



UNIVERSIDAD DE TALCA
FACULTAD DE INGENIERÍA
ESCUELA DE INGENIERÍA CIVIL EN BIOINFORMÁTICA

**Ensamble y Anotación del Genoma de la Rata de los
Pinares menor (*Aconaemys sagei*)**

CRISTOPHER DANIEL FIERRO RIQUELME

Profesor Tutor: DR. GONZALO RIADI MAHÍAS

Profesor Informante: DR. BRAULIO VALDEBENITO MATURANA

Memoria para optar al título de Ingeniero Civil en Bioinformática.

Talca – Chile

Junio 19, 2020

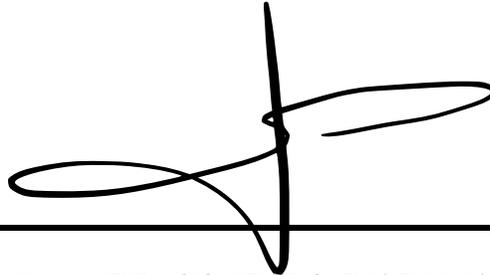
CONSTANCIA

La Dirección del Sistema de Bibliotecas a través de su unidad de procesos técnicos certifica que el autor del siguiente trabajo de titulación ha firmado su autorización para la reproducción en forma total o parcial e ilimitada del mismo.



Talca, 2021

FIRMAS COMISIÓN EVALUADORA

A handwritten signature in black ink, consisting of a vertical line with a horizontal stroke crossing it, and a large loop on the left side.

Profesor Tutor: DR. GONZALO RIADI MAHIAS

A handwritten signature in black ink, featuring a large, stylized 'B' and 'M' with a horizontal line through them, and a small dot at the end.

Profesor Informante: DR. BRAULIO FAVIÁN VALDEBENITO MATURANA

Esta tesis esta dedicada a:

La familia que me dió las alas y que me ayudó a volar.

A la familia que me acogió en su nido y hoy volamos juntos hacia un futuro mejor

AGRADECIMIENTOS

En primera instancia agradezco a quienes influyeron en mi formación profesional: a mi tutor Gonzalo Riadi por su paciencia y confianza depositadas durante el desarrollo de este proyecto. A Braulio, mi informante, quien siempre tuvo la disposición de enseñarme cosas nuevas en esta área de la bioinformática.

Agradezco el esfuerzo, dedicación, paciencia y confianza que con mucho cariño mis padres, Daniel y Susana, depositaron en mi; dándome el apoyo y las fuerzas para aventurarme en algo que un principio nos era desconocido.

Doy gracias también a tres pequeños motores que hoy en día enriquecen mi mundo: Tomás, Valentina y Bastián. Sus palabras y acciones siempre quedarán en mi corazón, por siempre son y serán el impulso que me motiva en ser una mejor persona, un hermano o una figura paterna.

Agradezco a mis amigos de universidad, Álvaro y Erwin, con quienes compartí muy gratos momentos durante mi formación académica. Doy gracias también a los padres de ambos, quienes fueron una fuente de sabiduría inagotable sobre lo que es vivir y que también me acogieron como uno más de su familia.

Quiero también agradecer a mi amigo Bastián quien tuvo la paciencia para estar conmigo toda una mañana cuando descubrí que era la bioinformática, dándome las primeras palabras de aliento para embarcarme en esto.

Y por sobre todo agradecer a Yetzabel por siempre estar ahí y ser ella tal como es: sincera, amable y bondadosa; no existe mejor compañía en mi mundo, y pasó a ser la persona quien admiro y amo con todas mis fuerzas. Agradezco la fortaleza que tuvo para contenerme en los peores momentos y la determinación para convertir nuestra vida en los mejores momentos.

De todas estas personas aprendí, que por mucho que vuele lejos de aquí, seguirán siempre junto a mí.

TABLA DE CONTENIDOS

	página
Firmas Comisión Evaluadora	I
Dedicatoria	II
Agradecimientos	III
Tabla de Contenidos	IV
Índice de Figuras	VI
Índice de Tablas	VII
Resumen	VIII
Abstract	IX
1. Superfamilia de roedores <i>Octodontoidea</i>	1
1.1. Colonización de ambientes	1
2. Contexto histórico de secuenciación en especies	4
3. ¿Que es el ensamble de un genoma?	7
3.1. Ensamble <i>de novo</i>	7
3.2. Ensamble por referencia	9
4. Evaluación del ensamble	13
5. Anotación de genomas	15
5.1. Importancia de la anotación	17
6. Anotación y Ensamble en organismos de la superfamilia <i>Octodontoidea</i>	19
7. Problema	21
8. Solución propuesta	22
9. Objetivos	23
9.1. Objetivo General	23

9.2. Objetivos Específicos	23
10. Metodología	24
10.1. Elección de modelo de organismo por referencia.	25
10.2. Ensamble por referencia	25
10.3. Ensamble de transcritos	27
10.4. Anotación de genomas	27
10.5. Post-procesamiento de la anotación	29
11. Resultados	30
12. Discusión	39
13. Conclusión	42
14. Anexos	51
14.1. Anexo 1.- Comandos de ejecución de programas	51
14.1.1. FASTQC	51
14.1.2. TRINITY	51
14.2. Anexo 2.- Script de ejecución de Trimmomatic	52
14.3. Anexo 3.- Comandos de ensamble por referencia	53
14.4. Anexo 4.- Archivo de configuración maker_exe.ctf	55
14.5. Anexo 5.- Archivo de configuración maker_bopts.ctf	56
14.6. Anexo 6.- Archivo de configuración maker_opts.ctf	58
14.7. Anexo 7.- Comandos anotación funcional	62
14.8. Anexo 8.- Información preliminar de <i>Octodon degus</i>	65
14.9. Anexo 9.- Análisis BUSCO en <i>Octodon degus</i>	65

ÍNDICE DE FIGURAS

página

1.1. Disposición teórica de filogenia en organismos de la superfamilia Octodontoidea a lo largo del tiempo.	2
2.1. Cantidad de genomas ensamblados y depositados en NCBI desde 1988 hasta la actualidad.	6
3.1. Esquema del ensamble de novo.	10
3.2. Tipos de cobertura.	11
3.3. Esquema del ensamble por referencia.	12
5.1. Anotación de genomas.	16
10.1. Trimming de reads.	25
10.2. Ensamble por referencia.	26
10.3. Pipeline de anotación MAKER	29
11.1. Resultados de calidad posterior al trimming.	31
11.2. Ausencia de adaptadores después del trimming.	32
11.3. Análisis de cobertura.	33
11.4. Análisis de calidad de ensamble.	34
11.5. Distribución de acumulada de largo de scaffolds en ensamble por referencia.	35
11.6. Resultados BUSCO del ensamble de transcritos de <i>A. sagei</i> en Trinity.	36
11.7. Diagramas de Venn en registros de elementos presentes en la anotación.	37
11.8. Resultados BUSCO del ensamble de transcritos de <i>Aconaemys sagei</i> en MAKER.	38

ÍNDICE DE TABLAS

	página
6.1. Detalle de organismos Octodontidae ensamblados.	20
11.1. Detalle de librerías de secuenciación Illumina HiSeq 2500 en muestras de <i>A.sagei</i>	30
11.2. Resultados de alineamiento en candidatos a modelo de ensamble por referencia.	32
11.3. Resumen de anotación del organismo <i>Aconaemys sagei</i>	37

RESUMEN

Los registros genómicos que hoy en día encontramos en las bases de datos biológicas son una fuente extensa de recursos que muchos investigadores ocupan en el estudio evolutivo de distintas especies. Esta información no existiría sin el desarrollo de tecnologías de secuenciación y el avance en paralelo de metodologías de ensamble de secuencia. Del mismo modo, el registro y descripción de los elementos genéticos resulta ser un paso crucial en la investigación de especies unidas por un ancestro común.

