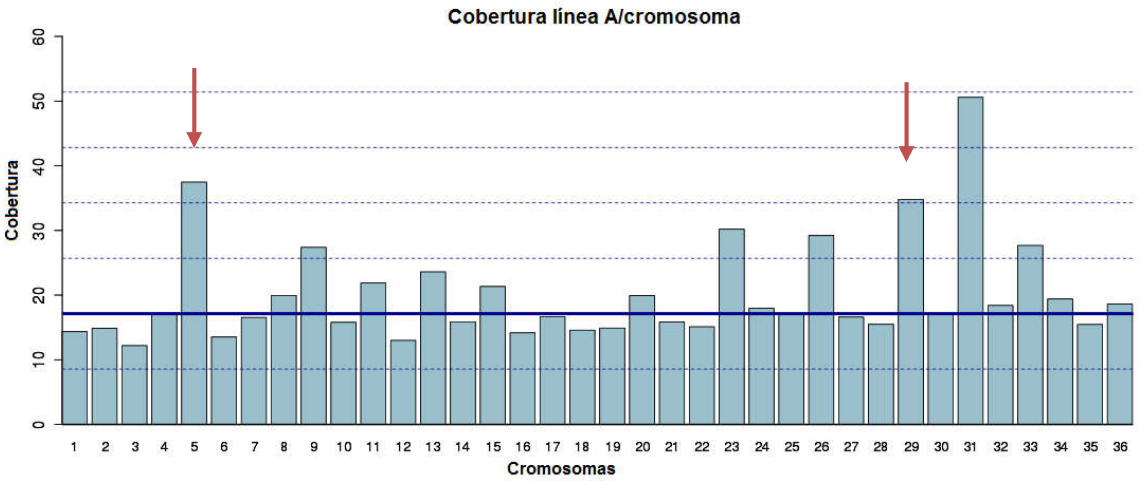

Figura 4. *Gráfico de somía para la línia resistente a anfotericina (A)*

En este gráfico de barras se muestra la somía esperada para cada cromosoma de la línia resistente a anfotericina (representados en el eje de ordenadas), según la cobertura media de cada cromosoma (eje de abcisas). Los cromosomas 5 y 29 presentan cambios de somías respecto a la línea parental (indicados con flechas).

Los datos resultantes del análisis de cobertura media por cromosoma fueron representados en R con la función barplot (). La línea uniforme muestra la mediana de cobertura en cada línea, que corresponde con la disomía teórica esperada. Las líneas discontinuas corresponden con el resto de somías teóricas. Estas líneas han sido dibujadas en el gráfico mediante la función abline() de R.

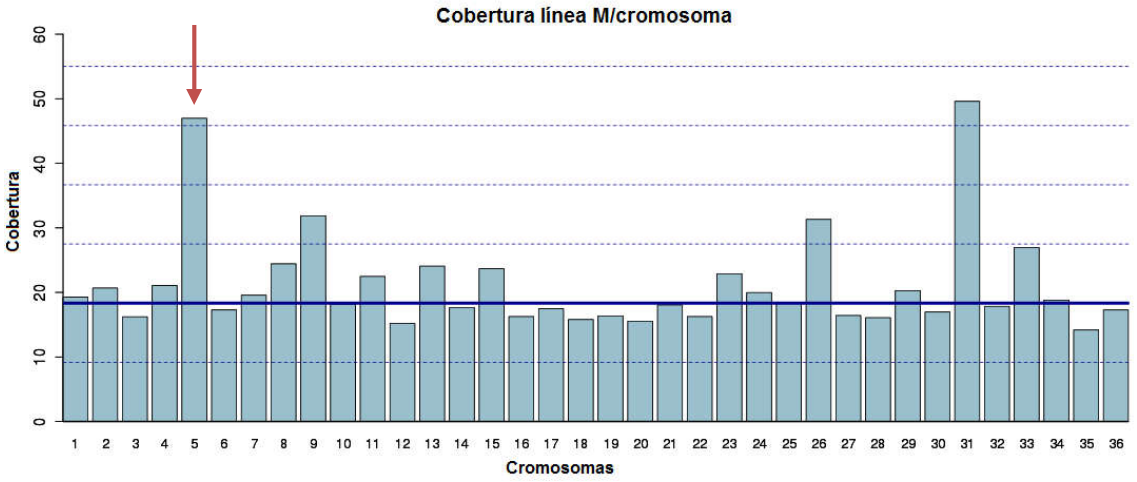

Figura 5. *Gráfico de somía para la línia resistente a miltefosina (M)*

En este gráfico de barras se muestra la somía esperada para cada cromosoma de la línia resistente a miltefosina (representados en el eje de ordenadas), según la cobertura media de cada cromosoma (eje de abcisas). El cromosoma 5 presenta cambios de somías respecto a la línea parental (indicados con flechas).

Los datos resultantes del análisis de cobertura media por cromosoma fueron representados en R con la función barplot (). La línea uniforme muestra la mediana de cobertura en cada línea, que corresponde con la disomía teórica esperada. Las líneas discontinuas corresponden con el resto de somías teóricas. Estas líneas han sido dibujadas en el gráfico mediante la función abline() de R.

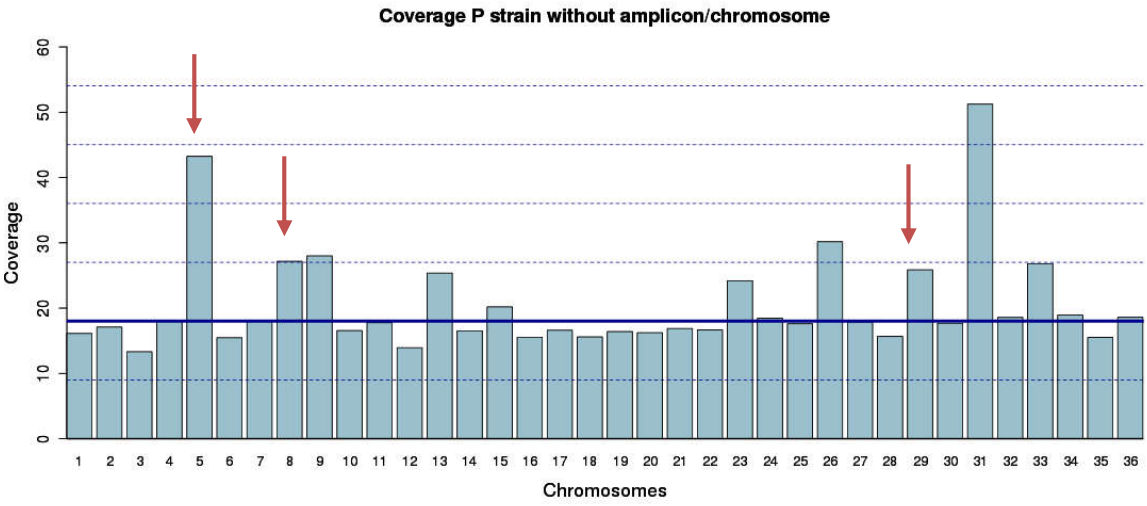

Figura 6. *Gráfico de somía para la línia resistente a paromomicina (P)*

En este gráfico de barras se muestra la somía esperada para cada cromosoma de la línia resistente a paromomicina (representados en el eje de ordenadas), según la cobertura media de cada cromosoma (eje de abcisas). Los cromosomas 5, 8 y 29 presentan cambios de somías respecto a la línea parental (indicados con flechas).

Los datos resultantes del análisis de cobertura media por cromosoma fueron representados en R con la función barplot (). La línea uniforme muestra la mediana de cobertura en cada línea, que corresponde con la disomía teórica esperada. Las líneas discontinuas corresponden con el resto de somías teóricas. Estas líneas han sido dibujadas en el gráfico mediante la función abline() de R.

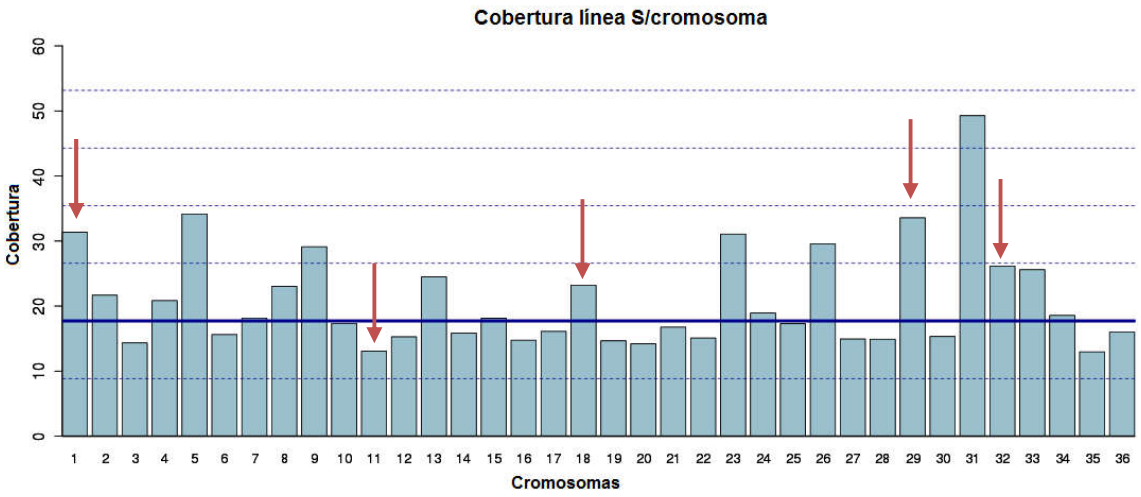

Figura 7. *Gráfico de somía para la línia resistente a antimoniales (S)*

Nota: *Se ha realizado una corrección para el cromosoma 27 en la línea resistente a paromomicina. Esta corrección consiste en la deleción de un amplicón detectado (coordenadas 857,320-865,727) en el cromosoma 27 del genoma de referencia usado en el alineamiento en la línea resistente a paromomicina y, tan solo, para la generación de este gráfico.*

En este gráfico de barras se muestra la somía esperada para cada cromosoma de la línia resistente a antimoniales (representados en el eje de ordenadas), según la cobertura media de cada cromosoma (eje de abcisas). Los cromosomas 1, 11, 18, 29 y 31 presentan cambios de somías respecto a la línea parental (indicados con flechas).

Los datos resultantes del análisis de cobertura media por cromosoma fueron representados en R con la función barplot (). La línea uniforme muestra la mediana de cobertura en cada línea, que corresponde con la disomía teórica esperada. Las líneas discontinuas corresponden con el resto de somías teóricas. Estas líneas han sido dibujadas en el gráfico mediante la función abline() de R.

Para interpretar el significado de un número variable de copias para cada cromosoma debe tenerse en cuenta que en *Leishmania* pueden coexistir subpoblaciones que tengan diferente somía para un mismo cromosoma (28). Estos cromosomas, también llamados cromosomas intermedios, se detectan en los gráficos presentados ya que tienen una somía intermedia. En estos casos, se estima que no habrá una somía única en toda la población para este cromosoma sino que habrá una mezcla de al menos dos somías.

Se muestra una tabla resumen con los cambios de somía de los cromosomas de las líneas resistentes respecto a la línea WT (*Figura 13, p.54*).

Línea resistente	*Cromosomas con diferente somía*
Anfotericina	5 y 29
Miltefosina	5
Paromomicina	5, 8 y 29
Antimoniales	1, 11, 18, 29 y 32

Tabla 16. *Cambios de somía respecto a la línia parental*

Cabe apuntar que la línea Wt no presenta una somía estimada completamente disómica. Los cromosomas 9, 23, 26 y 33 aparecen como trisómicos, el cromosoma 5 como tetrasómico y el cromosoma 31 como hexasómico.

Para estimar el impacto de estos cambios a nivel de implicación biológica en las resistencias, sería necesaria la realización de estudios complementarios.

3.1.2. A nivel de regiones cromosomales

Además de los estudios de somía, se realizó un análisis de distribución de lecturas a lo largo de cada uno de los cromosomas en la línea parental y en las líneas resistentes. Para ello, se analizó la cobertura por ventanas a lo largo de los cromosomas. Los gráficos obtenidos se encuentran recogidos en el Anexo 1 del trabajo.

En la *Tabla 16, p.44* se muestran las CNVs encontradas en las diferentes líneas resistentes.

	Número de CNVs en cada línea resistente			
Cromosoma	*A*	*M*	*P*	*S*
1				2
2				
3		1		1
4		1		1
5		1		2
6				1
7				
8		5	1	5
9		3		2
10		1		1
11	1	1		5
12		3	1	14
13				1

14	1			
15		1		1
16		3		1
17				1
18	1			1
19			1	5
20	2	2		1
21				1
22				
23		1		3
24				1
25				
26		2		2
27		2	16	1
28		1		
29	86		42	106
30	2	4	1	7
31		1	1	
32				
33				
34	1	5		5
35	1	1	2	1
36	1			5

Tabla 17. *CNVs encontradas por cromosoma y línea resistente respecto a la Wt*

Además, se ha generado una hoja de cálculo donde se ha anotado la localización de las CNVs, su tamaño y la variación en cobertura. Una parte de esta hoja se muestra en la *Tabla 18, p. 46.*

La tabla de CNVs tal como se obtiene de CNV-seq tiene un formato determinado.

cnv	chromosome	start	end	size	log2	p.value
CNVR_1	chrLinJ.11	2048	2758	711	-Inf	9.399677e-167
CNVR_1	chrLinJ.14	562425	563395	971	-0.6575195	2.07559e-18
CNVR_1	chrLinJ.18	7199	8173	975	-0.8181175	2.705724e-26
CNVR_1	chrLinJ.20	3070	4557	1488	-0.6977139	5.94E-039
CNVR_2	chrLinJ.20	5488	6231	744	-0.8242903	1.19E-026
CNVR_1	chrLinJ.29	346169	359305	13137	0.9123217	0
CNVR_2	chrLinJ.29	359457	361117	1661	0.7464301	2.48E-059
CNVR_3	chrLinJ.29	361419	362325	907	0.7694821	7.18E-035
CNVR_4	chrLinJ.29	363231	364589	1359	0.7082892	1.01E-044
CNVR_5	chrLinJ.29	364893	367005	2113	0.7600297	2.38E-077
CNVR_6	chrLinJ.29	367761	368969	1209	0.7263137	1.12E-041
CNVR_7	chrLinJ.29	369121	370629	1509	0.7400108	2.56E-053
CNVR_8	chrLinJ.29	370933	371535	603	0.6587272	1.81E-018
CNVR_9	chrLinJ.29	373499	375159	1661	0.7496879	9.03E-060
CNVR_10	chrLinJ.29	376217	376821	605	0.7549964	4.45E-023
CNVR_11	chrLinJ.29	377425	378029	605	0.7628243	1.82E-023
CNVR_12	chrLinJ.29	379841	380595	755	0.6488589	3.97E-022
CNVR_13	chrLinJ.29	381201	381955	755	0.6649278	4.77E-023
CNVR_14	chrLinJ.29	415175	415929	755	0.6368623	1.89E-021
CNVR_15	chrLinJ.29	77	679	603	-0.8859995	3.25E-034
CNVR_16	chrLinJ.29	1737	8985	7249	-1.05099	0
CNVR_17	chrLinJ.29	9137	11401	2265	-1.167232	1.19E-196

Tabla 18. *Fragmento de la lista de CNVs obtenida con CNV-seq para la línea resistente a anfotericina*

La primera columna corresponde al número de CNV por cromosoma. La segunda columna a la ID del cromosoma. La cuarta y la quinta son las coordenadas de inicio y final de la CNV. La sexta corresponde al tamaño de la CNV, la séptima al log 2 ratio *de esta y la última a su p-valor.*

Por otro lado, se procedió a visualizar la localización de estas CNVs en *IGV* y, con la información obtenida de la base de datos *TriTrypDB*, se han anotado los genes que están dentro de ellas con el fin de encontrar posibles relaciones funcionales entre los genes amplificados/delecionados y el tipo de resistencia existente en las líneas estudiadas. Dicha información se ha añadido a la tabla de CNVs y se puede consultar en el Anexo 2 del TFM. Esta nueva información, consiste en tres columnas nuevas, una que indica si la CNV implica ganancia (+) o pérdida de material (-) de DNA para la cepa resistente según *CNV-seq* y tras su visualización en *IGV*, otra que indica las IDs de los genes que están en esas CNVs y, por último, otra que indica su anotación en la base de datos *TriTrypDB*.

De este modo, se obtiene una lista completa de CNVs para cada una de las líneas resistentes a los diferentes fármacos: anfotericina, miltefosina, paromomicina y antimoniales.

A continuación se muestran y comentan ciertos casos peculiares.