La superfamilia Octodontoidea, perteneciente al infraorden Caviomorpha, es una familia de roedores endémica de Sudamérica. A lo largo de su historia evolutiva, ciertas especies pertenecientes a este grupo ha desarrollado características determinantes para la colonización en el ambiente subterráneo. Esto ha implicado cambios a nivel morfológico para su adaptación al subsuelo, por lo que se especula que dichas especies muestran variaciones a nivel genético en comparación con sus especies más cercanas que habitan en la superficie. Muchos de los rasgos genéticos son casi imposibles de comparar hoy en día debido a la escasez de información disponible en bases de datos biológicas sobre este grupo de roedores.

Con el fin de aportar en la falta de información genética que pueda generar estudios evolutivos entre dichas especies, se llevó a cabo el ensamble por referencia de la especie *Aconaemys sagei*, conocida comúnmente como rata de los pinares menor, así como una primera anotación estructural y funcional de dicho genoma.

ABSTRACT

The genomic records present in biological databases are an extensive source from where many researchers study the evolution of the species. The development in sequencing technologies, new assembly methods along with the mapping, annotation and analysis of genetic features have allowed the study of species with a common ancestor.

The Octodontoidea is a superfamily of the caviomorph infraorder endemic to south america, throughout the ages certain rodents of this clade have developed crucial features for the colonisation on the underground environment. the adaptation process has increased the genetic differentiation compared to the closest relative in the surface, rendering impossible to compare due to the lack of information available in databases regarding this rodents.

In order to close the knowledge gap required to develop evolutionary studies between those species, a reference assembly of *Aconaemys sagei*, also known as “Sage’s Rock Rat”, was made, as well as the first structural and functional annotation of the genome.

1. Superfamilia de roedores *Octodontoidea*

Los roedores del infraorden *Caviomorpha*, pertenecientes a la orden *Rodentia*, son uno de los grupos de mamíferos con mayor número de linajes y alta capacidad de reproducción conocidos en la tierra. Este linaje es un grupo anatómicamente variable con una marcada diversidad de tipos morfológicos y ecológicos [Elissamburu and Vizcaíno, 2004] y se compone de cuatro superfamilias principales: *Cavioidea*, *Chinchilloidea*, *Erethizontoidea* y *Octodontoidea*; donde esta última es considerada endémica de Sudamérica, localizada desde Perú al sur de Brasil, llegando hasta Tierra del Fuego y abarcando también zonas de Argentina, Bolivia, Paraguay y Uruguay. Ciertas especies de la superfamilia *Octodontoidea* han adquirido características determinantes para su adaptación en ambiente subterráneo, mostrando diferencias a niveles genéticos, bioquímicos, anatómicos y social con respecto a especies más cercanas habitantes de la superficie [Nevo, 1979] [Ruiz, 2011].

1.1. Colonización de ambientes

El término radiación adaptativa se conoce como el proceso en el cual una especie se inserta en un ecosistema nuevo y procede a través de una rápida especiación, es decir, la aparición de rasgos que diferencian dos especies próximas; con el fin de llenar muchos nichos ecológicos existentes en esa zona, lo que explica que especies ancestrales se diversificaran y divergieran en las especies que conocemos hoy en día [Losos, 2010]. La superfamilia *Octodontidae* es un linaje que surgió entre las épocas desde Oligoceno y el Mioceno con el propósito de ocupar nichos herbívoros a los cuales logró adaptarse durante radiación adaptativa. Esta superfamilia contiene a las familias *Abrocomidae*, *Octodontidae*, *Echymyidae* y *Ctenomyidae*. Sobre estas familias, la invasión dirigida al

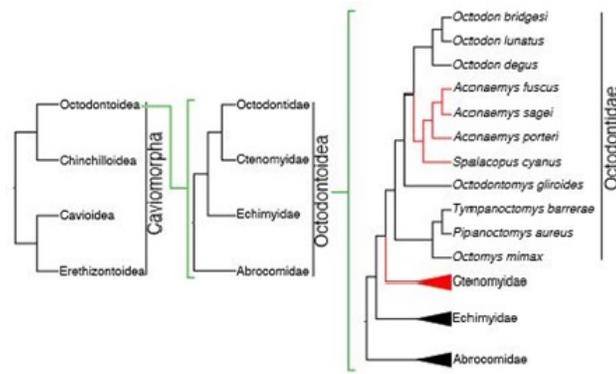


Figura 1.1: Disposición teórica de filogenia en organismos de la superfamilia Octodontoidea a lo largo del tiempo.

Los clados coloreados en rojo corresponden a especies colonizadoras del espacio subterráneo. La imagen representa a las familias contenidas dentro de la superfamilia *Octodontoidea*, mostrando en detalle a la familia *Octodontidae*.

ambiente subterráneo ocurrió de manera independiente en dos ocasiones a lo largo del tiempo. La primera colonización al ambiente subterráneo ocurre alrededor de los 18 y 9 millones de años atrás con la familia *Ctenomyidae* (más conocidos como tucos-tucos), la cual posee alrededor de 60 especies reconocidas que habitan dicho entorno cuyas condiciones filogenéticas han sido ampliamente estudiadas. Por otro lado, existe un clado compuesto por las especies *Spalacopus cyanus* (chululo o cururo), *Aconaemys fuscus* (tunduco común), *A. sagei* (rata de los pinares menor) y *A. porteri* (rata de los pinares o tunduco de Porter), especies de la familia *Octodontidae*, que también invadieron el ambiente subterráneo en un plazo entre 2.8 y 3.7 millones de años atrás. Hasta el presente, existen hipótesis filogenéticas basadas en las características de estos organismos (ver Figura 1.1) que permiten ver la historia evolutiva de la superfamilia, sin embargo, a pesar de estas conjeturas, no existen pruebas genéticas capaces de afirmar este modelo.

En los últimos años se han dedicado varios estudios orientados a determinar la divergencia morfológica con respecto a la anatomía de diversos organismos pertenecientes a la superfamilia *Octodontoidea*, haciendo énfasis en su estructura craneodental y postcranial. Estos estudios han logrado demostrar que uno de los primeros pasos en la divergencia evolutiva morfológica de estas especies ha sido la adaptación del ancho mandibular [Lessa et al., 2008]. Además, diversos rasgos craneanos indican su especialización en la dentoexcavación al adaptar incisivos robustos y aumentar las dimensiones de músculos masetéricos [Morgan and Verzi, 2011, Morgan and Verzi, 2006]. Las modificaciones reconocidas en los labios de las especies subterráneas ayudan a que no entre tierra en su boca durante la excavación. Por otro lado, se ha constatado que la diversificación

en el cariotipo de estas especies varía de $2n = 10$ a $2n = 70$ cromosomas en total; la evidencia experimental demuestra estas diferencias basándose en la masa promedio medida en picogramos de DNA, donde el largo del genoma promedio de los roedores subterráneos es cercano a los $4.18 \text{ pg} \pm 0.92$ y aquellos no subterráneos con $3.4 \text{ pg} \pm 0.71$, evidencia que puede ser interpretada como una mayor longitud en genomas de roedores subterráneos en comparación con no subterráneos. Las variaciones en los cromosoma de estas especies han sido propuestas como las causales del mecanismo de especiación [Parededa and Novello, 2012].

La adaptación al subsuelo también propone características beneficiosas para la subsistencia de especies subterráneas a través de la resistencia a diversas complicaciones que conlleva el hábitat subterráneo como la capacidad de soportar bajos niveles de oxígeno y altas concentraciones de CO_2 (hipoxia e hipercapnia) [Tomasco and Lessa, 2010]. Además, muchas especies que logran colonizar los ambientes subterráneos generan una excepcional resistencia al cáncer [Manov et al., 2013] y una alta tolerancia a enfermedades referentes a la vejez [Novikov et al., 2015].

Hasta la fecha, es escasa la evidencia que permitan realizar estudios de análisis comparativos en genomas de la superfamilia *Octodontoidea*. Sólo unos pocos genomas han sido secuenciados, ensamblados y anotados, asimismo, de algunos de ellos sólo se conoce su secuencia mitocondrial, esto impide la realización de estudios enfocados en estas especies que detallen diferencias a nivel genético. Para resolver las implicaciones genotípicas que puedan verse inmersas en las investigaciones sobre estas especies, es necesaria la realización de la secuenciación, ensamble y anotación del genoma de los organismos pertenecientes a esta superfamilia.

2. Contexto histórico de secuenciación en especies

A partir del modelo de doble hélice del DNA resuelto por Watson y Crick en el año 1953 [Watson and Crick, 1953], se obtuvieron las principales inferencias sobre el procedimiento de replicación, transcripción y traducción que hoy en día conforman el dogma central de la biología molecular. Con la información obtenida con el pasar de los años, se presentaba cada vez más la necesidad de identificar la composición nucleotídica, y por, sobre todo, el orden en que estas se encuentran para la producción de proteínas. Debido a la falta de tecnología y capacidad de aquella época, fue menester el desarrollo de nuevas tácticas que pudieran resolver dicho problema en un sistema más acotado similar al DNA, por lo que se optó por la investigación en RNA.

En el año 1966 el equipo de Fred Sanger dio el primer paso a la estrategia de detección de fragmentos de secuencias de carácter nucleotídico digeridos de manera parcial. Esta estrategia tendría por objetivo la investigación e identificación de secuencias cortas de RNA ribosomal y de transferencia. Posteriormente, en el año 1972, el equipo de Walter Fiers [Fiers et al., 1976] fue el encargado de secuenciar las primeras secuencias de genes codificantes a proteínas de cubierta viral en el bacteriófago MS2 utilizando dicho método. Cuatro años después, se dio paso a la completa secuenciación del genoma viral MS2. Es a partir de este tipo de estrategias que una gran cantidad de investigadores de la época resolvieron en desarrollar nuevos métodos de secuenciación de secuencia nucleotídicas [Brownlee and Sanger, 1967], abriendo paso a las tecnologías de secuenciación de primera generación y consolidando un objetivo cuya investigación sigue desarrollándose y mejorando hasta el día de hoy [Mullis and Faloona, 1987].

Sin embargo, la totalidad de estas estrategias que en un principio surgieron para suplir esta necesidad no eran capaces de responder al cuestionamiento del orden en que estas

bases nucleotídicas se encontraban en el genoma del organismo objetivo. Los intentos de aquella época para la determinación real de las bases aún estaban sujetos a su identificación en tramos cortos de secuencias de DNA y demandaba una gran cantidad de tiempo y esfuerzo invertido en técnicas de química analítica y fraccionamiento. Además, la lectura única del genoma completo resulta imposible sólo a través de la secuenciación, puesto a que esta última utiliza técnicas de amplificación del DNA lo cual dificulta la resolución de la estructura final debido a la obtención de tantas copias que cubren el genoma. Bajo este esquema se comenzaron a desarrollar de forma paralela y casi sinérgica a las técnicas de secuenciación, las bases de los algoritmos y técnicas de ensamble que se ocupan hoy en día [Heather and Chain, 2016](#).

Los algoritmos de ensamblado de genomas surgen de la necesidad de identificar el orden específico de cada uno de las pares de bases de un organismo, bajo la condición de utilizar cantidades masivas de lecturas de corto tamaño. Estos algoritmos han estado en desarrollo gracias al avance de las nuevas técnicas de secuenciación de siguiente generación (*Next Generation Sequencing*), las cuales han dado el paso a la resolución de problemáticas referidas a la estructura del genoma. A diferencia de las tecnologías de primera generación de Sanger, cuyo enfoque era secuenciar el genoma fragmentándolo en múltiples secuencias cortas de nucleótidos extraídas de zonas al azar; las tecnologías de segunda generación lograron implementar un método en el cual dos secuencias cortas de nucleótidos son separadas por un tramo de pares de bases desconocido, pero de longitud definida, a este tramo se le denomina tamaño de inserto. [Bradnam et al., 2013](#), [Lischer and Shimizu, 2017](#).

La empresa estadounidense Illumina ha desarrollado una gran cantidad de protocolos de NGS, entre ellos el método de secuenciación por síntesis (SBS). Este método se basa en fragmentar el genoma en múltiples segmentos de DNA, la adición de adaptadores permite que los pequeños trazos de material genético se adhieran a un flowcell cuya superficie se encuentra llena de oligonucleótidos complementarios a los adaptadores. Al ser fijados en el flowcell, las polimerasas se encargan de generar la hebra molde restante de las secuencias adheridas con el uso de nucleótidos fluorescentes; estos nucleótidos emiten una señal que al ser captada nos permite determinar el orden de los nucleótidos en un tramo denominado read o lectura. Un paso adicional de bioetilinización en los extremos de los trazos de DNA permite generar pares de lecturas en base a fragmentos de mayor tamaño y, por ende, con un tamaño de inserto superior [Bronner et al., 2014](#), [Yoshinaga et al., 2018](#), [Zhang et al., 2011](#). El resultado de este proceso es una base de datos de lecturas de secuenciación de tamaño corto donde Idealmente las lecturas apareadas por un tamaño de inserto, a dicho par se le denomina fragmento.

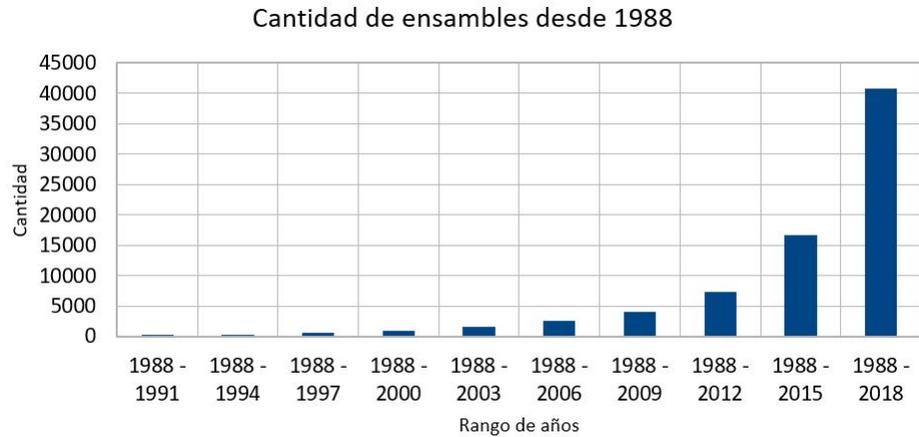


Figura 2.1: Cantidad de genomas ensamblados y depositados en NCBI desde 1988 hasta la actualidad.

La cantidad de genomas corresponde a la totalidad de los organismos, incluyendo eucariotas, procariotas, archaeas y virus.

La cantidad de genomas ensamblados ha crecido de manera exponencial en la última década (Figura 2.1), este crecimiento ha sido gracias al continuo avance de las tecnologías de secuenciación y al surgimiento de las denominadas tecnologías de secuenciación NGS.