- Cromosoma 12 de la línea resistente a antimoniales:

    - Las CNVs 1-9 están muy juntas, así que se han tratado como un bloque a la hora de indicar los genes que hay en ellas.

    - Las CNVs 12-14 están muy juntas, así que se han tratado como un bloque a la hora de indicar los genes que hay en ellas.

- Cromosoma 27 de la línea resistente a paromomicina:

    - Las CNVs 6-16 son falsos positivos. Esto se ha establecido mediante su comprobación en IGV y se ha confirmado con un análisis que se a describe más adelante en el trabajo.

- Cromosoma 29 de las líneas resistentes a anfotericina, paromomicina y antimoniales:

    - Estos cromosomas presentan un mismo perfil peculiar que se analizará en detalle a continuación. En la lista de CNVs se han anotado dos zonas diferentes en dicho cromosoma, pero no se ha anotado los genes que hay en cada una de ellas debido a su gran número.

Análisis de funcionalidad de los genes donde se han detectado CNVs.

A continuación se comentan aquellos genes que podrían estar implicados en el fenotipo de resistencia a los correspondientes fármacos de estudio.

<u>CNVs presentes en la línea resistente a anfotericina.</u>

Gen LinJ.18.0010: Proteína mitocondrial de unión a RNA 1- Pérdida de material (log2 ratio= -0.82)

Las proteínas de unión a RNA son el núcleo de la regulación génica postranscripcional, ya que coordinan el procesamiento, el almacenamiento y la traducibilidad de RNAs celulares.

En *Saccharomyces cerevisiae* se ha visto que los genes implicados en la síntesis de proteínas mitocondriales ribosomales estaban más reprimidos que aquellos del citoplasma, lo cual puede favorecer la citotoxicidad de la anfotericina (29).

La teoría que explicaría la pérdida de material genético para este gen es que esta pérdida de expresión sea, del mismo modo, favorable para nuestra línea de estudio.

Genes LinJ.36.2510 y **LinJ.36.2520: esterol 24-c-metiltransferasa – Pérdida de material (log 2 ratio = -2.06).**

La anfotericina, fue inicialmente diseñada como antifúngico y, posteriormente se vio que también servía como fármaco anti-leishmania. Además, el gen de la esterol 24-c-metiltransferasa es común en hongos y en leishmania y su producto, el ergosterol, se cree que es la diana del fármaco.

Los genes que codifican para esta proteína (LinJ.36.2510 y LinJ.36.2520) son idénticos en cuanto a secuencia y están dispuestos en tándem. Esta región se analizó en el visualizador IGV y se teorizó sobre una posible deleción de una de las dos copias del gen en la línea resistente a anfotericina *(Figura 8, p.48)*, mientras que en la línea parental las dos copias permanecen intactas.

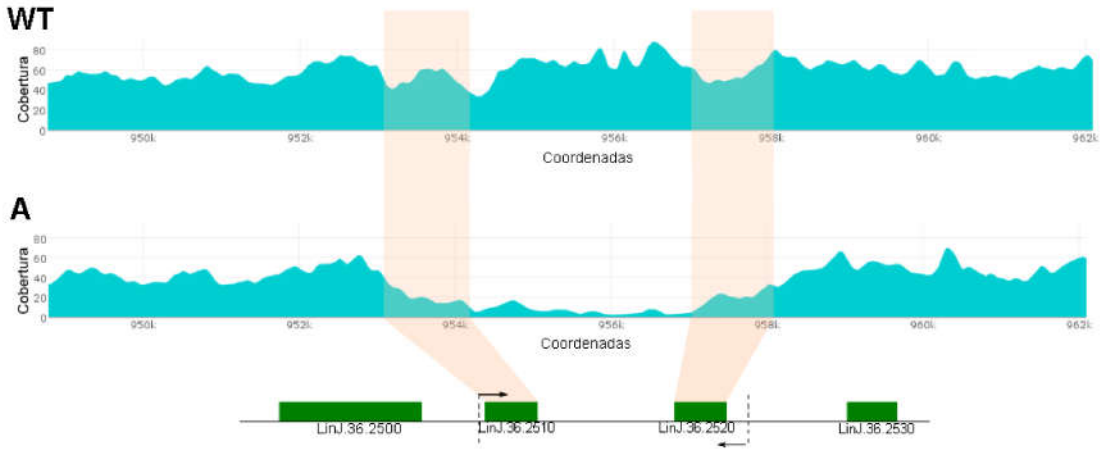

Figura 8. *Gráficos de cobertura de los genes LinJ.36.2510 y LinJ.36.2520 en la línia parental y la línea resistente a la anfotericina*

Nota. *'WT' hace referencia a la línea parental y 'A' a la línea resistente a la anfotericina.*

En la parte inferior del gráfico se representa un esquema de los genes a escala.

Esta hipótesis ha sido corroborada de manera experimental mediante la realización de una PCR. Esta deleción puede haber sido ocasionada por una recombinación homóloga intracromosomal.

La deleción en anfotericina de la esterol 24-c-metiltransferasa implicaría una menor expresión de esta proteína y, por tanto, menos nivel de ergosterol en la membrana celular del parásito, disminuyendo la actuación del fármaco, confiriendo así una mayor resistencia a este en dicha línea.

CNVs presentes en la línea resistente a antimoniales

Gen LinJ.05.1210: Beta tubulina- Pérdida de material genético (log 2 ratio= -0.82)

Se ha encontrado en *L.donovani* expresión alterada de la alfa y la beta tubulina en promastigotes resistentes a arsenito. El arsenito parece alterar las tubulinas porque a medida que se aumenta la dosis del fármaco, disminuye la expresión excepto en la línea resistente del estudio que se mantiene igual (30). Se sabe que los antimoniales también producen cambios en las tubulinas (1). Por eso, se podría hipotetizar que la pérdida de material genético en la línea resistente a antimoniales puede ser ventajosa para la resistencia al fármaco.

Gen LinJ.23.0240: proteína con dominio casete de unión a ATP, subfamília C, miembro 2 (ABCC2)- Ganancia de material (log 2 ratio = 0.75)

En los parásitos, los transportadores con casetes de unión a ATP (ABC) representan una importante familia de proteínas ligadas a resistencia de fármacos, además de desempeñar actividades biológicas relevantes.

El transportador ABCC2 es también conocido como proteína de resistencia a múltiples fármacos 2 (LdMRP2) de *L. donovani*. La proteína se localiza principalmente en la región del bolsillo flagelar y en las vesículas internas (31).

Se ha visto en estudios anteriores que los parásitos resistentes a la baicaleína (BLN) progresivamente adaptados (pB25R) muestran sobreexpresión del transportador ABCC2. Sobreexpresado, confiere resistencia a BLN en los parásitos por el rápido eflujo del fármaco al exterior (31).

En el caso de nuestro estudio, puede estar pasando algo similar en la línea resistente a antimoniales dado que se estima que hay alrededor de 2/3 más de material en la región donde se encuentra este gen, por tanto, también está sobreexpresado en esta línea.

Gen LinJ.36.2740: Sintetasa de folilpoliglutamato – Ganancia de material (log 2 ratio = 0.83)

Los folatos son metabolizados para formar folilpoliglutamato por esta enzima. La poliglutamilación puede desempeñar dos papeles: provocar la retención de folatos en las células y asistir en su compartimentalización en la célula (aunque esta función está menos estudiada). *Leishmania* tan solo posee una sola sintetasa de folilpoliglutamato. Se ha visto que en *L. tarentolae*, las líneas resistentes a metotrexato poliglutamilizan peor, sugiriendo que este gen puede tener un papel importante en la actuación del fármaco (32).

En nuestro caso de estudio, sucede todo lo contrario. Hay más presencia de este gen en la línea resistente a antimoniales, sugiriendo que el aumento de la metabolización del folato puede ser favorable para la resistencia al fármaco.

CNVs presentes en las líneas resistentes a miltefosina y paromomicina

Gen LinJ.27.1940: D-lactato deshidrogenasa- Pérdida de material en miltefosina (log 2 ratio = -0.68) y ganancia de material en paromomicina (log 2 ratio = 4.59).

Gen LinJ.27.1950: Aminotransferasa de aminoácidos de cadena ramificada – Pérdida de material (log 2 ratio = -0.68) y ganancia de material en paromomicina (log 2 ratio = 4.59).

Estos dos genes se encuentran infraexpresados en miltefosina pero lo que realmente llama la atención es lo amplificados que están en la línea resistente a paromomicina.

En la *Figura 9 p. 50* se muestra la amplificación de este locus en la línea resistente a paromomicina.

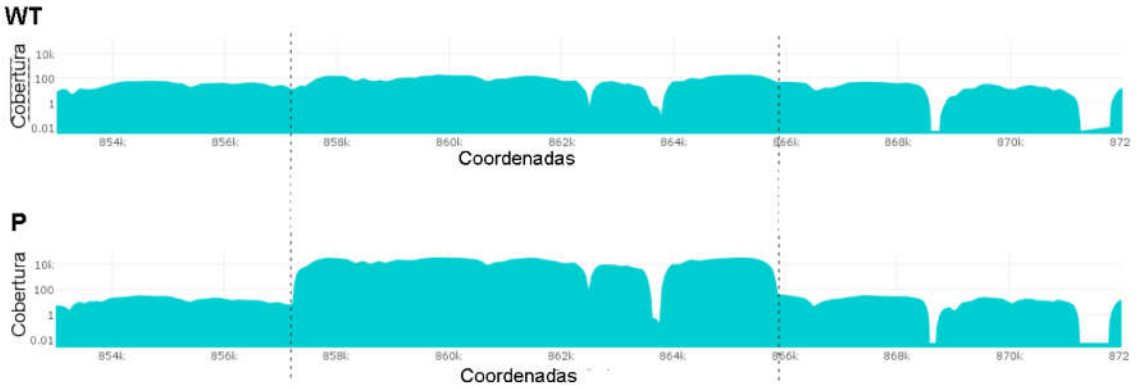

Figura 9. *Gráficos de cobertura de la región con el amplicón detectado en la línea resistente a la paromomicina*

Nota. *'WT' hace referencia a la línea parental y 'P' a la línea resistente a la paromomicina. Fijarse en que ambos gráficos están representados en escala logarítmica.*

La teoría a la que se ha llegado es que esta región se encuentra amplificada en forma de amplicón extracromosomal en la línea resistente a paromomicina, confiriéndole a la línea mayor resistencia a la paromomicina.

CNVs presentes en las líneas resistentes a paromomicina y antimoniales

Gen LinJ.05.1210: proteína tipo antígeno de superfície – Ganancia de material en paromomicina (log 2 ratio = 0.76) y pérdida de material en antimoniales (log 2 ratio= -0.72).

Esta proteína es un componente integral de membrana. Hay estudios que demuestran que el aumento de esta proteína en L. donovani aumenta la resistencia a antimoniales (Bhandari et al., 2013). En nuestro caso, cabría esperar lo contrario para la línea resistente a antimoniales. En la línea resistente a paromomicina, sí que se observa una ganancia de material genético compatible con el estudio anterior que implicaría una ventaja en la línea resistente a paromomicina.

LinJ.30.3610: Proteína tipo citocromo p450 – Pérdida de material en la línea resistente a paromomicina (log 2 ratio = -0.733) y en la línea resistente a antimoniales (log 2 ratio = -0.76).

Los citocromos usan gran variedad de moléculas como sustrato. Los de tipo p450 son proteínas de la subfamilia de las hemoproteínas y, también, potenciales dianas a fármacos como se demostró en *Mycobacterium tuberculosis*. Están asociados a resistencias a fármacos en *Candida albicans* (33).

CNVs presentes en las líneas resistentes a anfotericina, paromomicina y antimoniales

La cobertura de lecturas en el cromosoma 29 de estas tres líneas resistentes mostró una distribución peculiar, como se ilustra en las figuras *Figura 10, p. 51; Figura 11, p. 52 y Figura 12, p. 52*.

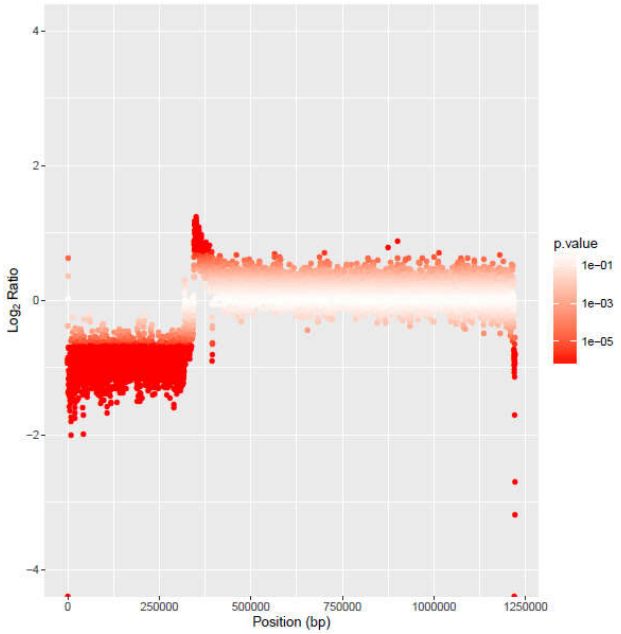

Figura 10. *Gráfico de coberturas de la línea resistente a anfotericina en relación a la cepa WT en el cromosoma 29*

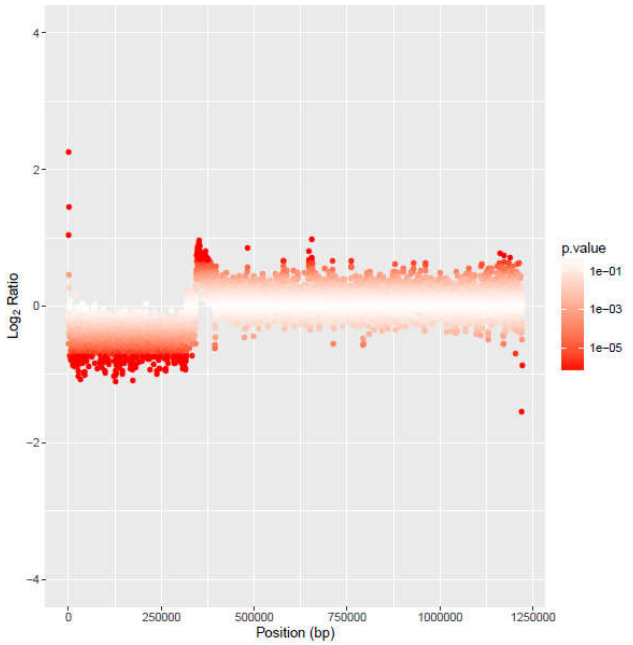

Figura 11. *Gráfico de coberturas de la línea resistente a paromomicina en relación a la cepa WT en el cromosoma 29*

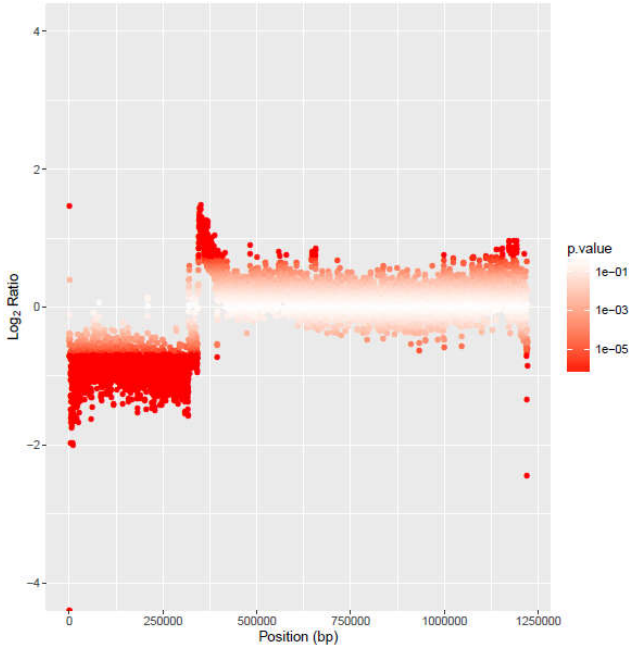

Figura 12. *Gráfico de coberturas de la línea resistente a antimoniales en relación a la cepa WT en el cromosoma 29*

Para analizar las similitudes en la distribución de la cobertura del cromosoma 29, directamente, en las tres líneas, se realizó una representación conjunta de coberturas en estas tres líneas junto con la línea parental (*Figura 13, p.54*).

Con esta información, se ha establecido la hipótesis de que hay una duplicación de aproximadamente dos tercios en el cromosoma 29 de estas tres líneas resistentes, que puede ser ventajosa para la adquisición de resistencia a estos fármacos por parte de las líneas. En la paromomicina, esta duplicación parcial, se encontraría solo en la mitad de los cromosomas 29. Esta explicación es compatible con lo que se ha visualizado en los cromosomas 29 de los gráficos de somía (*Figura 4. Gráfico de somía para la línia resistente a anfotericina (A), p.42; Figura 5. Gráfico de somía para la línia resistente a miltefosina (M), p.42; Figura 6. Gráfico de somía para la línia resistente a paromomicina (P), p.43*). Viéndose un aumento de una somía en las líneas resistentea a anfotericina y antimoniales respecto a la parental, y de media somía en la línea resistente a paromomicina.

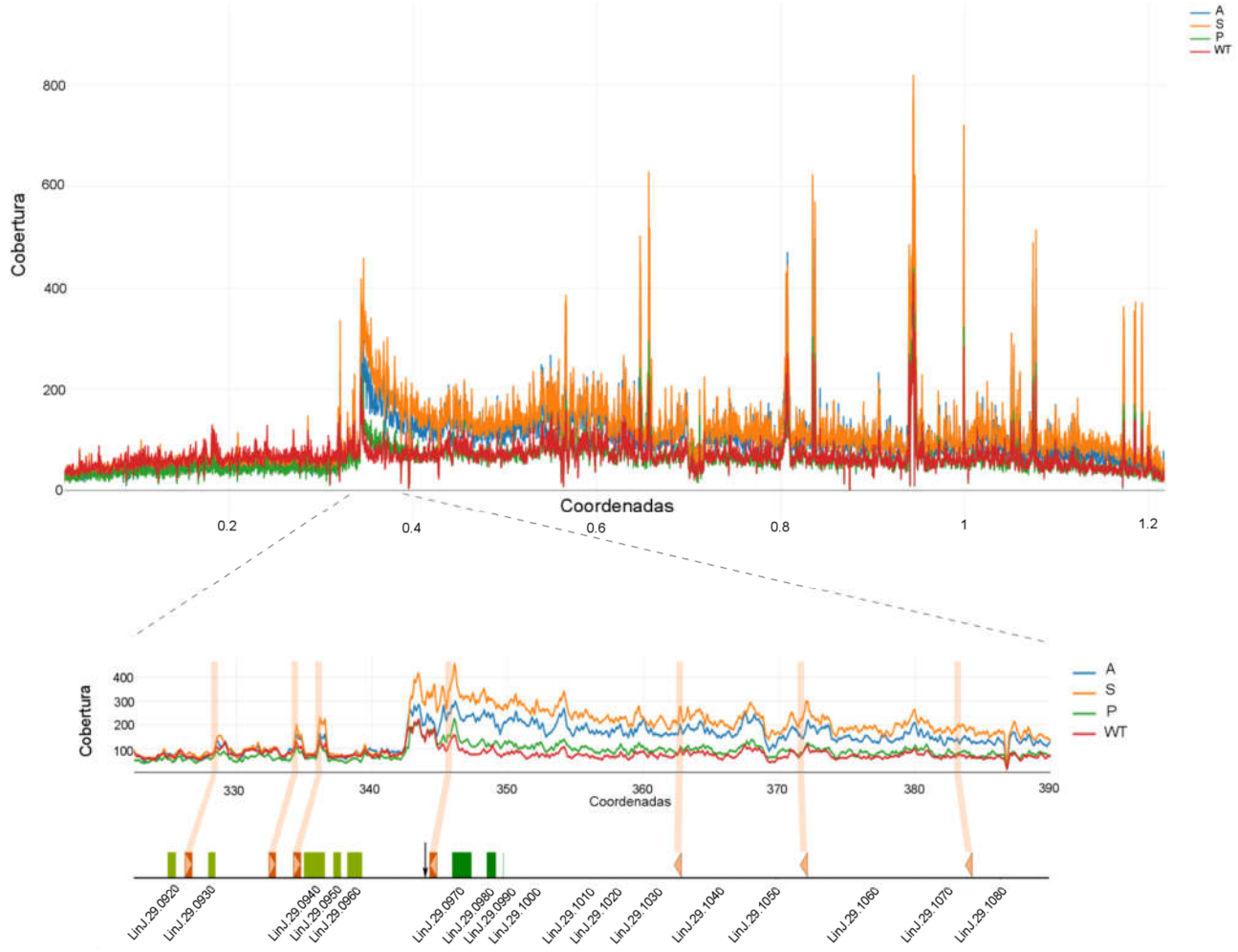

Figura 13. *Gráfico de cobertura del cromosoma 29 líneas resistentes a anfotericina, antimoniales y paromomicina y línea parental*

Nota. *'A' hace referencia a la línea resistente a anfotericina, 'P' hace referencia a la línea resistente a paromomicina, 'S' hace referencia a la línea resistente a antimoniales y 'Wt' hace referencia a la línea parental.*

El gráfico de cobertura superior muestra el cromosoma 29 entero. El segundo gráfico muestra la cobertura en la región de cambio de somía en las líneas resistentes. En la parte inferior aparece un esquema a escala donde aparecen los diferentes genes presentes en la región del segundo gráfico de coberturas, con una flecha en negro se indica la zona estimada de cambio de somía.

Los gráficos de cobertura estan realizados con el paquete plotly() (34) de R.

3.2. Análisis de SNPs

3.2.1. A nivel de genoma

Como se ha detallado en el apartado de materiales y métodos, se realizó primeramente una búsqueda de SNPs a lo largo del genoma. En la *Tabla 18, p. 54*, se muestra la cabecera del fichero generado con la lista de SNPs en el genoma y la frecuencia de bases por cada uno de ellos en cada línea de estudio.

SNPs A		WT				A				M				P				S			
Cromosoma	Posición	A	C	G	T	A	C	G	T	A	C	G	T	A	C	G	T	A	C	G	T
LinJ.07	14805	0,00	68,18	31,82	0,00	0,00	89,47	10,53	0,00	0,00	50,00	50,00	0,00	0,00	66,67	33,33	0,00	0,00	60,87	39,13	0,00
LinJ.07	14872	0,00	45,00	0,00	55,00	0,00	26,67	0,00	73,33	0,00	66,67	0,00	33,33	0,00	28,57	0,00	71,43	0,00	43,75	0,00	56,25
LinJ.07	59401	30,00	0,00	70,00	0,00	41,67	0,00	58,33	0,00	31,43	0,00	68,57	0,00	37,04	0,00	62,96	0,00	27,03	0,00	72,97	0,00
LinJ.07	77897	0,00	0,00	100,00	0,00	0,00	11,11	88,89	0,00	0,00	0,00	100,00	0,00	0,00	4,55	95,45	0,00	0,00	0,00	100,00	0,00
LinJ.07	214646	43,75	0,00	0,00	56,25	46,67	0,00	0,00	53,33	43,75	0,00	0,00	56,25	46,15	0,00	0,00	53,85	47,06	0,00	5,88	47,06
LinJ.07	214653	55,56	0,00	44,44	0,00	61,11	0,00	38,89	0,00	52,94	0,00	47,06	0,00	53,85	0,00	46,15	0,00	60,00	0,00	40,00	0,00
LinJ.07	215794	0,00	0,00	48,39	51,61	0,00	0,00	59,09	40,91	0,00	0,00	49,37	50,63	0,00	0,00	45,10	54,90	0,00	0,00	43,43	56,57
LinJ.07	215847	0,00	0,00	86,00	14,00	0,00	0,00	66,67	33,33	0,00	0,00	78,43	21,57	0,00	0,00	75,86	24,14	0,00	0,00	81,36	18,64
LinJ.07	216063	52,38	0,00	47,62	0,00	47,06	0,00	52,94	0,00	63,16	0,00	36,84	0,00	61,90	0,00	38,10	0,00	62,50	0,00	37,50	0,00
LinJ.07	254553	0,00	10,71	0,00	89,29	0,00	21,05	0,00	78,95	0,00	7,41	0,00	92,59	0,00	10,00	0,00	90,00	0,00	10,00	0,00	90,00
LinJ.07	254559	0,00	12,00	0,00	88,00	0,00	23,53	0,00	76,47	0,00	7,69	0,00	92,31	0,00	10,00	0,00	90,00	0,00	14,29	0,00	85,71
LinJ.07	254582	0,00	16,67	83,33	0,00	0,00	26,67	73,33	0,00	0,00	11,11	88,89	0,00	0,00	12,50	87,50	0,00	0,00	16,67	83,33	0,00
LinJ.07	341112	88,24	0,00	11,76	0,00	75,00	0,00	25,00	0,00	88,24	0,00	11,76	0,00	100,00	0,00	0,00	0,00	81,82	0,00	18,18	0,00

Tabla 19. *Fragmento de la tabla de frecuencias de SNPs obtenida (en este fragmento solo se ven parte de los SNPs en la línea resistente a anfotericina).*

Nota. *WT hace referencia a la línea parental, A hace referencia a la línea resistente a anfotericina, M hace referencia a la línea resistente a miltefosina, P hace referencia a la línea resistente a paromomicina y S hace referencia a la línea resistente a antimoniales. La tabla completa se puede consultar en el Anexo 3 adjunto.*

Para visualizar el solapamiento de SNPs se realizó un diagrama de *Venn* con el programa de diagramas de *Venn* de la Universidad de Ghent (35). El resultado se muestra en la *Figura 14 p. 54*.

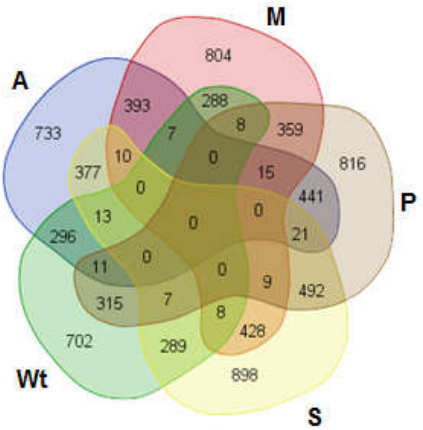

Figura 14. *Diagrama de Venn de la distribución de SNPs en todo el genoma*

Nota. *'A' hace referencia a la línea resistente a anfotericina, 'M' hace referencia a la línea resistente a miltefosina, 'P' hace referencia a la línea resistente a paromomicina, 'S' hace referencia a la línea resistente a antimoniales y 'Wt' hace referencia a la línea parental.*

Por otro lado, se realizó un análisis gráfico sobre la distribución de SNPs en los distintos cromosomas y en las distintas líneas resistentes. Un ejemplo se muestra en la *Figura 15, p. 56.*

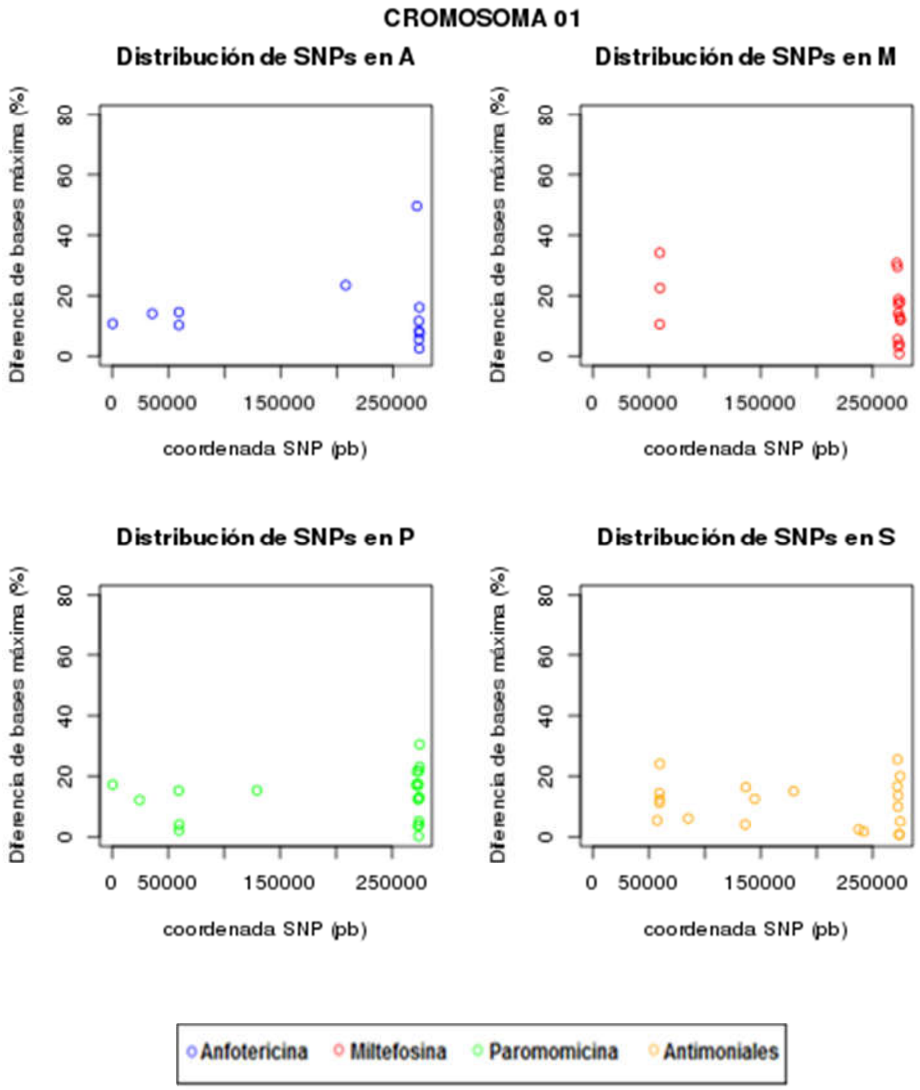

Figura 15. *Ejemplo de gráfico de distribución de SNPs para el cromosoma 1*

Nota. *'A' hace referencia a la línea resistente a anfotericina, 'M' hace referencia a la línea resistente a miltefosina, 'P' hace referencia a la línea resistente a paromomicina y 'S' hace referencia a la línea resistente a antimoniales.*

El resto de gráficos están en el Anexo 4 adjunto.

A continuación, se tratan aquellos casos en los que se han encontrado diferencias significativas entre alguna línea resistente y la línea parental.

Regiones con diferente distribución de SNPs.

Cromosoma 5 → región con más SNPs en la línea resistente a antimoniales (coordenadas 225,089-227,811).

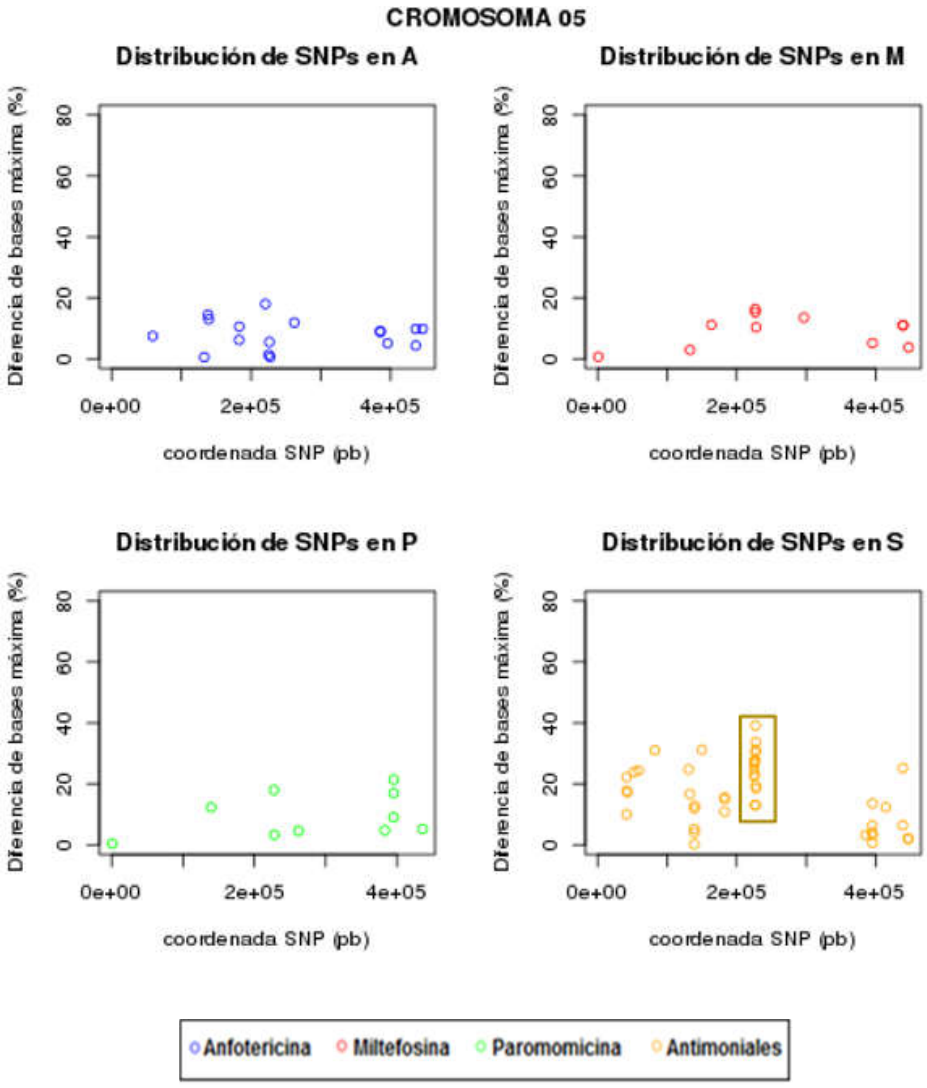

Figura 16. *Gráfico de distribución de SNPs en el cromosoma 5*

Nota. *'A' hace referencia a la línea resistente a anfotericina, 'M' hace referencia a la línea resistente a miltefosina, 'P' hace referencia a la línea resistente a paromomicina y 'S' hace referencia a la línea resistente a antimoniales.*

Este gráfico tiene resaltada una zona con acumulación diferenciada de SNPs en la línea resistente a antimoniales.