El continuo desarrollo de las técnicas NGS ha generado gran interés en estudiar organismos con un mayor grado de complejidad en su distribución de nucleótidos, siendo de esta manera un aporte en la mapeabilidad de las lecturas de secuenciación y en la construcción de genomas de tamaño extenso. Es de este modo, que se ha logrado formalizar el ensamblaje de genomas como una técnica necesaria para el estudio del genotipo y de los múltiples factores inmiscuidos en el genoma del organismo.

3. ¿Que es el ensamble de un genoma?

El ensamble de secuencias es un proceso técnico donde se realiza una reconstrucción o modelado del genoma de un organismo objetivo, el cual puede ser considerado como su mapa físico describiendo al genoma a nivel de pares de bases. Este modelo es mapeado y sintetizado en una base de datos de secuencias o lecturas obtenidas a partir del procedimiento experimental denominado secuenciación. El proceso surge de la necesidad de resolver el orden de las pares de bases en la secuencia de ADN que permiten el funcionamiento de los organismos vivos y la estructura del mismo [Khan et al., 2018].

El objetivo técnico del ensamblado de genomas es el de armar a partir de una gran cantidad de secuencias cortas denominadas lecturas o reads para generar secuencias contiguas de mayor tamaño (contigs) [Khan et al., 2018], con el fin ideal de obtener de manera completa la secuencia de nucleótidos contenida por cada cromosoma del organismo. La generación de este modelo puede ser llevada a cabo mediante las técnicas de novo y por referencia. Para la obtención de la secuencia genómica de un organismo no ensamblado (esto es, que no existe un genoma publicado) existen dos tipos de metodologías que se explicarán a continuación.

3.1. Ensamble *de novo*

El ensamble de novo es una técnica que permite la generación de las secuencias de un genoma sin el uso de una referencia, usando como método el ordenamiento de lecturas de DNA secuenciadas o reads. Como resultado de esto se espera un set de secuencias nucleotídicas que representen el genoma del organismo de interés.

Los tipos de reads que pueden ser utilizados en el ensamble depende tanto de la cantidad en que estos se ocupen como la distancia entre ellos; para esto es necesaria la distinción entre Single-end y Paired-end reads. Los Single-end reads son aquellas secuencias nucleótidas individuales que provienen de la secuenciación de sólo un extremo de un fragmento de DNA, mientras que los Paired-end reads son secuencias “emparejadas”, ya que provienen de la secuenciación de los dos extremos de un mismo fragmento de DNA. Los Paired-end reads corresponden a dos secuencias nucleotídicas de similar tamaño en pares de bases que están distanciadas por una cantidad de pares de bases específicas ya conocidas; esta técnica permite abordar la solución a problemáticas como la construcción de zonas repetidas dentro de un genoma [Miller et al., 2010](#).

La construcción de un modelo de genoma se basa, en primera instancia, en el ordenamiento y alineamiento de reads mediante el traslape entre ellos. Algunos métodos aprovechan el traslape de reads con un corto tamaño de inserto para generar fragmentos de mayor tamaño. Otros métodos en cambio dan uso de técnicas matemáticas para la elaboración de secuencias de mayor tamaño a partir de los fragmentos de lecturas, tal es el caso del uso de ensambladores basados en el grafo De Bruijn (DBG). En este método, ambas secuencias del fragmento son cortadas en subsecuencias contiguas de tamaño k denominadas K -mers. Debido a la gran cantidad de secuencias que son fragmentadas, los K -mers son dispuestos en un DBG siendo ordenados a partir del traslape de $k-1$ nucleótidos. De este modo, el ensamblador obtiene la secuencia recorriendo el nodo o las aristas que presentan más evidencia o peso dentro del grafo. Muchos ensambladores de novo difieren entre sí por el uso de un DBG hamiltoniano o euleriano, siendo su principal diferencia la ubicación de la subsecuencia en las aristas o en los nodos del grafo respectivamente, esto impacta en gran manera en la eficiencia y tiempo de ejecución del proceso (Pevzner et al., 2001; Sohn & Nam, 2018) [Pevzner et al., 2001](#), [Sohn and Nam, 2018](#).

Como resultado de esta primera etapa, es posible la obtención de constructos mayores denominados “Contigs” que representan la parte más básica del ensamblado de genomas. Posteriormente estos contigs pueden ser enlazados utilizando información de Paired-end reads, cuya distancia entre reads es larga (¿1000 nucleótidos) mediante espacios vacíos o gaps (representados por Ns), los cuales no presentan información alguna sobre el modelo construido para el genoma del organismo más que la distancia establecida entre contigs. A estas secuencias enlazadas de mayor tamaño se les denomina “Scaffolds”. Por último, existe la posibilidad de rellenar los gaps presentes en las secuencias generadas mediante el alineamiento de reads adicionales pertenecientes al mismo organismo con el fin de llenar los vacíos no conocidos del genoma, con el fin de obtener una mejor contigüidad y

calidad en el ensamble (ver Figura 3.1). Como resultado, el ensamble de genomas debería dilucidar información sobre el orden de las secuencias nucleotídicas contenidas en cada cromosoma del organismo secuenciado a través de los scaffolds construidos en el proceso de ensamble [Kitts et al., 2016].

3.2. Ensamble por referencia

El ensamble por referencia es una metodología de ensamblado de genomas que utiliza un genoma pre-existente, o una referencia, para la construcción del genoma de un organismo objetivo. Un paso crucial para este proceso es la elección de un organismo cercano evolutivamente al secuenciado. En este método, los reads de secuenciación son alineados contra el genoma de referencia, los cuales al ser mapeados cubren tanto en longitud como en profundidad dicha secuencia. La cobertura de profundidad (o Depth coverage) se refiere a la cantidad de veces en que un único nucleótido es secuenciado, en consecuencia, es posible analizar la distribución y el promedio de depth coverage en un set de secuencias de lecturas de secuenciación. Por otro lado, existe la amplitud de cobertura, que se refiere a la proporción del genoma de referencia cubierto por los reads (Ver Figura 3.2). Cuando dos especies son evolutivamente cercanas analizando su tiempo de divergencia, es de esperar que se presente un alto porcentaje de similitud al comparar el genoma de ambas especies; la cobertura, tanto de profundidad como de amplitud, dependen de dicha similitud. [Olofsson et al., 2019].

Los reads que se alinean sobre el genoma de referencia conforman una única secuencia consenso o canónica, que representará el genoma del organismo objetivo basado en la referencia del organismo cercano (Ver Figura 3.3) [Lischer and Shimizu, 2017]. Una secuencia consenso se genera mediante el cálculo frecuencia de aparición de nucleótidos en una posición específica del genoma; a mayor frecuencia existe una mayor conservación del residuo, es por tanto que resulta canónico en dicha secuencia.

Los algoritmos utilizados para este proceso se encargan de determinar el alineamiento óptimo para el mejor candidato a posición para el read. Al igual que en cualquier alineamiento de secuencias, las inserciones y deleciones aumentan la complejidad del mismo [Langmead and Salzberg, 2012].