Afecta a una proteína hipotética (LinJ.05.0670). En este caso, dado que no se conoce una posible función para esta proteína, no es posible inferir la ventaja que podría conferir esta acumulación de SNPs en la línea resistente a antimoniales.

Cromosoma 10 → región con más SNPs en M (coordenadas 535,874-545,986). Genes afectados:

- Transportador de pteridina (LinJ.10.1450)
- Desaturasa de ácidos grasos (LinJ.10.1460). Posición polimórfica en Wt.

En la literatura hay un estudio que muestra que el aumento del transporte de las pteridinas contribuye a la resistencia al metotrexato en *L. tarentolae* (36). Esta ventaja podría ser común en nuestra línea de estudio resistente a miltefosina, aunque aún está por determinar las razones biológicas.

Los resultados de otro estudio (37) sugieren que las membranas plasmáticas de los parásitos de *Leishmania* resistentes a miltefosina podrían ser menos fluidas que en Wt, debido a su contenido de ácidos grasos insaturados, provocando el aumento de *fitness* en nuestra línea. Esta bajada del contenido de cadenas alquil insaturadas se podría explicar por una inactivación parcial de las desaturasas. La reacción de desaturación, sin embargo, no se ha descrito aún en Leishmania.

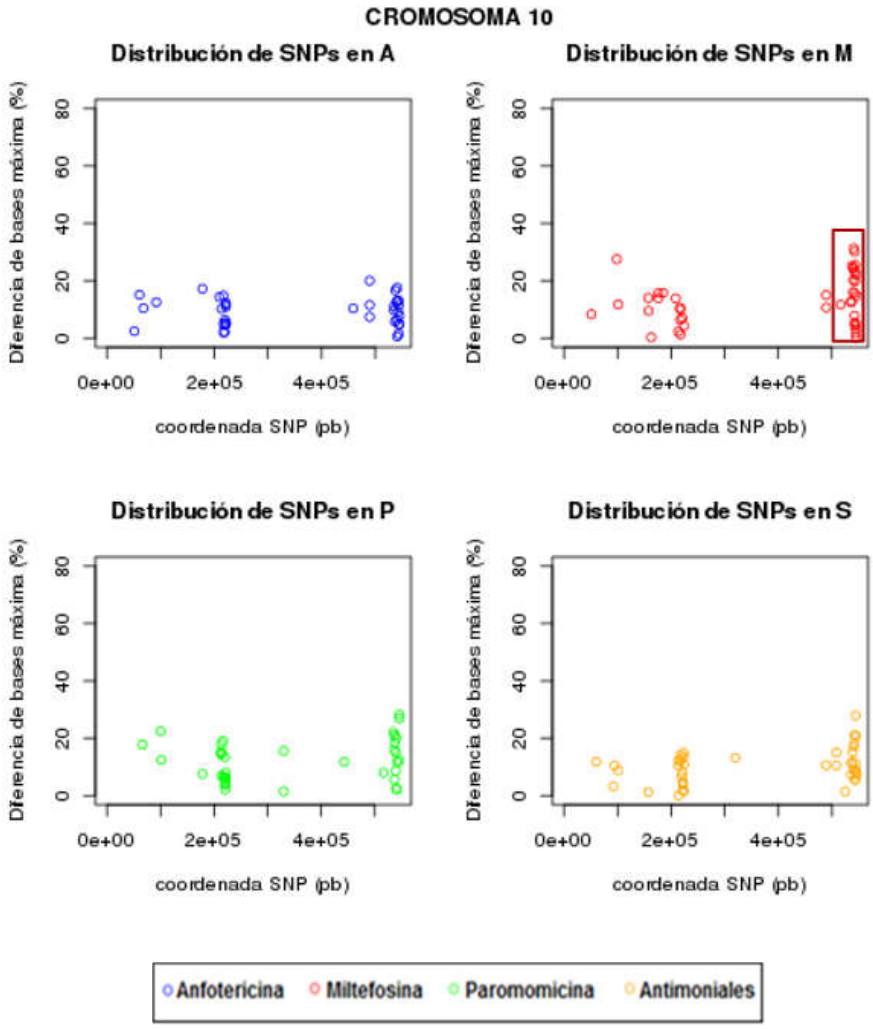

Figura 17. *Gráfico de distribución de SNPs en el cromosoma 10*

Nota. *'A' hace referencia a la línea resistente a anfotericina, 'M' hace referencia a la línea resistente a miltefosina, 'P' hace referencia a la línea resistente a paromomicina y 'S' hace referencia a la línea resistente a antimoniales.*

Este gráfico tiene resaltada una zona con acumulación diferenciada de SNPs en la línea resistente a miltefosina.

Cromosoma 21→ región con más SNPs en antimoniales (coordenadas 195,687-219,423).

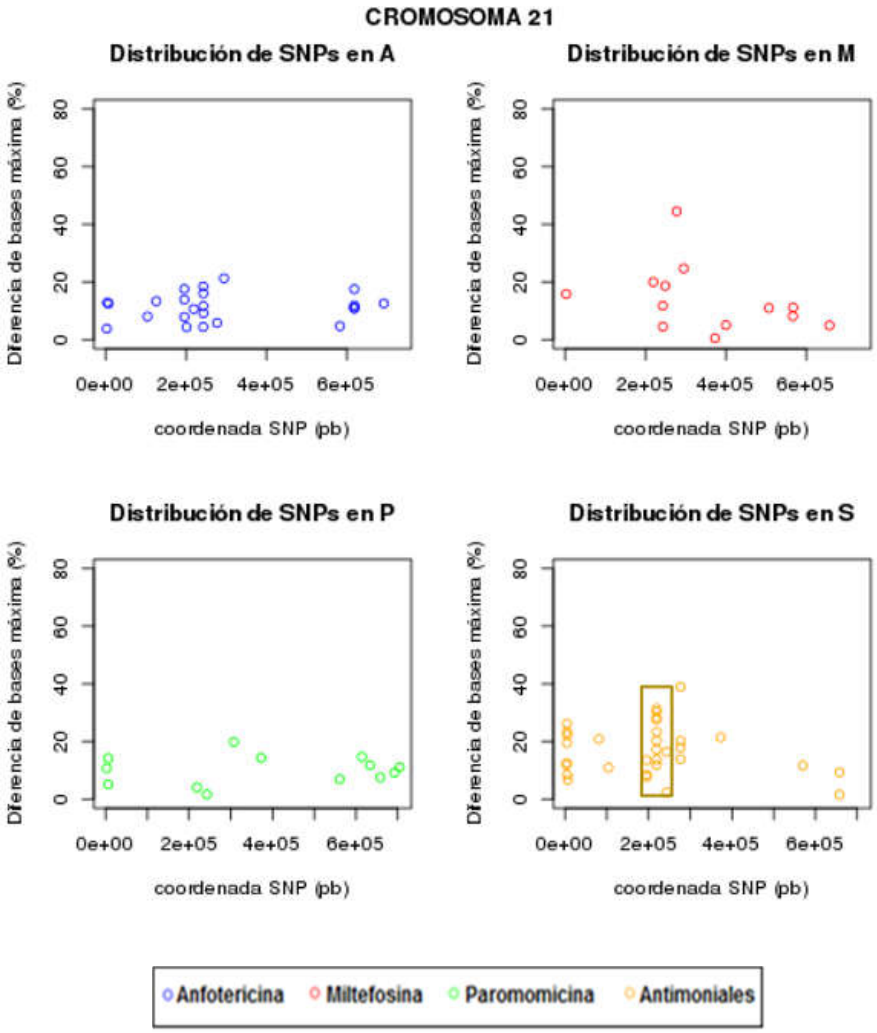

Figura 18. Gráfico de distribución de SNPs en el cromosoma 21

Nota. *'A' hace referencia a la línea resistente a anfotericina, 'M' hace referencia a la línea resistente a miltefosina, 'P' hace referencia a la línea resistente a paromomicina y 'S' hace referencia a la línea resistente a antimoniales.*

Este gráfico tiene resaltada una zona con acumulación diferenciada de SNPs en la línea resistente a antimoniales.

Es una zona con muchos polimorfismos y parece que está colapsada en el genoma de referencia, de modo que parece que no está bien anotada. No afecta a ningún gen. Por el momento, no se puede determinar posibles implicaciones de esta acumulación de SNPs en la línea resistente a antimoniales.

Cromosoma 26 → región con más SNPs en antimoniales (coordenadas 126,944-167,577).

En esta zona hay acumulación de muchos polimorfismos. Los genes afectados son afectados:

- Proteína hipotética conservada, pseudogen (LinJ.26.0500).

- Proteína tipo galactofuranosyltransferasa lpg1 (LinJ.26.0520).

- Proteína hipotética conservada (LinJ.26.0530).

- Proteína hipotética conservada (LinJ.26.0570).

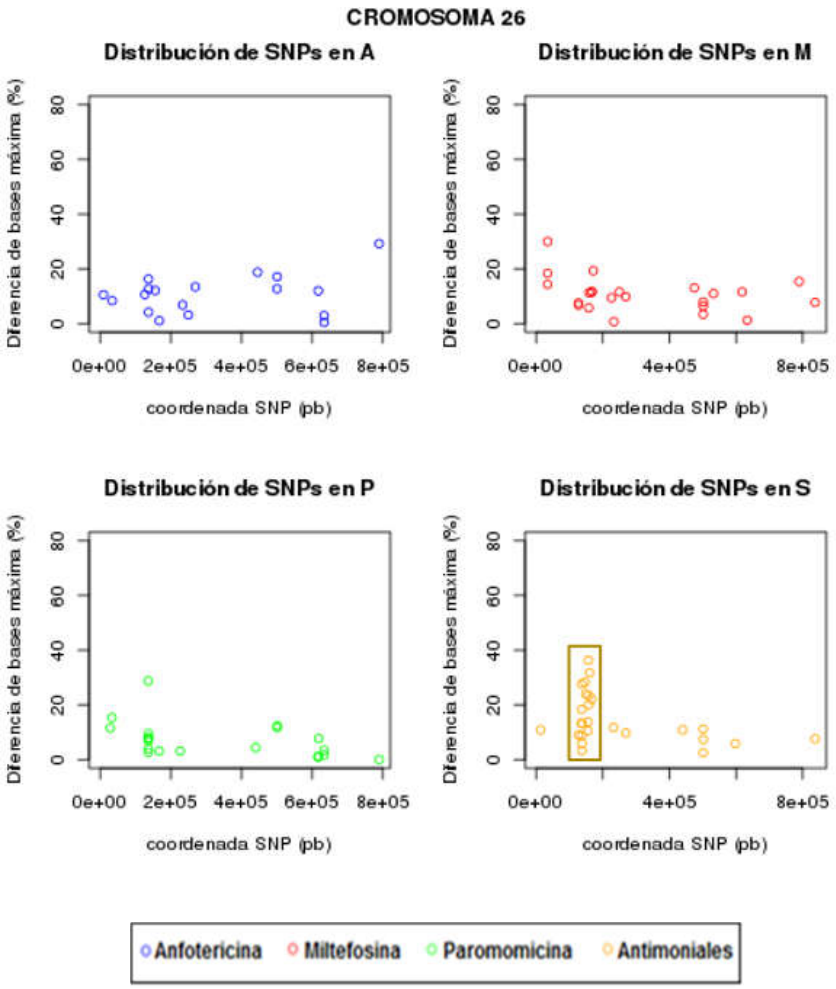

Figura 19. *Gráfico de distribución de SNPs en el cromosoma 26*

Nota. *'A' hace referencia a la línea resistente a anfotericina, 'M' hace referencia a la línea resistente a miltefosina, 'P' hace referencia a la línea resistente a paromomicina y 'S' hace referencia a la línea resistente a antimoniales.*

Este gráfico tiene resaltada una zona con acumulación diferenciada de SNPs en la línea resistente a antimoniales.

De todos los genes afectados, tan solo está descrito el que codifica para la proteína tipo galactofuranosyltransferasa lpg1 (LinJ.26.0520). No se ha encontrado ninguna relación de este gen con resistencias a fármacos.

Cromosoma 27 → región con más SNPs en P (coordenadas 857,320-961,765).

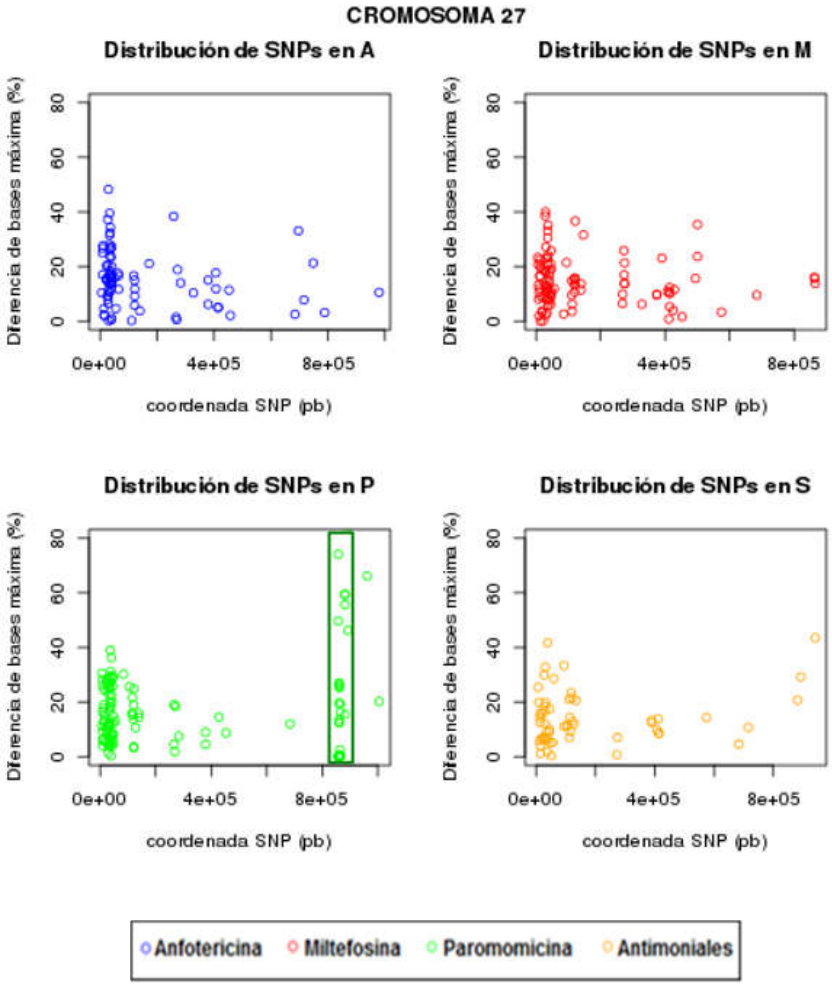

Figura 20. *Gráfico de distribución de SNPs en el cromosoma 27*

Nota. *'A' hace referencia a la línea resistente a anfotericina, 'M' hace referencia a la línea resistente a miltefosina, 'P' hace referencia a la línea resistente a paromomicina y 'S' hace referencia a la línea resistente a antimoniales.*

Este gráfico tiene resaltada una zona con acumulación diferenciada de SNPs en la línea resistente a paromomicina.

Esta zona corresponde al amplicón de la paromomicina determinado en el análisis de CNVs (*3.1. Análisis de CNVs, p. 40*). Los genes afectados son:

- D-lactato deshidrogenasa (LinJ.27.1940).

- Aminotransferasa de aminoácidos de cadena ramificada (LinJ.27.1950)

Cromosoma 28 → región con más SNPs en M (coordenadas 113,618-119,556).

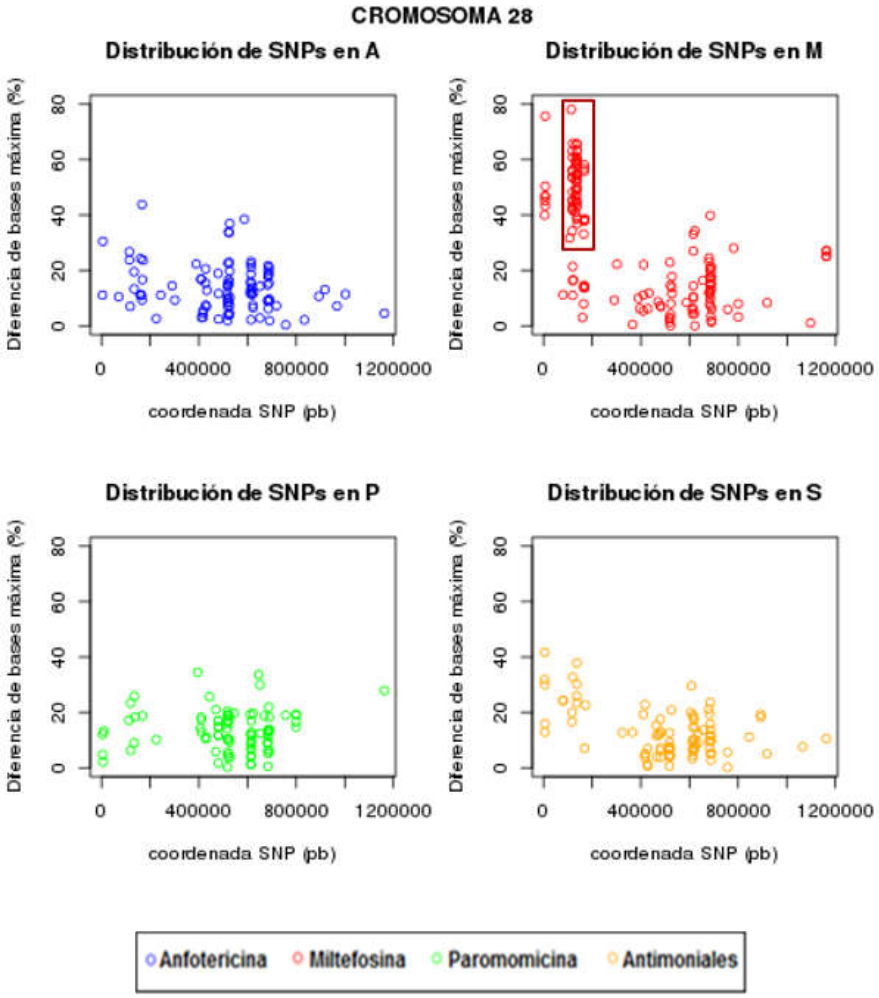

Figura 21. *Gráfico de distribución de SNPs en el cromosoma 28*

Nota. *'A' hace referencia a la línea resistente a anfotericina, 'M' hace referencia a la línea resistente a miltefosina, 'P' hace referencia a la línea resistente a paromomicina y 'S' hace referencia a la línea resistente a antimoniales.*

Este gráfico tiene resaltada una zona con acumulación diferenciada de SNPs en la línea resistente a miltefosina.

Esta zona tiene muchos **polimorfismos** que pasan a ser posiciones **homocigóticas** en la línea resistente a miltefosina. En la siguiente tabla (*Tabla 20, p.65)* se puede ver un resumen de este fenómeno.

Coord.	*Gen ID*	*Anotación funcional del gen*	*Información*
4655	NO	-	SNP que pasa de polimórfico a homocigoto (100%)
6122	NO	-	SNP que pasa de polimórfico a homocigoto (100%)
6221	NO	-	SNP que pasa de polimórfico a homocigoto
6685	NO	-	SNP que pasa de polimórfico a homocigoto
6728	NO	-	SNP que pasa de polimórfico a homocigoto
6762	NO	-	SNP que pasa de polimórfico a homocigoto

6819	NO	-	SNP que pasa de polimórfico a homocigoto
80564	*NO*	*-*	*Posición polimórfica en todas las líneas*
106795	*NO*	*-*	*Posición polimórfica en todas las líneas. Pero la base mayoritaria (79%) no es la misma.*
113618	LinJ.28.0350	hypothetical protein, conserved	SNP que pasa de homocigoto a homocigoto
117616	NO	-	SNP que pasa de polimórfico a homocigoto
117625	NO	-	SNP que pasa de polimórfico a homocigoto
117890	NO	-	SNP que pasa de polimórfico a homocigoto
118379	NO	-	SNP que pasa de polimórfico a homocigoto
118488	NO	-	SNP que pasa de polimórfico a homocigoto
118605	NO	-	SNP que pasa de polimórfico a homocigoto
118744	NO	-	SNP que pasa de polimórfico a homocigoto
118799	NO	-	SNP que pasa de polimórfico a homocigoto (100)
119128	NO	-	SNP que pasa de polimórfico a homocigoto
119173	NO	-	SNP que pasa de polimórfico a homocigoto
119211	NO	-	SNP que pasa de polimórfico a homocigoto
119222	NO	-	SNP que pasa de polimórfico a homocigoto
119346	NO	-	SNP que pasa de polimórfico a homocigoto
119556	NO	-	SNP que pasa de polimórfico a homocigoto
119662	*NO*	*-*	*Posición polimórfica en todas las líneas*
119825	*NO*	*-*	*Posición polimórfica en todas las líneas*
120085	*NO*	*-*	*Posición polimórfica en todas las líneas*
122320	*NO*	*-*	*Posición polimórfica en todas las líneas (excepto P)*
131649	LinJ.28.0380	hypothetical protein, conserved	SNP que pasa de polimórfico a homocigoto
131918	LinJ.28.0380	hypothetical protein, conserved	SNP que pasa de polimórfico a homocigoto
133049	NO	-	SNP que pasa de polimórfico a homocigoto (100%)
133129	NO	-	SNP que pasa de polimórfico a 87%
134533	NO	-	SNP que pasa de polimórfico a homocigoto
134570	NO	-	SNP que pasa de polimórfico a homocigoto
134662	NO	-	SNP que pasa de polimórfico a homocigoto (100%)
134918	NO	-	SNP que pasa de polimórfico a homocigoto (100%)
134962	NO	-	SNP que pasa de polimórfico a homocigoto
134984	NO	-	SNP que pasa de polimórfico a homocigoto (100%)
135047	NO	-	SNP que pasa de polimórfico a homocigoto (100%)

135146	NO	-	SNP que pasa de polimórfico a homocigoto
135456	NO	-	SNP que pasa de polimórfico a homocigoto
135465	NO	-	SNP que pasa de polimórfico a homocigoto
135555	NO	-	SNP que pasa de polimórfico a homocigoto (100%)
135796	NO	-	SNP que pasa de polimórfico a homocigoto
136054	NO	-	SNP que pasa de polimórfico a homocigoto
136061	NO	-	SNP que pasa de polimórfico a homocigoto
136138	NO	-	SNP que pasa de polimórfico a homocigoto
136595	NO	-	SNP que pasa de polimórfico a homocigoto
136612	NO	-	SNP que pasa de polimórfico a homocigoto
136712	NO	-	SNP que pasa de polimórfico a homocigoto
136817	NO	-	SNP que pasa de polimórfico a homocigoto
136856	NO	-	SNP que pasa de polimórfico a homocigoto (100%)
136999	NO	-	SNP que pasa de polimórfico a homocigoto
137000	NO	-	SNP que pasa de polimórfico a homocigoto
137077	NO	-	SNP que pasa de polimórfico a homocigoto
137078	NO	-	SNP que pasa de polimórfico a homocigoto
137181	NO	-	SNP que pasa de polimórfico a 89%
137315	NO	-	SNP que pasa de polimórfico a homocigoto
137417	NO	-	SNP que pasa de polimórfico a homocigoto (100%)
137445	NO	-	SNP que pasa de polimórfico a homocigoto
159865	*NO*	-	*Posición polimórfica en todas las cepas.*
164861	*NO*	-	*Posición polimórfica en todas las cepas.*
164990	*NO*	-	*Posición polimórfica en todas las cepas.*
165255	*NO*	-	-
165258	*NO*	-	-
166291	*LinJ.28.0490*	*Sugar efflux transporter for intercellular exchange, putative*	*SNP que pasa de polimórfico a 83%*
167130	NO	-	SNP que pasa de polimórfico a homocigoto
167401	LinJ.28.0500	ubiquitin activating enzyme, putative	SNP que pasa de polimórfico a homocigoto
167713	LinJ.28.0500	ubiquitin activating enzyme, putative	SNP que pasa de polimórfico a homocigoto
168516	NO	-	SNP que pasa de polimórfico a homocigoto
168537	NO	-	SNP que pasa de polimórfico a homocigoto
168834	*NO*	-	*Posición polimórfica en todas las cepas.*
168997	*NO*	-	*Posición polimórfica en todas las cepas.*
169699	NO	-	SNP que pasa de polimórfico a homocigoto

Tabla 20. *SNPs de la zona que presenta acumulación en el cromosoma 28 de la línea resistente a la miltefosina*

Nota: las posiciones marcadas en gris **no** pasan de polimórficas a homocigóticas.

Estos datos sugieren que en esta región debe estar codificada alguna proteína que confiere alguna ventaja para la línea resistente a miltefosina.

La razón de esta selección positiva de SNPs puede no estar ligada a la expresión de variantes proteicas, ya que la mayoría de los SNPs que han derivado a una homocigósis del 100% están en regiones no codificantes, así como los SNPs que más han variado sus frecuencias.

Los genes que se encuentran en esta zona son:

- Proteína hipotética (LinJ.28.0350)
- Proteína hipotética (LinJ.28.0380)
- Transportador de eflujo de azúcar para intercambio intercelular (LinJ.28.0490)
- Enzima activadora de ubiquitina (LinJ.28.0500)

Los dos primeros genes mencionados (LinJ.28.0350 y LinJ.28.0380) son proteínas hipotéticas que no están caracterizadas. Por otro lado, hasta el momento, no se conoce posibles implicaciones en la resistencia a fármacos del transportador de azúcar (LinJ.28.0490), ni tampoco la enzima activadora de ubiquitina (LinJ.28.0500). Esta última proteína, sin embargo, al ser un activador de ubiquitina, puede tener un gran impacto en muchas vías y que la variante seleccionada en nuestra línea resistente a miltefosina sea favorable por verse afectado uno de sus mecanismos de actuación de manera indirecta. De todos modos, harían falta más estudios experimentales para validar esta hipótesis.

Cromosoma 34 → región con más SNPs en M (coordenadas 1,370,997-1,621,947).

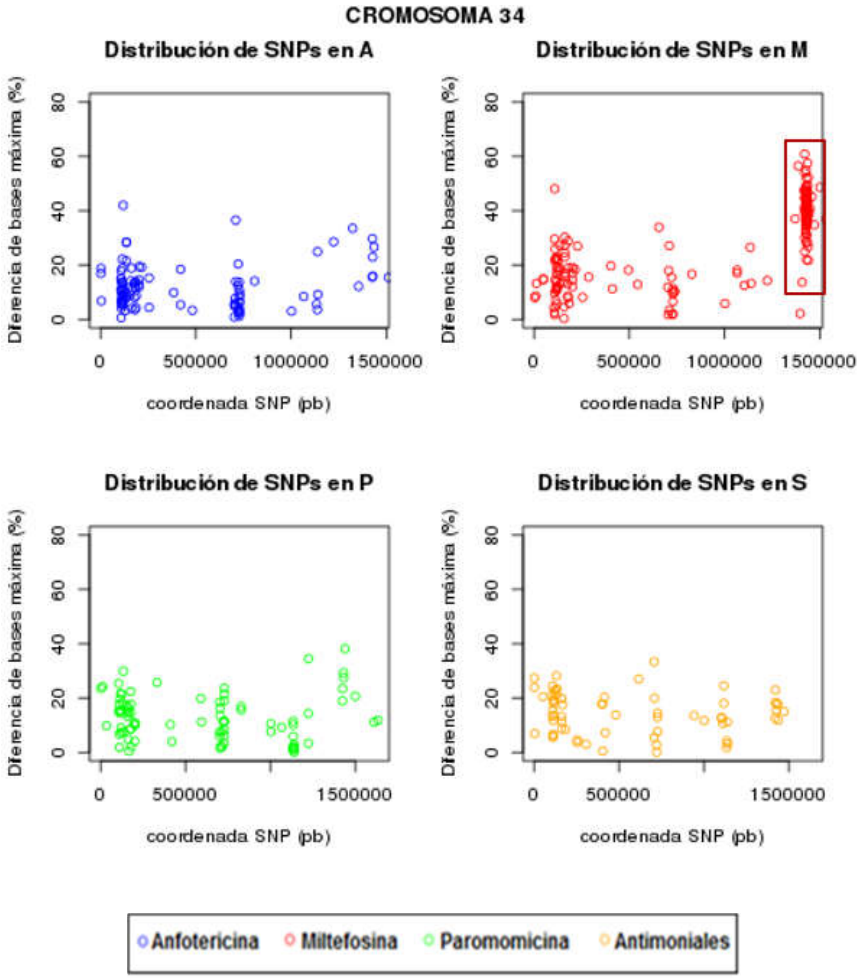

Figura 22. *Gráfico de distribución de SNPs en el cromosoma 34*

Nota.. *'A' hace referencia a la línea resistente a anfotericina, 'M' hace referencia a la línea resistente a miltefosina, 'P' hace referencia a la línea resistente a paromomicina y 'S' hace referencia a la línea resistente a antimoniales.*

Este gráfico tiene resaltada una zona con acumulación diferenciada de SNPs en la línea resistente a miltefosina.

Esta zona contiene muchos **polimorfismos** que pasan a ser posiciones **homocigóticas** *(Tabla 21, p.69)*.

Coord.	***Gen ID***	***Anotación funcional del gen***	***Información***
1370997	LinJ.34.3330	Programmed cell death protein 2, C-terminal putative domain containing protein, putative	SNP que pasa de polimórfico a homocigoto
1386373	NO	-	SNP que pasa de polimórfico a homocigoto
1396507	NO	-	-
1407539	NO	-	SNP que pasa de polimórfico a 88%
1419702	NO	-	SNP que pasa de polimórfico a 88%

1419737	NO	-	SNP que pasa de polimórfico a 87%
1419769	NO	-	SNP que pasa de polimórfico a homocigótico
1420960	LinJ.34.3480	hypothetical protein, conserved	SNP que pasa de polimórfico a homocigótico
1421146	LinJ.34.3480	hypothetical protein, conserved	SNP que pasa de polimórfico a 82%
1421378	LinJ.34.3480	hypothetical protein, conserved	SNP que pasa de polimórfico a homocigótico
1421388	LinJ.34.3480	hypothetical protein, conserved	SNP que pasa de polimórfico a 88%
1421444	LinJ.34.3480	hypothetical protein, conserved	SNP que pasa de polimórfico a 83%
1421529	LinJ.34.3480	hypothetical protein, conserved	SNP que pasa de polimórfico a homocigótico
1421590	LinJ.34.3480	hypothetical protein, conserved	SNP que pasa de polimórfico a 89%
1421760	LinJ.34.3480	hypothetical protein, conserved	SNP que pasa de polimórfico a homocigótico
1424831	NO	-	SNP que pasa de polimórfico a 87%
1424943	NO	-	SNP que pasa de polimórfico a homocigótico
1425243	NO	-	SNP que pasa de polimórfico a homocigótico
1425279	NO	-	SNP que pasa de polimórfico a homocigótico
1425285	NO	-	SNP que pasa de polimórfico a homocigótico
1425364	LinJ.34.3490	inositol-pentakisphosphate 2-kinase, putative	SNP que pasa de polimórfico a 83%
1425457	LinJ.34.3490	inositol-pentakisphosphate 2-kinase, putative	SNP que pasa de polimórfico a 88%
1425487	LinJ.34.3490	inositol-pentakisphosphate 2-kinase, putative	SNP que pasa de polimórfico a 84%
1425599	LinJ.34.3490	inositol-pentakisphosphate 2-kinase, putative	SNP que pasa de polimórfico a 89%
1425771	LinJ.34.3490	inositol-pentakisphosphate 2-kinase, putative	SNP que pasa de polimórfico a 87%
1425846	LinJ.34.3490	inositol-pentakisphosphate 2-kinase, putative	SNP que pasa de polimórfico a homocigótico
1426376	LinJ.34.3490	inositol-pentakisphosphate 2-kinase, putative	SNP que pasa de polimórfico a homocigótico
1426463	LinJ.34.3490	inositol-pentakisphosphate 2-kinase, putative	SNP que pasa de polimórfico a homocigótico
1426647	LinJ.34.3490	inositol-pentakisphosphate 2-kinase, putative	SNP que pasa de polimórfico a 86%
1426664	LinJ.34.3490	inositol-pentakisphosphate 2-kinase, putative	SNP que pasa de polimórfico a homocigótico
1426880	NO	-	SNP que pasa de polimórfico a 84%
1426935	NO	-	SNP que pasa de polimórfico a homocigótico