Una desventaja de este método es a causa de la diferencia de tamaños entre ambos genomas. Debido a que el ensamble por referencia se basa principalmente en la secuencia de un organismo cercano, el tamaño de secuencia resultante de este proceso será menor

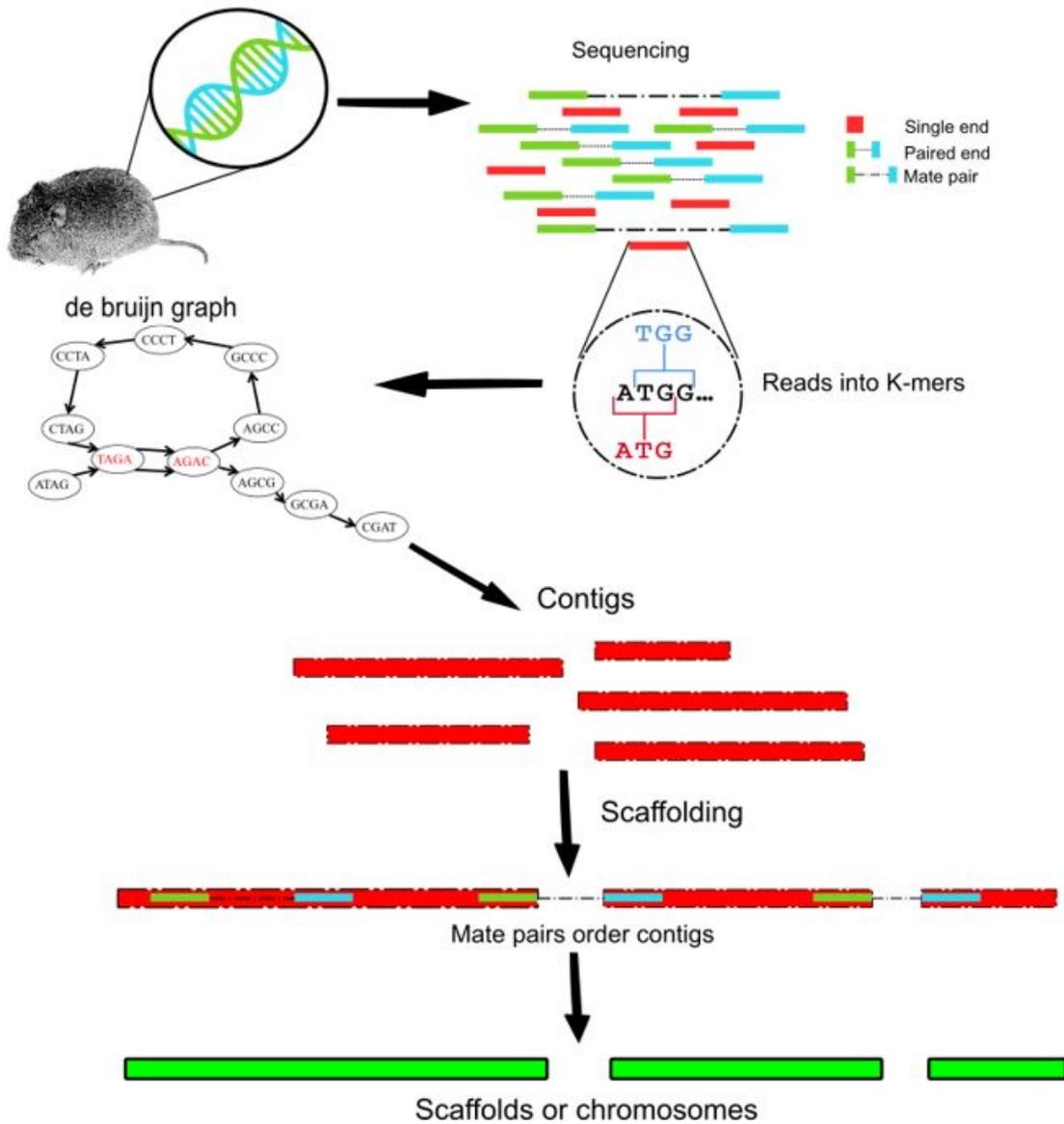


Figura 3.1: Esquema del ensamble de novo.

Las secuencias de lecturas son cortadas en K-mers de tamaño K. el grafo De Bruijn le permite al software encontrar la conformación más viable para la construcción de contigs. Los contigs son mapeados en scaffolds lo que posteriormente podría considerarse como cromosomas.

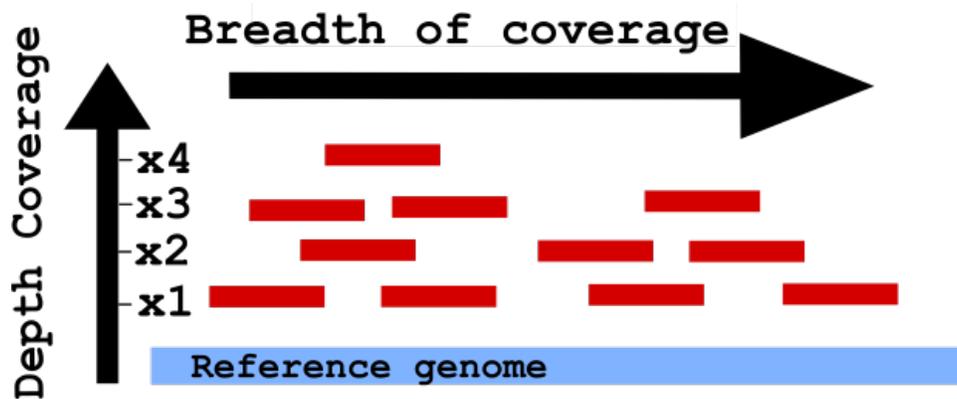


Figura 3.2: Tipos de cobertura.

En rojo se muestran las lecturas de secuenciación alineadas sobre el genoma de referencia. La cobertura de profundidad se considera como la cantidad de veces que una sola base es secuenciada. La cobertura de amplitud se mide como la porción del genoma de referencia que cubierta por los reads durante el alineamiento.

o igual al tamaño del organismo de referencia, dependiendo claramente de la tasa de alineamiento de las lecturas de secuenciación; del mismo modo, la disposición de los genes a lo largo del genoma estará basada en el organismo de referencia. En consecuencia, la calidad del ensamble final estará altamente sujeta al genoma ocupado como molde, por lo que una baja calidad de ensamble en la referencia resultará en un ensamble erróneo para el organismo objetivo [Sayers et al., 2020].

Ambas metodologías de ensamble pueden ser utilizadas y resultan ser convenientes según las preguntas que se desean responder y la calidad, cantidad y cobertura del proceso de secuenciación. Sin embargo, con el propósito de producir una secuencia genómica que pueda suplir la falta de información genética en organismos de la superfamilia Octodontoidea que posteriormente puedan generar estudios comparativos, se ha optado por escoger una especie cercana cuyas características se encuentren registradas, es decir, incurrir en la ejecución de un ensamble por referencia. Dicho esto, es importante cuestionar y evaluar, ya finalizado el ensamble, la calidad de la secuencia generada.

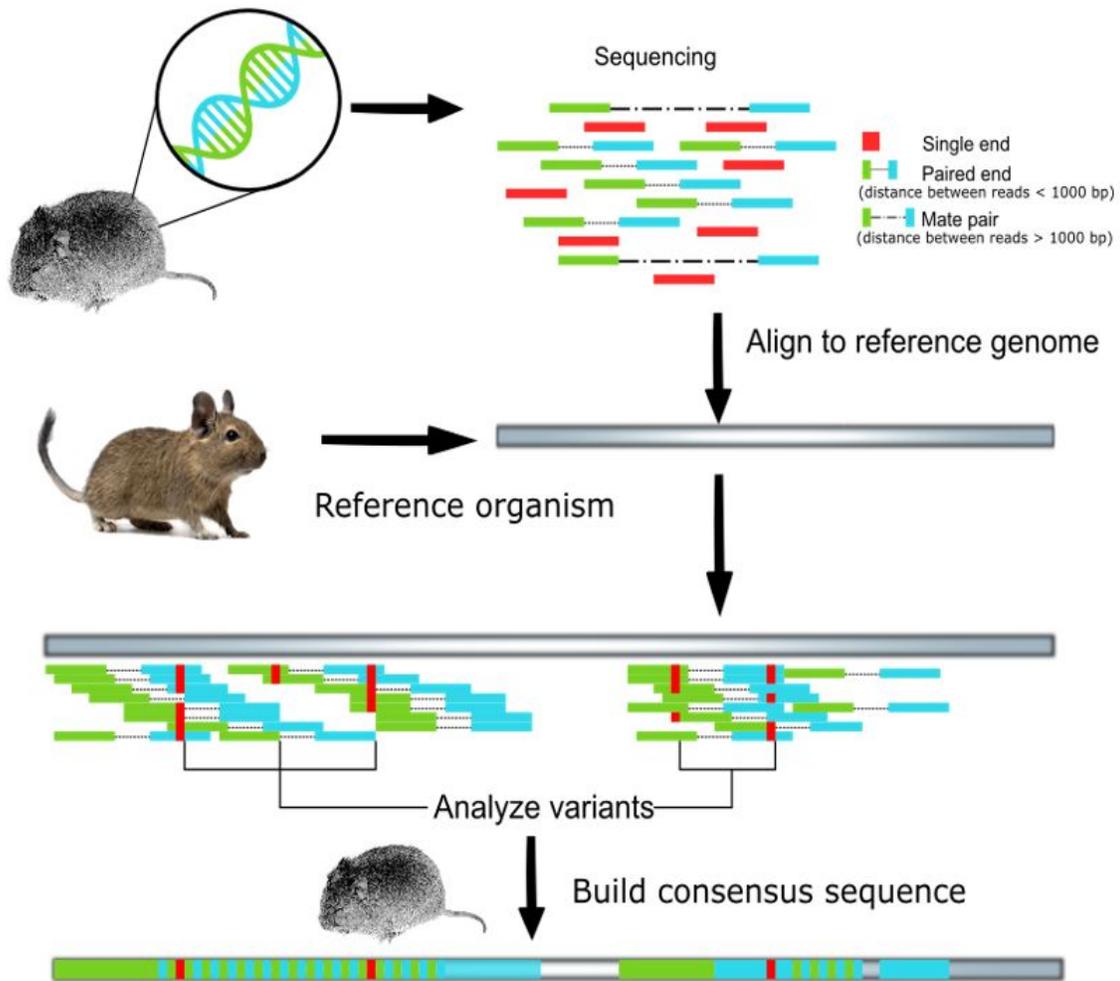


Figura 3.3: Esquema del ensamble por referencia.

Las secuencias de lectura son alineadas en un genoma de referencia. Las variantes por cada posición son analizadas y busca encontrar una secuencia consenso por mayor cantidad de repeticiones de un nucleótido por posición específica. Se crea el consenso y el genoma por referencia.

4. Evaluación del ensamble

Entender algunos conceptos que son esenciales para la evaluación de la calidad del ensamble. Estas métricas responden a características de tamaño, contigüidad y cantidad de las secuencias ensambladas:

- N50: Este parámetro indica qué tan contiguos son los contigs o scaffolds ensamblados. Su medida se basa en el tamaño de la menor secuencia ensamblada ubicada en el 50% del total del genoma ensamblado si las secuencias estuvieran ordenadas de mayor a menor tamaño. Cuanto más alto sea un N50 más contiguo es el ensamble. Por otro lado, un bajo N50 puede ser resultado de una baja cobertura en el set de datos [\[Miller et al., 2010\]](#).
- Largo de secuencias: El primer paso del ensamble genera secuencias en base a reads contiguos. Estas secuencias no son generadas con un tamaño específico, por lo que existe una distribución de largo de contigs y scaffolds. Puesto que el objetivo final del ensamble de genomas es la construcción de cada uno de los cromosomas del organismo objetivo, se espera que la distribución ideal del ensamble sea la de generar un Scaffold por cada uno de los cromosomas del organismo.
- BUSCO: El puntaje BUSCO es una medida que busca estandarizar la integridad de un ensamble de transcritos. Este software se dedica a buscar la proporción de los genes ortólogos altamente conservados. Esto se basa en la suposición de que existe un conjunto de genes que poseen una única copia (single-copy) en el organismo y que se encuentran altamente conservados a lo largo del tiempo; en apoyo a esta idea, BUSCO cuenta con diversos conjuntos de genes clasificados por el linaje del organismo (p.e: mamífero, vertebrado, insecto, eucariota, etc.). Para su funcionamiento, BUSCO convierte los transcritos ensamblados a evaluar en bases de datos y alinea contra ellos el set de transcritos single-copy usado por el

programa [Seppy et al., 2019]. El alineamiento puede resultar en 4 clasificaciones distintas por cada transcrito single-copy alineado:

- * Complete and single copy: El transcrito ensamblado es encontrado de manera única en la base de datos de genes single-copy del linaje buscado ($\geq 95\%$ de similitud).
- * Complete and duplicated: El transcrito ensamblado se encuentra completo, sin embargo, es encontrado más de una vez en la base de datos de genes single-copy del linaje, cuando idóneamente debería ser encontrado sólo una vez.
- * Fragmented: El transcrito ensamblado es encontrado, pero no es considerado íntegro. ($\geq 95\%$ de similitud)
- * Missing: El transcrito ensamblado no es encontrado en la base de datos.

Además de muchas otras métricas comúnmente utilizadas, como lo son el promedio de largos de reads, cantidad de N's en secuencias, secuencias de mayor y menor tamaño en el ensamble, etc. Esto posibilita cuantificar la calidad del ensamble en términos de integridad y contigüidad.

5. Anotación de genomas

La anotación corresponde al proceso computacional que vincula la información biológicamente importante a los datos de secuencia del genoma. En el desarrollo de este proceso se busca identificar y caracterizar elementos importantes contenidos en la secuencia lineal del DNA con el fin de entender el significado de su estructura y funcionalidad. En gran manera, la anotación se ha vuelto un gran reto en la investigación de organismo a nivel genómico, pues si bien las últimas generaciones de secuenciación masiva les han otorgado a los científicos la capacidad de construir una gran variedad de genomas gracias al bajo costo que ha acarreado a lo largo de los años y a las características que poseen las lecturas (contigüidad, cobertura, etc.) que mejoran con cada nueva tecnología desarrollada; el proceso de anotación se ha convertido en un desafío debido a la naturaleza exótica y variedad de muchos genomas recientemente secuenciados, es por ello que muchos grupos consideran esta información como un consenso de las características genómicas correspondientes a un organismo secuenciado [Rust et al., 2002, Yandell and Ence, 2012].

El desarrollo de una anotación se encarga de caracterizar el genoma de un organismo a dos niveles: el estructural, el cual registra la disposición de las secciones de un gen en términos de coordenadas dentro de un cromosoma, y la anotación funcional de genes, que identifica la función del producto del gen analizado tanto en forma de términos GO -ontología de genes- como asignándole rutas metabólicas ligadas, ‘gene name’ y descripciones (Ver figura 5.1). Ambos niveles son utilizados por los investigadores para enfocar los registros a su campo de estudio teniendo en cuenta la mayor cantidad de información posible sobre el organismo de interés [Kitts et al., 2016].