1427410	NO	-	SNP que pasa de polimórfico a 88%
1427487	NO	-	SNP que pasa de polimórfico a 87%
1427569	NO	-	SNP que pasa de polimórfico a 81%
1427888	NO	-	SNP que pasa de polimórfico a 86%
1428149	LinJ.34.3500	hypothetical protein, conserved	SNP que pasa de polimórfico a 85%
1428307	LinJ.34.3500	hypothetical protein, conserved	SNP que pasa de polimórfico a homocigótico
1428591	LinJ.34.3500	hypothetical protein, conserved	SNP que pasa de polimórfico a homocigótico
1428605	NO	-	SNP que pasa de polimórfico a homocigótico
1428748	NO	-	SNP que pasa de polimórfico a homocigótico
1429033	NO	-	SNP que pasa de polimórfico a homocigótico
1429326	NO	-	SNP que pasa de polimórfico a homocigótico
1429416	NO	-	SNP que pasa de polimórfico a homocigótico
1429465	LinJ.34.3510	Protein of unknown function (DUF501), putative	SNP que pasa de polimórfico a homocigótico
1429639	LinJ.34.3510	Protein of unknown function (DUF501), putative	SNP que pasa de polimórfico a homocigótico
1430103	LinJ.34.3510	Protein of unknown function (DUF501), putative	SNP que pasa de polimórfico a 89%
1430751	NO	-	SNP que pasa de polimórfico a 89%
1430817	NO	-	SNP que pasa de polimórfico a 84%
1430930	NO	-	SNP que pasa de polimórfico a 86%
1431655	LinJ.34.3520	hypothetical protein, conserved	SNP que pasa de polimórfico a 88%
1431969	LinJ.34.3520	hypothetical protein, conserved	SNP que pasa de polimórfico a 79%
1432060	LinJ.34.3520	hypothetical protein, conserved	SNP que pasa de polimórfico a 85%
1432427	NO	-	SNP que pasa de polimórfico a 86%
1432529	NO	-	SNP que pasa de polimórfico a 86%
1432606	NO	-	SNP que pasa de polimórfico a 92%
1432730	NO	-	SNP que pasa de polimórfico a 82%
1432785	NO	-	SNP que pasa de polimórfico a 83%
1433250	LinJ.34.3530	serine palmitoyltransferase-like protein	SNP que pasa de polimórfico a 92%
1433667	LinJ.34.3530	serine palmitoyltransferase-like protein	SNP que pasa de polimórfico a 95%
1434231	LinJ.34.3530	serine palmitoyltransferase-like protein	SNP que pasa de polimórfico a 91%
1434453	NO	-	SNP que pasa de polimórfico a 88%
1435518	NO	-	SNP que pasa de polimórfico a 89%

1435580	NO	-	SNP que pasa de polimórfico a 87%
1435858	NO	-	SNP que pasa de polimórfico a homocigótico
1435940	NO	-	SNP que pasa de polimórfico a homocigótico
1436017	NO	-	SNP que pasa de polimórfico a homocigótico
1436553	NO	-	SNP que pasa de polimórfico a 88%
1436997	NO	-	SNP que pasa de polimórfico a 89%
1437094	NO	-	SNP que pasa de polimórfico a 80%
1437097	NO	-	SNP que pasa de polimórfico a 77%
1437119	NO	-	SNP que pasa de polimórfico a 85%
1437231	NO	-	SNP que pasa de polimórfico a 93%
1437893	LinJ.34.3540	hypothetical protein	SNP que pasa de polimórfico a 89%
1438607	LinJ.34.3540	hypothetical protein	SNP que pasa de polimórfico a 86%
1439248	NO	-	SNP que pasa de polimórfico a 84%
1439603	NO	-	SNP que pasa de polimórfico a homocigótica
1446303	NO	-	SNP que pasa de polimórfico a 88%
1448450	NO	-	SNP que pasa de polimórfico a 86%
1458160	NO	-	SNP que pasa de polimórfico a 90%
1471683	NO	-	SNP que pasa de polimórfico a 79%
1500295	LinJ.34.3740	expression site-associated protein 5 (ESAG5), putative	SNP que pasa de polimórfico a homocigótico
1533547	*LinJ.34.3800*	*AP-1 adapter complex gamma subunit, putative*	*SNP que pasa de homocigótico a heterocigótico*
1581110	NO	-	SNP que pasa de polimórfico a homocigótico
1582774	LinJ.34.3970	hypothetical protein, conserved	SNP que pasa de polimórfico a 80%
1595779	LinJ.34.3990	Cytoplasmic dynein 2 heavy chain (DYNC2H1), putative	SNP que pasa de polimórfico a 84%
1621947	LinJ.34.4090	**Unc104-like kinesin, putative**	SNP que pasa de polimórfico a 86%

Tabla 21. *SNPs de la zona que presenta acumulación en el cromosoma 28 de la línea resistente a la miltefosina*

Nota. *Las posiciones marcadas en gris* ***no*** *pasan de polimórficas a homocigóticas.*

Los genes que se encuentran en esta zona son:

- Proteína 2 de muerte celular programada (LinJ.34.3330).
- Quinasa-2 inositol-pentakisfosfato (LinJ.34.3490).
- Proteína de función desconocida (DUF501) (LinJ.34.3530).
- Proteína 5 de expresión asociada a sitio de expresión (ESAG5) (LinJ.34.3740).

- Adaptador AP-1 de la subunidad del complejo gamma (LinJ.34.3800).
- Cadena pesada de dineina citoplasmática 2 (DYNC2H1) (LinJ.34.3990).
- Proteína quinesina tipo Unc104 (LinJ.34.4090).

Para ninguno de estos genes se han encontrado estudios previos que los correlacionen con resistencias a fármacos en *Leishmania*. Sin embargo, cabe destacar que el SNP en el gen relacionado con la programación de muerte celular (LinJ.34.3330) sí podría estar implicado en el fenotipo de resistencia a miltefosina, ya que la vía de actuación de la droga es la inducción a muerte celular tipo apoptosis en *Leishmania*. La hipótesis de actuación es que el SNP impidiese generar una proteína completamente funcional de manera que se viera reducida la apoptosis inducida, ocasionando la resistencia. Por último, cabe destacar que en el gen de la quinesina Unc104 hay un SNP que implica truncamiento de la proteína (E1494*) que se encuentra en mayor frecuencia en la línea resistente a miltefosina.

3.2.2. A nivel de ORFs

Con los resultados obtenidos de SNPs en ORFs se realizó un nuevo diagrama de *Venn*:

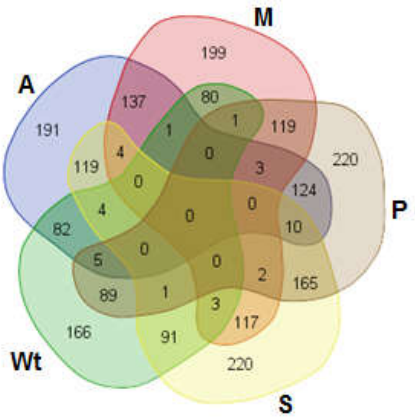

Figura 23 . *Diagrama de Venn de la distribución de SNPs en todo el genoma*

Nota. *'A' hace referencia a la línea resistente a anfotericina, 'M' hace referencia a la línea resistente a miltefosina, 'P' hace referencia a la línea resistente a paromomicina, 'S' hace referencia a la línea resistente a antimoniales y 'Wt' hace referencia a la línea parental.*

De esta manera, se puede visualizar rápidamente la distribución de SNPs en las diferentes líneas de estudio. Se puede ver cómo en la mayoría de los casos, los SNPs son propios de la línea o comunes solo en dos de las líneas.

A continuación, en la *Tabla 22, p.71* se comparan los SNPs obtenidos a nivel de genoma y en ORFs.

Línea	*SNPs nivel genoma*	*SNPs en ORFs*	*Relación (VNO/VNG)*
A	2309	680	29.45%
M	2339	666	28.47%
P	2494	739	29.63%
S	2552	736	28.84%
WT	1944	523	26.90%

Tabla 22. *Tabla comparativa de resultados de SNPs a nivel de genoma y en ORFs*
Nota. *'A' hace referencia a la línea resistente a anfotericina, 'M' hace referencia a la línea resistente a miltefosina, 'P' hace referencia a la línea resistente a paromomicina, 'S' hace referencia a la línea resistente a antimoniales y 'Wt' hace referencia a la línea parental.*

Como se puede comprobar, la relación (SNPs nivel genoma/SNPs en ORFs) se mantiene cercana a 1/3, cosa que entra dentro de lo esperado. Fijarse que en la línea WT esta relación es ligeramente menor, esto puede ser debido a que esta línea no ha sido expuesta a ninguna presión selectiva.

En este punto, se consideró interesante realizar un diagrama de Venn complementario, esta vez con los genes donde había SNPs (*Figura 24, p.71*).

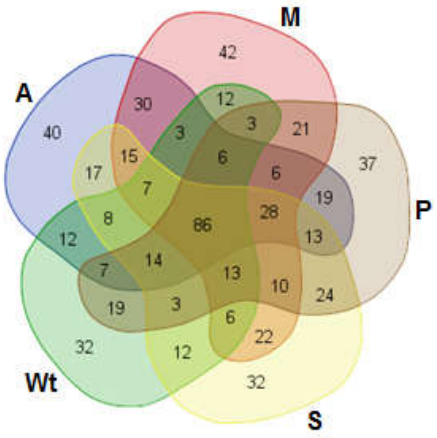

Figura 24. *Diagrama de Venn de la distribución de SNPs en los diferentes genes anotados*

Nota. *'A' hace referencia a la línea resistente a anfotericina, 'M' hace referencia a la línea resistente a miltefosina, 'P' hace referencia a la línea resistente a paromomicina, 'S' hace referencia a la línea resistente a antimoniales y 'Wt' hace referencia a la línea parental.*

En el genoma de referencia de *L. infantum* hay 8381 genes. Curiosamente, solo una pequeña parte de estos presentaba SNPs (599) y, a su vez, muchos de estos genes tenían más de un SNP asociado. En el diagrama, se puede ver

como, por ejemplo, hay 86 genes que presentan SNPs en todas las líneas de estudios y 32 que solo presentan SNPs en la línea WT.

Otro aspecto importante, que se ha analizado es la tabla de mutaciones obtenida de los SNPs detectados en ORFs. La tabla completa se puede consultar en el Anexo 5 de este TFM.

Como conclusiones de la información contenida en esta tabla, cabe destacar:

- 8 mutaciones que llevan a parada prematura de la traducción.

- 1175 mutaciones que implican un cambio aminoacídico en la proteína mutada.

- 970 mutaciones sinónimas en las proteínas mutadas.

Las mutaciones que implican cambio aminoacídico y las sinónimas no se analizaran en profundidad en este TFM.

Las 8 mutaciones que implican parada prematura son las que se recogen en la *Tabla 23, p.72*.

Lines	Gen ID	Coordinate in chromosome	Position in gene	Mutation	Kind of mutation	Gene function
P	LinJ.11.0040	12518	258	K87*	STOP	ATP-binding cassette protein subfamily H, member 1, putative (ABCH1)
M	LinJ.13.1590	619572	1040	W347*	STOP	phospholipid-transporting ATPase 1-like protein
M,S	LinJ.12.0667	413958	534	Q179*	STOP	hypothetical protein
M	LinJ.31.1470	656000	1122	R375*	STOP	hypothetical protein, unknown function
S	LinJ.30.2050	732514	57	R20*	STOP	ferric reductase transmembrane protein-like protein
A,M	LinJ.34.4090	1621947	4479	E1494*	STOP	Unc104-like kinesin, putative
S	LinJ.28.1730	614186	351	R118*	STOP	helicase-like protein
M	LinJ.27.0500	146174	165	Q56*	STOP	calpain-like cysteine peptidase, putative,cysteine peptidase, Clan CA, family C2, putative

Tabla 23. *Tabla de mutaciones de parada*

Nota. *'A' hace referencia a la línea resistente a anfotericina, 'M' hace referencia a la línea resistente a miltefosina, 'P' hace referencia a la línea resistente a paromomicina, 'S' hace referencia a la línea resistente a antimoniales y 'Wt' hace referencia a la línea parental.*

En la primera columna aparece la línea donde hay el SNP, en la segunda la ID del gen, en la tercera la coordenada del cromosoma donde se encuentra, en la cuarta la posición en el gen, en la quinta la mutación, en la sexta el tipo de mutación que supone y en la última la función del gen.

A continuación se muestra un análisis más exhaustivo de los genes en los que hay estos SNPs.

Análisis de las mutaciones de parada

Gen LinJ.11.0040

En este gen se ha encontrado una mutación en la proteína que codifica que implica una parada prematura en su síntesis. Esta mutación se encuentra al

inicio de la proteína en el aminoácido 87, posición que en la proteína Wt codifica para una lisina.

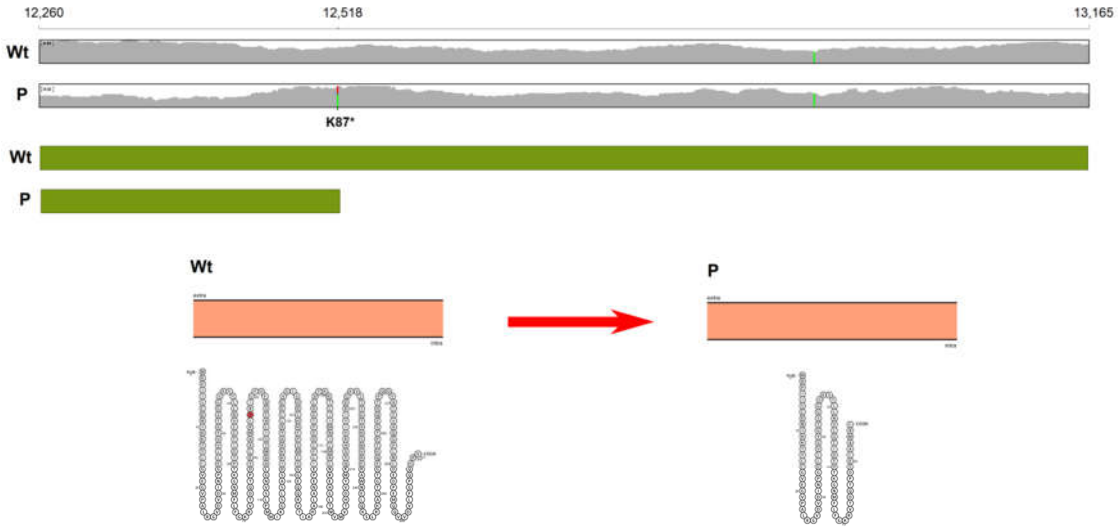

Figura 25. *Esquema de la mutación de parada encontrada en el gen LinJ.11.0040*

Nota. *'P' hace referencia a la línea resistente a paromomicina y 'Wt' hace referencia a la línea parental.*

En primer lugar, en la figura se muestra la cobertura del gen en el visualizador IGV donde se marca el SNP encontrado, En segundo lugar, se muestra la relación de longitud de la proteína entera y de la proteína truncada. En tercer lugar, se muestra un esquema de localización de la proteína realizado con el programa Protter (38), donde la parte superior indica el espacio extracelular, la parte salmón la membrana celular y la parte inferior el citosol.

El alelo mutado, se encuentra en la línea resistente a paromomicina en una frecuencia aproximada del 34,5%.

El gen LinJ.11.0040 codifica para la proteína con casete de unión a ATP (transportador ABC) de la subfamilia H, miembro 1 (ABCH1). Se han encontrado transportadores ABC ligados con resistencias a antimoniales (39). La mutación encontrada en la línea resistente a paromomicina podría favorecer la resistencia al fármaco, disminuyendo su entrada por la membrana celular del parásito.

Gen LinJ.13.1590

En este gen, la mutación se encuentra en el triplete para el aminoácido 347, que en la línea Wt codifica para un triptófano. El alelo mutado se encuentra en una frecuencia aproximada del 40% (*Figura 26, p.74*).

El gen codifica para la proteína tipo ATPasa 1 transportadora de fosfolípidos. Este gen se ha establecido, por estudios anteriores (40) como el transportador de la miltefosina (LdMT). En esos estudios previos, se han caracterizado líneas resistentes a miltefosina en *Leishmania* que presentan mutaciones en este gen,

tanto de parada como de cambio aminoacídico, siendo las de parada las que suelen conferir mayor resistencia al fármaco. De hecho, este gen es considerado un marcador de resistencia a miltefosina. Con todos estos datos, esta mutación por si sola serviría de validación de la resistencia en nuestra línea de estudio.