Este proceso puede ser abordado de distintas maneras según la cantidad de información con la cual se disponga, por lo que, es preciso distinguir entre la anotación de genes basada en evidencia y la predicción de genes para la anotación. La anotación por evidencia busca el alineamiento y el registro de elementos en el genoma previamente secuenciados o

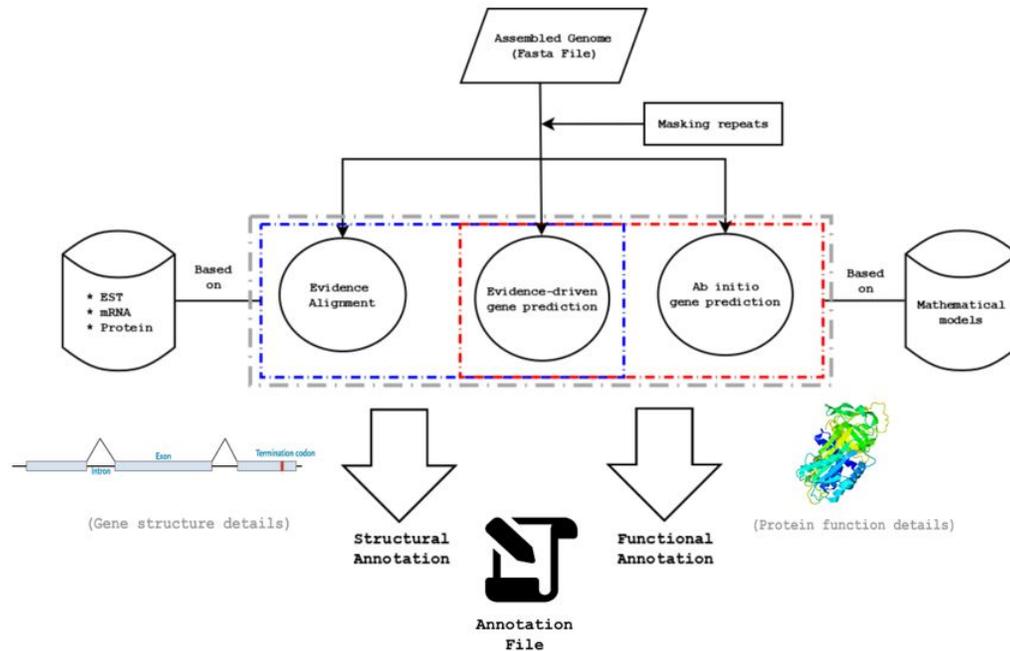


Figura 5.1: Anotación de genomas.

Proceso de caracterización de los diversos elementos contenidos en el genoma, basados en distintos tipos de información. En el recuadro azul se presentan los métodos basados en evidencia extraídas a partir de procesos experimentales. En el recuadro rojo se presentan los métodos basados en modelos matemáticos, por ejemplo, *Hidden Markov Models*.

reconocidos por el investigador que tengan cabida en el genoma o su función; registros como secuencias de proteínas, secuenciación de RNA o marcadores de secuencias expresadas (EST), funcionan como evidencia disponible para apuntar en la anotación.

Por otro lado, la anotación por predicción *ab initio* se encarga de identificar mediante modelos matemáticos la ubicación de un elemento genético en función de la información conocida en organismos cercanos filogenéticamente al objeto de estudio. Es debido a ello que la mayoría de los softwares de predicción de genes poseen módulos encargados de entrenar el modelo matemático previo al uso del programa, mejorando así la precisión en la anotación de elementos genéticos sin evidencia. Del mismo modo, existen múltiples herramientas *ab initio* que dan uso a evidencia externa para mejorar aún más dicha precisión, a esto se le denomina predicción de genes basado en evidencia. Este último proceso da una gran mejora en la anotación del genoma de organismo que no han sido caracterizados previamente, debido a que comúnmente los métodos *ab initio* utilizados para ellos dependen fuertemente del set de entrenamiento, el alineamiento de evidencia externa otorga veracidad a las predicciones obtenidas por estos softwares y disminuyen el error de la predicción [Scalzitti et al., 2020].

La construcción de la anotación del ensamble final es en gran medida útil para la realización de técnicas de análisis comparativo entre dos organismos. La descripción de la función, tamaño y ubicación de los elementos del genoma no sólo contribuyen a la detección de homología entre ambas especies, sino que también son esenciales en la descripción y caracterización de otros genomas. Es por ello que la información registrada para un organismo debiera ser periódicamente actualizada a la luz de nueva información descrita a partir de nuevos organismos anotados o en la inclusión de nueva evidencia registrada para el uso público.

5.1. Importancia de la anotación

Comprender que el orden específico de los nucleótidos y aminoácidos confiere propiedades estructurales o funcionales a una secuencia biológica, nos permite definir y entender ciertas regiones o segmentos que, dentro de una extensa secuencia, confieren características importantes al organismo en el cual se alojan. De este modo, la anotación podría entenderse como la escritura de un plano que almacena las coordenadas de dichas secciones dentro de una secuencia mayor (como lo podría ser un gen almacenado en un cromosoma); y que nos permite a nosotros como investigadores enfocar el estudio a secciones específicas en vez de interactuar con secuencias demasiado extensas [Abril and Castellano, 2018].

Además de entender el significado de las secuencias, tanto por su estructura y funcionalidad, las anotaciones de genomas dan sustento y recursos al trabajo de una gran variedad de laboratorios en todo el mundo. Múltiples organizaciones se publican sus resultados en bases de datos biológicas disponibles para el público, entre ellas se encuentran:

- RefSeq
- Ensembl
- UniProt
- Encyclopedia of DNA Elements (ENCODE)
- GENCODE, entre otros.

Estas bases de datos generalmente son consultadas por los investigadores para documentar y sustentar sus proyectos en su campo de estudio [Sayers et al., 2020].

La descripción de cada uno de los elementos contenidos en un genoma resulta útil para definir la interacción y el papel que juegan en un sistema de mayor tamaño y complejidad, como la participación de sus productos en rutas metabólicas o la inclusión de ligandos con acciones regulatorias con afinidad a estos elementos; esto logra integrarse en un reciente campo de estudio denominado “System Biology” [Huttlin et al., 2017]. Por último, la anotación de elementos funcionales de un genoma es un componente de vital importancia en los estudios de filogenia y análisis de expresión de genes de un organismo; donde en este último precisamos de un elemento target con el cual trabajar. Por otro lado, el estudio evolutivo in silico es fuertemente influenciado por la comparación de todas las características registradas en genoma del organismo, en paralelo con otras especies cercanas a él, por lo que el detalle de las diferencias entre dicho grupo suele ser una de las preguntas frecuentes en este tipo de estudios.

6. Anotación y Ensamble en organismos de la superfamilia *Octodontoidea*

El panorama actual de los organismos pertenecientes a la superfamilia *Octodontoidea* nos otorga una escasa cantidad de información genómica depositada en las bases de datos biológicas. Tanto así que dentro de la familia de los equimiídos (*Echymidae*) y abrocomis (*Abrocomyidae*) no es posible encontrar ningún registro de proyecto de secuenciación o de ensamble. Para los roedores tucos-tucos (*Ctenomyidae*) existe registro de ensamble de genoma mitocondrial en el organismo *Ctenomys sociabilis* (social tuco-tuco) en el registro ‘Nucleotide Database’ de NCBI (NC_020658.1).

Por otro lado, para la familia *Octodontidae* existe registro de más de un organismo, los cuales son (Ver Tabla 6.1): *Octomys mima*, conocida como la rata vizcacha del monte, es endémica de Argentina y la única especie en el género de los *Octomys*; *Tympanoctomys barrerae*, conocida como la rata vizcacha roja, endémica de las estepas secas y salinas del centro-oeste de Argentina y única en el género *Tympanoctomys* [Evans et al., 2017]; y *Octodon degus*, conocido como degú (del mapudungún dewü), especie no subterránea endémica de Chile, son recurrentemente objeto de estudio dada su alta capacidad de organización y sociabilidad.