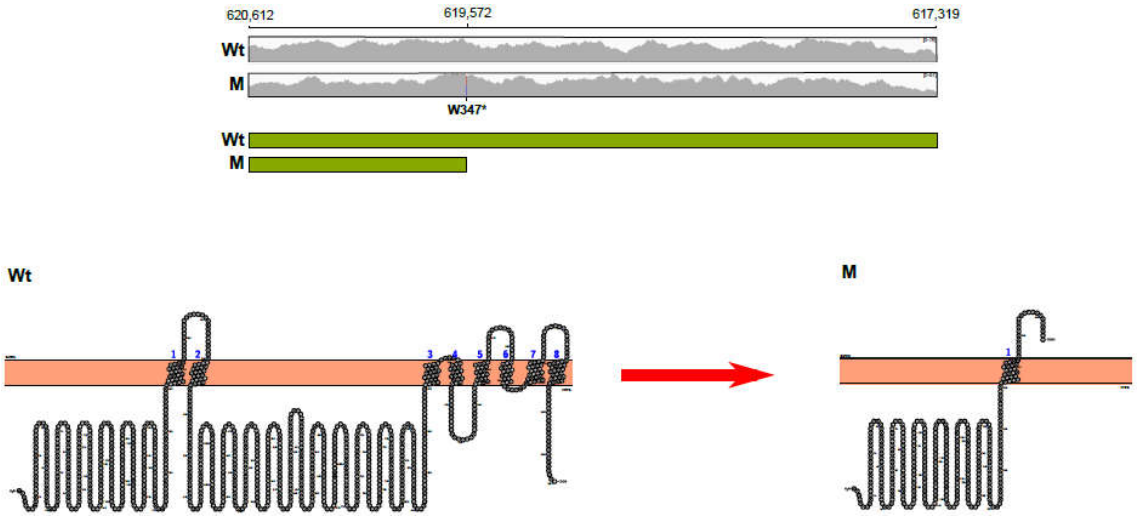

Figura 26. *Esquema de la mutación de parada encontrada en el gen LinJ.13.1590*

Nota. *'M' hace referencia a la línea resistente a miltefosina y 'Wt' hace referencia a la línea parental.*

En primer lugar, en la figura se muestra la cobertura del gen en el visualizador IGV donde se marca el SNP encontrado, En segundo lugar, se muestra la relación de longitud de la proteína entera y de la proteína truncada. En tercer lugar, se muestra un esquema de localización de la proteína realizado con el programa Protter, donde la parte superior indica el espacio extracelular, la parte salmón la membrana celular y la parte inferior el citosol.

Gen LinJ.12.0667

La posición de parada detectada en este gen se encuentra en una de las dos variantes que presenta de manera natural la cepa HU3 de *L. donovani*. En las líneas resistentes a miltefosina y anfotericina, la variante que implica parada se encuentra en mayor proporción que en la línea WT (WT: 48,89%, miltefosina: 58,82%, anfotericina: 57,58%).

Este gen, codifica para una proteína hipotética no caracterizada. Haría falta realizar más estudios para saber si el aumento proporcional del SNP de parada confiere alguna ventaja en las líneas de resistencia a miltefosina y anfotericina.

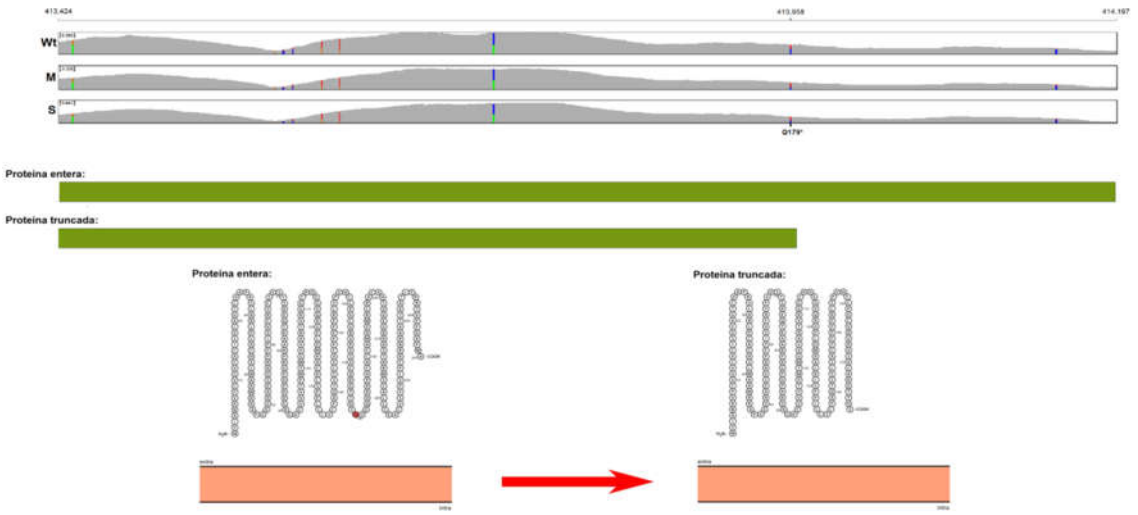

Figura 27. *Esquema de la mutación de parada encontrada en el gen LinJ.12.0667*

Nota. *'M' hace referencia a la línea resistente a miltefosina, 'S' hace referencia a la línea resistente a antimoniales y 'Wt' hace referencia a la línea parental.*

En primer lugar, en la figura se muestra la cobertura del gen en el visualizador IGV donde se marca el SNP encontrado, En segundo lugar, se muestra la relación de longitud de la proteína entera y de la proteína truncada. En tercer lugar, se muestra un esquema de localización de la proteína realizado con el programa Protter donde la parte superior indica el espacio extracelular, la parte salmón la membrana celular y la parte inferior el citosol.

Gen LinJ.31.1470

La posición de parada detectada en este gen también es una de las variantes naturales que se encuentran en la cepa HU3 de *L. donovani*. En la posición 375 de la proteína puede haber una arginina o un truncamiento de la proteína. En la línea resistente a la miltefosina, el alelo que implica un truncamiento se encuentra en mayor proporción que en la Wt (Wt: 56,12%, miltefosina: 60,56%).

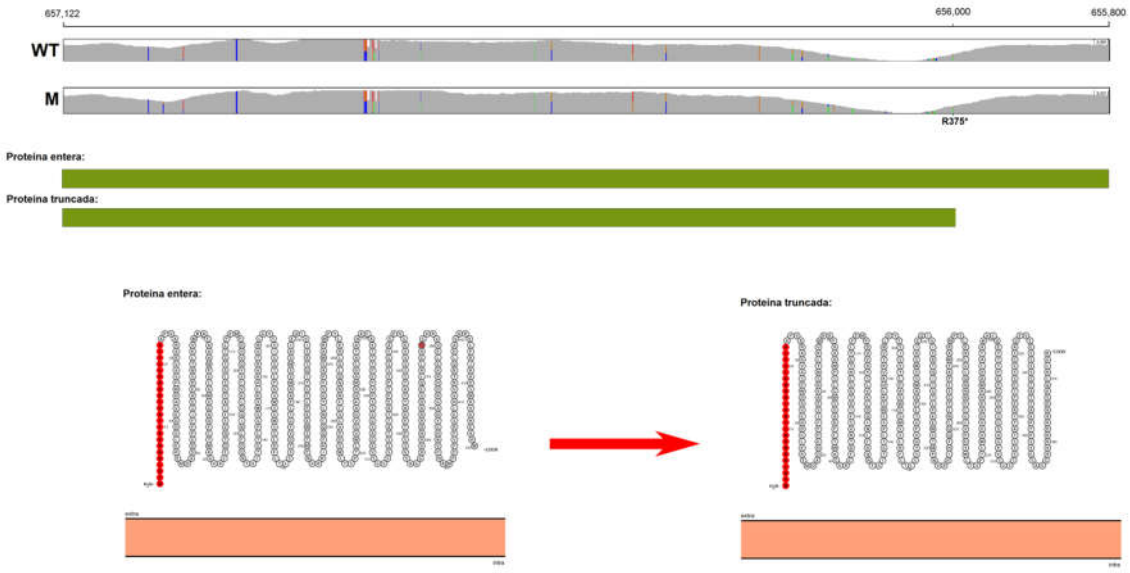

Figura 28. *Esquema de la mutación de parada encontrada en el gen LinJ.31.1470*

Nota. *'M' hace referencia a la línea resistente a miltefosina, y 'Wt' hace referencia a la línea parental.*

En primer lugar, en la figura se muestra la cobertura del gen en el visualizador IGV donde se marca el SNP encontrado, En segundo lugar, se muestra la relación de longitud de la proteína entera y de la proteína truncada. En tercer lugar, se muestra un esquema de localización de la proteína realizado con el programa Protter donde la parte superior indica el espacio extracelular, la parte salmón la membrana celular y la parte inferior el citosol.

Esta proteína también codifica para una proteína hipotética de función desconocida, aunque se predice la existencia de un péptido señal al principio de la proteína (conjunto de aminoácidos marcados en rojo, *Figura 28, p.76*).

Gen LinJ.30.2050

Este gen tiene de manera natural un polimorfismo en la posición 20 de la proteína que codifica. Una de las variantes de ese polimorfismo implica un truncamiento de la proteína. Dicha variante se encuentra en mayor proporción en la línea resistente a antimoniales que en la WT (Wt: 75%, antimoniales: 84%).

Este gen codifica para la proteína tipo reductasa férrica transmembrana. Tal como se puede comprobar en el esquema (*Figura 29, p.77*), la mutación de truncamiento está al inicio de la proteína, afectando a todos los dominios transmembrana.

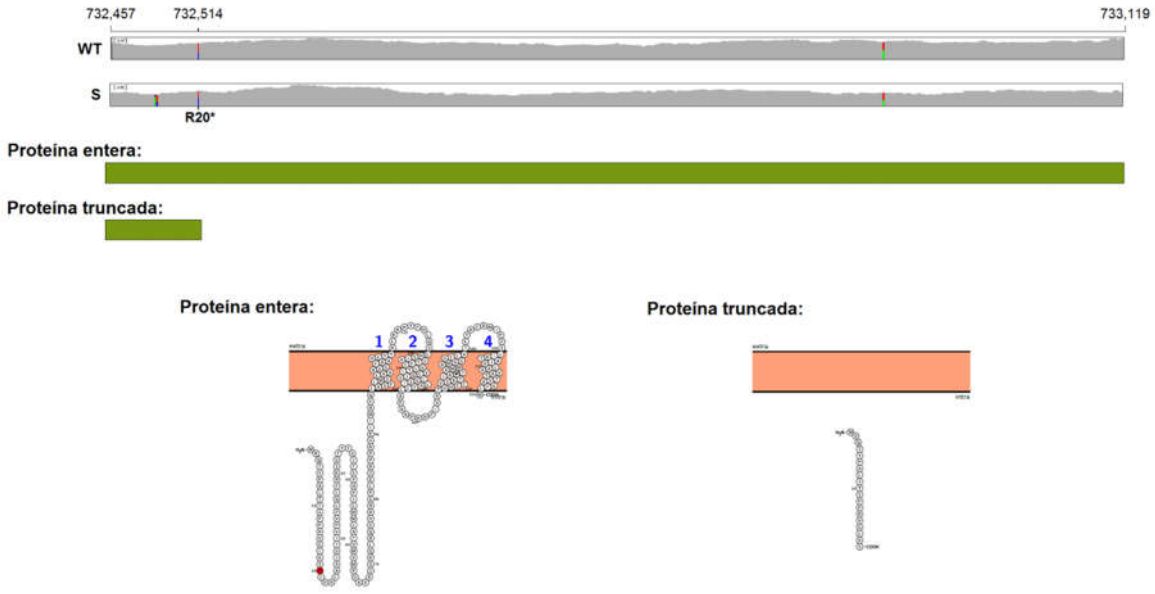

Figura 29. *Esquema de la mutación de parada encontrada en el gen LinJ.30.2050*

Nota. *'S' hace referencia a la línea resistente a antimoniales y 'Wt' hace referencia a la línea parental.*

En primer lugar, en la figura se muestra la cobertura del gen en el visualizador IGV donde se marca el SNP encontrado, En segundo lugar, se muestra la relación de longitud de la proteína entera y de la proteína truncada. En tercer lugar, se muestra un esquema de localización de la proteína realizado con el programa Protter donde la parte superior indica el espacio extracelular, la parte salmón la membrana celular y la parte inferior el citosol.

Se ha encontrado resistencia cruzada a antimoniales en cultivos ocasionadas por derivados del arsénico que no se encuentran de manera clínica (41). Los arsénicos tienen indudablemente, características comunes con los antimoniales en su modo de acción como por ejemplo mayor cantidad de especies reactivas del oxígeno (ROS), pérdida de potencial en la membrana mitocondrial y colapso de la cantidad de ATP seguido de muerte celular que ocurre por encogimiento celular y fragmentación del DNA. Tanto el arsénico como antimonio requieren hierro para su efecto citotóxico (41). Es plausible, entonces, que un aumento del alelo que implica parada ocasione un menor número de reductasas de hierro funcionales y, por tanto, confiera una mejora biológica en la línea resistente a antimoniales de este estudio.

Gen LinJ.34.4090

Este gen tiene en la posición 1494 de la proteína que codifica un polimorfismo *Figura 30, p.78*. Una de las variantes implica un truncamiento de la proteína, mientras que en la proteína larga esa posición corresponde a un glutámico. Esta posición está hacia el final de la proteína. En la línea resistente a anfotericina se está fijando la base que deriva en la codificación del glutámico. Sin embargo, en la línea resistente a miltefosina se está fijando la base que

implica el truncamiento de la proteína (variante que implica parada Wt: 61,53%, anfotericina: 27,27%, miltefosina: 92,30%).

Este gen codifica para la quinesina tipo Unc-104. No se encuentra ningún estudio que relacione esta proteína en la implicación de resistencias. Para valorar su implicación, deberían realizarse más estudios experimentales al respecto.

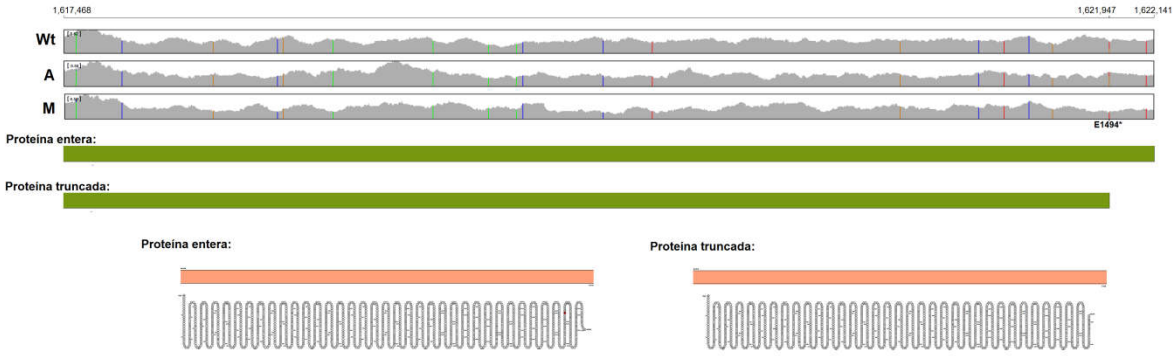

Figura 30. *Esquema de la mutación de parada encontrada en el gen LinJ.34.4090*

Nota. *'A' hace referencia a la línea resistente a anfotericina, 'M' hace referencia a la línea resistente a miltefosina y 'Wt' hace referencia a la línea parental.*

En primer lugar, en la figura se muestra la cobertura del gen en el visualizador IGV donde se marca el SNP encontrado, En segundo lugar, se muestra la relación de longitud de la proteína entera y de la proteína truncada. En tercer lugar, se muestra un esquema de localización de la proteína realizado con el programa Protter donde la parte superior indica el espacio extracelular, la parte salmón la membrana celular y la parte inferior el citosol.

Gen LinJ.28.1730

Este gen codifica para una proteína tipo helicasa y presenta una posición polimórfica que afecta al aminoácido 56 en la proteína. Este aminoácido, puede resultar en una arginina o en el truncamiento de la proteína. La línea WT presenta una mayor proporción del alelo que implica parada de la traducción (Wt: 43,48, antimoniales: 39,02%). De todos modos, la diferencia es muy pequeña y la posición es estadísticamente diferente en antimoniales. Presumiblemente, este SNP es fruto de un error en el método y no se debe tener en cuenta.

Gen LinJ.27.0500

El gen aparece con la misma anotación en varias partes del genoma, presumiblemente, estos genes tienen una identidad de secuencia cercana al 100%.

El gen donde hay la mutación STOP (LinJ.27.0500) es contiguo a otro gen con la misma anotación (LinJ.27.0510). Mirando en IGV se ve que hay una zona considerable en LinJ.27.0500 que tiene mayor cobertura que el resto. En LinJ.27.0510 la cobertura es uniforme (alineamiento Bowtie por defecto, en todas las líneas). Esto, hace pensar que las regiones con más cobertura son multireads de otras calpainas muy similares. Para abordar esto, se visualiza la misma zona con el alineamiento sin multi-reads y, efectivamente, la zona con mayor cobertura es zona con multi-reads. En LinJ.27.0510, hay dos pequeñas zonas con *multi-reads* también.

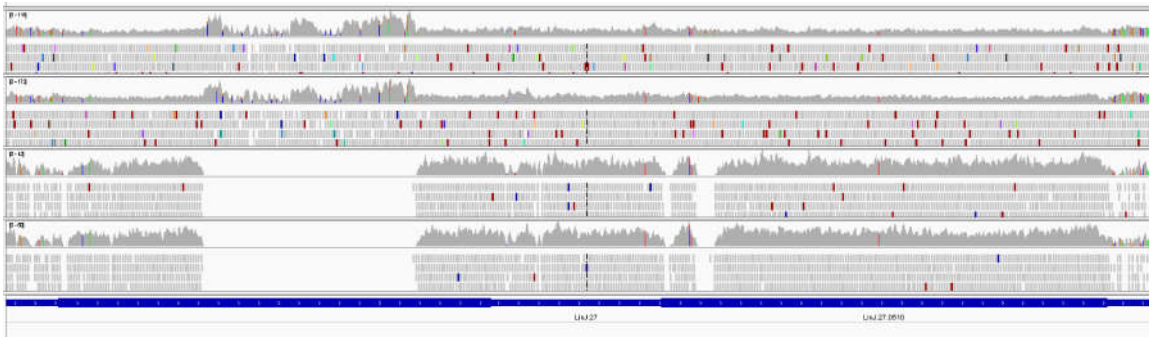

Figura 31. *Visualización de los genes LinJ.27.0500 y LinJ.27.0510 en el visualizador IGV*

Las dos primeras visualizaciones corresponden al alineamiento con multi-reads (WT y línea resistente a la miltefosina, respectivamente), las dos últimas visualizaciones corresponden al alineamiento sin multi-reads.

El SNP de la mutación STOP de LinJ.27.500 cae al inicio de la región con multi-reads, llega a tener cobertura, pero ya es zona de bajada de cobertura considerable en el alineamiento sin multi-reads (cobertura ~6). Por este motivo, no se puede considerar que este SNP sea real.

4. Conclusiones

Este TFM me ha permitido desarrollar a nivel práctico muchas de las áreas trabajadas en el Máster de Bioinformática y Bioestadística de la UOC. En este sentido, considero que el proyecto escogido ha resultado ser muy completo y me ha ayudado a ganar experiencia a nivel profesional. Quiero dar las gracias a mi tutor del máster, Álex Sanchez, por haber permitido que realizase este proyecto *ad-hoc* como TFM y a mi consultora del TFM, Ivette Olivares, así como al Dr. Jose María Requena, mi supervisor científico en el CBMSO, por haber tutelado y revisado mi trabajo. Sin sus aportaciones, la calidad del TFM no hubiera sido la misma.

Al empezar el TFM, era consciente de que el proyecto elegido era ambicioso teniendo en cuenta las limitaciones de tiempo marcadas por el calendario. Sin embargo, la planificación temporal del mismo, permitió una buena optimización del tiempo, de manera que, finalmente, no solo se lograron los objetivos marcados sino que estos se ampliaron, realizando una tabla con las implicaciones a nivel proteico de los SNPs detectados en ORFs.

Manejando esta gran cantidad de datos, uno se da cuenta de lo importante que es realizar un buen diseño experimental, que muchas veces hay que ir optimizando sobre la marcha según el comportamiento de los datos obtenidos. En este sentido, la experiencia hace mucho, sobre todo a la hora de ganar tiempo en el procesado de los datos.

Estoy contenta con los resultados obtenidos en el TFM porque se han descubierto alteraciones en el genoma de nuestras líneas de estudio, y algunas de estas se han asociado a implicaciones biológicas de las resistencias a los fármacos. Además, se han obtenido muchos datos, como por ejemplo, SNPs en ORFs que implican mutaciones sinónimas o de cambio aminoacídico, cuya implicación biológica no se ha podido analizar en profundidad debido a la falta de tiempo y al complejo diseño de análisis para poder validarlas. Por otro lado, muchas de las hipótesis de las alteraciones encontradas relacionadas con la adquisición de resistencias a fármacos, deben ser validadas experimentalmente. Así, el trabajo de análisis computacional va de la mano con el experimental, combinando ambos, se pueden llegar a obtener hipótesis más sólidas.

Este trabajo ha abierto muchas vías de análisis de cara al futuro, pasando por la validación de todas las alteraciones encontradas que no se han podido analizar en profundidad en este TFM, así como la realización de protocolos parecidos para otras líneas de estudio.

He aprendido y disfrutado mucho durante el desarrollo de este proyecto. Me ha ayudado a ganar experiencia y soltura al manejar grandes cantidades de datos. Espero haber conseguido plasmar tanto el desarrollo del análisis como los frutos obtenidos de este y que la lectura de la memoria escrita haya resultado amena e interesante.

5. Glosario

SNP- del inglés *Single Nucleotide Polymorphism*.

CNV- del inglés *Copy Number Variation.*

ORF- del inglés *Open Reading Frame.*

TFM- Trabajo de Fin de Máster.

UOC- del catalán *Universitat Oberta de Catalunya.*

CBMSO- Centro de Biología Molecular Severo Ochoa.

6. Bibliografía

1. *Molecular mechanisms of antimony resistance in Leishmania.* Ashutosh, Sundar S., Goyal N. 143-153, : Journal of Medical Microbiology, 2007, Vol. 56.
2. *Spread of vector-borne diseases and neglect of Leishmaniasis, Europe.* Dujardin, J.C., Campino, L., Cañavate, C., Dedet, J.P., Gradoni, L., Soteriadou, K., Mazeris, A.,Ozbel, Y., Boelaert, M. 1013-1018, : Emerg Infect Dis, 2008, Vol. 14.
3. *Immune Regulation during Chronic Visceral Leishmaniasis.* Faleiro, R.J., Kumar R., Hafner, L.M., Engwerda, C.R. 7, : PLOS, Neglected Tropical Diseases, 2014, Vol. 8.
4. *Leishmaniasis: vaccine candidates and perspectives.* Singh, B., Sundar, S. 3834-3842, : Vaccine, 2012, Vol. 30.
5. *Leishmaniasis: drugs in the clinic, resistance and new developments.* Ouellette, M., Drummelsmith, J., Papadopoulou, B. 257-266, : Drug Resist Updat, 2004, Vol. 7.
6. *Drug resistance analysis by next generation sequencing in Leishmania.* Leprohon, P., Fernandez-Prada, C., Gazanion, É., Monte-Neto, R., Ouellette, M. 1, : Int J Parasitol Drugs Drug Resist, 2015, Vol. 5.
7. *Plasticity of the Leishmania genome leading to gene copy number variations and drug resistance.* Laffitte, M.C.N., Leprohon, P. , Papadopoulou, B., Ouellette M. 2350, : F1000Research, 2016, Vol. 5.
8. *Drug resistance in leishmaniasis.* Croft, S.L., Sundar, S., and Fairlamb, A.H. 111-126, : Clin Microbiol Rev, 2006, Vol. 19.
9. *Leishmaniasis: current status of available drugs and new potential drug targets.* Singh, N., Kumar, M., and Singh, R.K. 485-497, : Asian Pacific journal of tropical medicine, 2012, Vol. 5.
10. *Therapeutic options for visceral leishmaniasis.* Monge-Maillo, B., and Lopez-Velez, R. 1863-1888, : Drugs, 2013, Vol. 73.
11. *Recent developments and future prospects in the treatment of visceral leishmaniasis.* Sundar, S., and Singh, A. 98-109, : Ther Adv Infect Dis, 2016, Vol. 3.
12. *Mechanism of amphotericin B resistance in Leishmania donovani promastigotes.* Mbongo, N., Loiseau, P.M., Billion, M.A., and Robert-Gero, M. 352-357, : Antimicrob Agents Chemother, 1998, Vol. 42.
13. *Characterisation of Leishmania donovani promastigotes resistant to hexadecylphosphocholine (miltefosine).* Seifert, K., Matu, S., Javier Perez-Victoria, F., Castanys, S., Gamarro, F., and Croft, S.L. 380-387, : International journal of antimicrobial agents, 2003, Vol. 22.
14. *Phospholipid translocation and miltefosine potency require both L. donovani miltefosine transporter and the new protein LdRos3 in Leishmania parasites.* Perez-Victoria, F.J., Sanchez-Cañete, M.P., Castanys, S., and Gamarro, F. 23766-23775, : J Biol Chem, 2006, Vol. 281.
15. *Development and characterization of paromomycin-resistant Leishmania donovani promastigotes.* Maarouf, M., Adeline, M.T., Solignac, M., Vautrin, D., and Robert-Gero, M. 5, : Parasite, 1998. 167-173.
16. *Paromomycin: uptake and resistance in Leishmania donovani.* Jhingran, A., Chawla, B., Saxena, S., Barrett, M.P., and Madhubala, R. 111-117, : Mol Biochem Parasitol, 2009, Vol. 164.

17. *Mechanism of amphotericin B resistance in clinical isolates of Leishmania donovani.* Purkait, B., Kumar, A., Nandi, N., Sardar, A.H., Das, S., Kumar, S., Pandey, K., Ravidas, V., Kumar, M., De, T., et al. 1031-1041, : Antimicrob Agents Chemother, 2012, Vol. 56.
18. *Up-regulation of silent information regulator 2 (Sir2) is associated with amphotericin B resistance in clinical isolates of Leishmania donovani.* Purkait, B., Singh, R., Wasnik, K., Das, S., Kumar, A., Paine, M., Dikhit, M., Singh, D., Sardar, A.H., Ghosh, A.K., et al. 1343-1356, : J Antimicrob Chemother, 2015, Vol. 70.
19. VarScan. [Online] [Cited: 04 01, 2017.] http://varscan.sourceforge.net/using-varscan.html.
20. *Comparative genomic analysis of three Leishmania species that cause diverse human disease.* Peacock, C.S., Seeger, K., Harris, D.,Murphy,L., Ruiz, J.C., Quail,M.A., Peters, N., Adlem, E., Tivey, A., Aslett, M., et al. 839-847, : Nat Genet, 2000, Vol. 39.
21. TriTrypDB Kinetiplastid Genomics Resource. [Online] [Cited: 04 01, 2017.] http://tritrypdb.org/tritrypdb/.
22. Bowtie2. [Online] [Cited: 04 01, 2017.] http://bowtie-bio.sourceforge.net/bowtie2/index.shtml.
23. SAMtools. [Online] http://www.htslib.org/doc/samtools.html.
24. Integrative Genomics Viewer. [Online] [Cited: 04 01, 2017.] http://software.broadinstitute.org/software/igv/.
25. *Chromosome and gene copy number variation allow major structural change between species and strains of Leishmania.* Rogers, M. B., Hilley, J. D., Dickens, N. J., Wilkes, J., Bates, P. A., Depledge, D. P. , Harris, D., Her, Y., Herzyk, P., Imamura, H., Otto, T. D., Sanders, M., Seeger, K., Dujardin, J. C., Berriman, M., Smith, D.F., Hertz-Fowler, C. and Mottram, J. C. 2129-2142, : Genome Res, 2011, Vol. 21.
26. *Python tutorial.* van Rossum, G. Amsterdam : Centrum voor Wiskunde en Informatica (CWI), 1995, Vols. Technical Report CS-R9526.
27. Tammi Lab. [Online] [Cited: 04 01, 2017.] http://tiger.dbs.nus.edu.sg/cnv-seq/.
28. *Novel insights into genome plasticity in Eukaryotes: mosaic aneuploidy in Leishmania.* Sterkers, Y., Lachaud, L. , Bourgeois, N., Crobu, L., Bastien, P., Pagès, M. 15-23, : Molecular Microbiology, 2012, Vol. 86.
29. *Response of gene expression in Saccharomyces cerevisiae to amphotericin B and nystatin measured by microarrays.* Zhang, L., Zhang, Y., Zhou, Y.M., An, S., Zhou, Y.X., Cheng, J. 905-915, : Journal of Antimicrobial Chemotherapy, 2002, Vol. 49 (6).
30. *Altered expression, polymerisation and cellular distribution of α-/β-tubulins and apoptosis-like cell death in arsenite resistant Leishmania donovani promastigotes.* Jayanarayan, K.G., Dey, C.S. 915-925, : s.n., 2004, Vol. 34 (8).
31. *Flavone-resistant Leishmania donovani Overexpresses LdMRP2 Transporter in the Parasite and Activates Host MRP2 on Macrophages to Circumvent the Flavone-mediated Cell Death.* Chowdhury, S., Mukhopadhyay, R., Saha, S., Mishra, A., Sengupta, S., Roy, S., Majumder, H. K. 646-651, : s.n., 2014, Vol. 440 (4).
32. *Folate metabolic pathways in Leishmania.* Vickers, T.J., Beverley, S.M. 63-80, : Essays in Biochemistry, 2011, Vol. 51.

33. *CYP5122A1, a novel cytochrome P450 is essential for survival of Leishmania donovani.* Verma, S., Mehta, A., Shaha, C. : PLoS ONE, 2011, Vol. 6(9).
34. *plotly: Create Interactive Web Graphics via 'plotly.js'.* Sievert, C., Parmer, C., Hocking, T., Chamberlain, S., Ram, K., Corvellec, M., Despouy, P. https://CRAN.R-project.org/package=plotly, 2016, Vol. R package version 4.5.6.
35. UGent, Bioinformatics & Evolutionary Genomics Group. [Online] [Cited: 04 27, 2017.] http://bioinformatics.psb.ugent.be/webtools/Venn/.
36. *Increased transport of pteridines compensates for mutations in the high affinity folate transporter and contributes to methotrexate resistance in the protozoan parasite Leishmania tarentolae.* Kündig, C., Haimeur, A., Légaré, D., Papadopoulou, B., Ouellete, M. 2342-2351, : EMBO Journal, 1999, Vol. 18 (9).
37. *Alteration of fatty acid and sterol metabolism in miltefosine-resistant Leishmania donovani promastigotes and consequences for drug-membrane interactions.* Rakotomanga, M., Saint-Pierre-Chazalet, M., Loiseau, P.M. 2677-2686, : Antimicrob Agents Chemother, 2005, Vol. 49.
38. Protter. [Online] [Cited: 05 08, 2017.] http://wlab.ethz.ch/protter/start/.
39. *ABC transporters in Leishmania and their role in drug resistance.* Ouellette, M., Légaré, D., Haimeur, A., Grondin, K., Roy, G., Brochu, C., Papadopoulou, B. 43-8, : Drug Resistance Updates: Reviews and Commentaries in ntimicrobial and Anticancer Chemotherapy, 1998, Vol. 1(1).
40. *Functional Cloning of the Miltefosine Transporter A NOVEL P-TYPE PHOSPHOLIPID TRANSLOCASE FROM LEISHMANIA INVOLVED IN DRUG RESISTANCE.* Pérez-Victoria, F.J., Gamarro, F., Ouellette, M., Castanys, S. 49965-49971, : The Journal of Biological Chemistry, 2003, Vol. 278.
41. *Antimony resistance in Leishmania focusing on experimental research.* Jeddi, F., Piarroux, R., Mary, C. : Journal of Tropical Medicine, 2011, Vol. 2011.
42. *Soybean Cyst Nematode Resistance.* Wu, X.Y., Zhou, G.C., Chen, Y.X., Wu, P., Liu, L.W., Ma, F.F., Wu M., Liu, C.C., Zeng, Y.J., Chu, A.E., Hang, Y.Y., Chen J.Q., Wang, B. 998, : Frontiers in Plant Science, 2016, Vol. 7.
43. *Drug resistance analysis by next generation sequencing in Leishmania.* Leprohon, P., Fernandez-Prada, C., Gazanion, É., Monte-Neto, R., Ouellette, M. 26–35, : International Journal for Parasitology: Drugs and Drug Resistance, 2015, Vol. 5(1).

7. Anexos

Debido a la gran extensión del anexo y para evitar que la lectura de la memoria sea dificultosa, los anexos de este trabajo se han adjuntado en un documento aparte con el consentimiento y visto bueno de la coordinadora del TFM Ivette Olivares Castiñeira.