Los estudios de filogenia entre las especies de la familia *Octodontidae* no han sido comprendido desde un punto de vista genético; por lo que, si bien se infiere que su especiación al hábitat subterráneo es en gran medida gracias a cambios genéticos, no existe base alguna que dé sustento a esta afirmación, y menos aún a señalar la causa de que esta diversificación de ambientes haya existido en dos puntos independientes uno de otro en espacios de tiempo separados. Para llevar a cabo esta comparación, se debe establecer las

	Octomys mimax	Tympanoctomys barrerae	Octodon degus
Versión	1	1	1
Fecha	17/10/2017	17/10/2017	01/05/2012
Tecnología de secuenciación	Illumina HiSeqX	Illumina HiSeqX	Illumina Hi-Seq
Nivel de ensamble	Scaffold	Scaffold	Scaffold
Método ensamble	ABYSS v1.9	ABYSS v1.9	AllPaths v. R40507
Cobertura	10x	10x	80x
Número de Scaffolds	1,257,804	1,558,800	7,135
N50 Scaffold	4,874	4,698	12,091,372
ID GenBank	GCA_002564305.1	GCA_002564285.1	GCA_000260255.1

Tabla 6.1: Detalle de organismos Octodontidae ensamblados.

Ensamble y anotación disponibles de genomas de organismos Octodontidae caracterizados hasta la fecha. Datos recopilados de la base de datos GenBank.

diferencias entre organismos aparentemente cercanos que coexistan en distintos hábitats, sin embargo, existen claras limitaciones con la información disponible en las bases de datos biológicas. Para solucionar esta problemática se ha optado por ensamblar y anotar a *A. sagei*, uno de los organismos subterráneos teóricamente más cercanos a uno de los pocos roedores caracterizados de este linaje (*Octodon degus*) que habita en la superficie. Es de esperar que el resultado de esta investigación permita afrontar de manera más clara la comparación entre ambas especies.

7. Problema

Ausencia de registros genómicos en organismos del linaje Octodontidae cuyo hábitat se encuentre en el ambiente subterráneo.

8. Solución propuesta

Aplicación de protocolos de ensamble y anotación de genomas en datos DNA-Seq y RNA-Seq del organismo *Aconaemys sagei*.

9. Objetivos

9.1. Objetivo General

Producir el ensamble y la anotación del genoma de la especie *Aconaemys sagei* perteneciente a la superfamilia *Octodontoidea*.

9.2. Objetivos Específicos

1. Elegir modelo de referencia entre especies candidatas cercanas a *A. sagei* para la preparación del ensamble por referencia.
2. Ensamble por referencia del organismo *A. sagei* a partir de set de lecturas de secuenciación y modelo por referencia elegido.
3. Ejecutar protocolo de anotación sobre el modelo de genoma (ensamblado).

10. Metodología

Dentro de las primeras consideraciones para el trabajo de lecturas de secuenciación, es la verificación de la integridad de estas mismas antes de su posterior procesamiento. Para ello se usó el software FASTQC [Andrews, 2010] en su versión 0.11.5. Este es un software orientado a extraer y visualizar información estadística en librerías de lecturas de secuenciación con el fin de comprobar la calidad de librerías antes y después del filtrado. Provee un resumen gráfico sobre el promedio y distribución de calidad de los nucleótidos secuenciados por posición específica en el read a lo largo de todo el archivo gracias a la información proveída por los archivos de extensión “.fastq”. Además de esto, FASTQC grafica la distribución de largos, el contenido de GC, contenido de N’s, contenido de bases por posición y posibles contaminaciones por vectores que pudieran presentarse durante la secuenciación de la muestra en el laboratorio. (Ver Anexo 1).

Toda esta información resulta útil en el control de calidad, puesto que, si el resultado de los gráficos fuera desfavorable, es necesaria la intervención mediante el corte y filtrado de las secuencias de lectura, proceso denominado “Trimming”. Trimmomatic (versión 0.36) es un software orientado al corte y filtrado de secuencias de reads en formato .fq o .fastq. Este programa posee distintas opciones de corte de acuerdo al criterio del usuario según se quiera operar bajo un largo de read específico o de acuerdo a la calidad que posee cada nucleótido en su posición dentro del read (Ver Figura 10.1). Del mismo modo, si un adaptador es encontrado durante el chequeo con FASTQC, es posible cortar hasta el término del adaptador con Trimmomatic [Bolger et al., 2014]. (Ver Anexo 2)

Este proceso es iterativo, se debe verificar la calidad y cortar si es necesario de forma continua hasta obtener la calidad deseada, idealmente sin perder muchos nucleótidos durante el proceso.

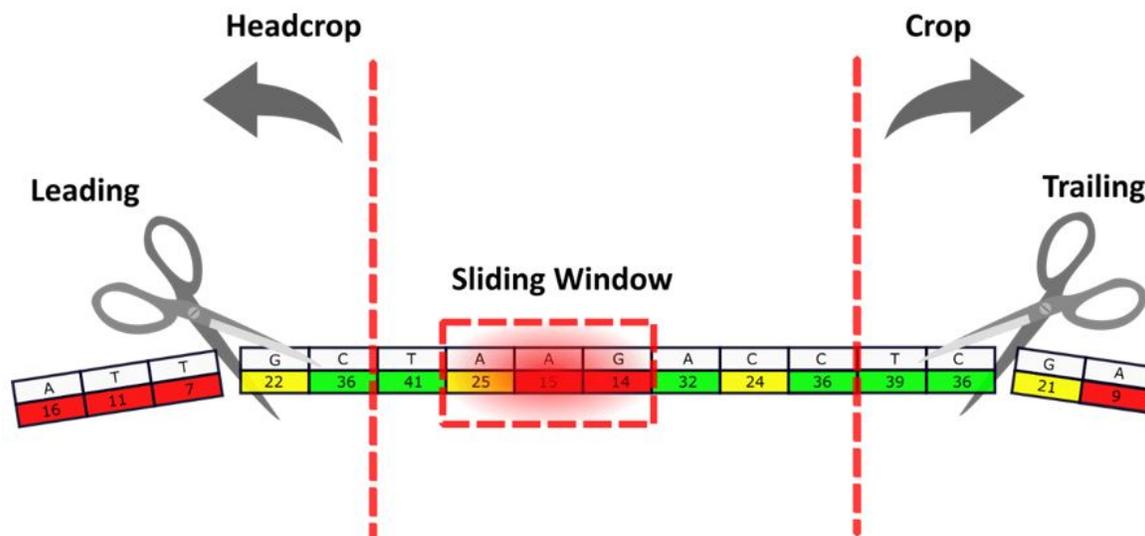


Figura 10.1: Trimming de reads.

Se muestran las opciones de cortar zonas de baja calidad en reads (LEADING, TRAILING y SLIDINGWINDOW de Trimmomatic). Y opciones de cortar hasta un tamaño específico (CROP y HEADCROP) entre otras.

10.1. Elección de modelo de organismo por referencia.

En la elección del modelo de referencia que se usó para el ensamblaje, se analizaron tres especies roedoras aparentemente cercanas a la superfamilia *Octodontoidea*, las cuales fueron *Cavia porcellus*, *Chinchilla lanígera* y *O. degus*. Para ello se usó el software Bowtie v2.3.4.3 [Bolger et al., 2014], sobre cuyo resultado se analizaron las métricas de cobertura y tasa de alineamiento. El tiempo de divergencia entre el set de lecturas de *A. sagei* y las especies *C. porcellus* y *C. lanígera* es superior a los 30 millones de años, en cambio para *O. degus* es cercano a los 3,7 millones de años atrás [Opazo, 2005].

10.2. Ensamblaje por referencia

El ensamblaje por referencia es, en principio, la elaboración de la secuencia consenso resultante al combinar las lecturas del organismo target con el genoma del organismo de referencia. Para poner en contacto ambos elementos se ejecutó el alineamiento de los reads de *A. sagei* sobre el genoma del organismo de referencia. Para ello se utilizó el software Bowtie2 con el cual se indexó el genoma de referencia y posteriormente se obtuvo el archivo binario de alineamiento .bam (Ver Anexo 3).

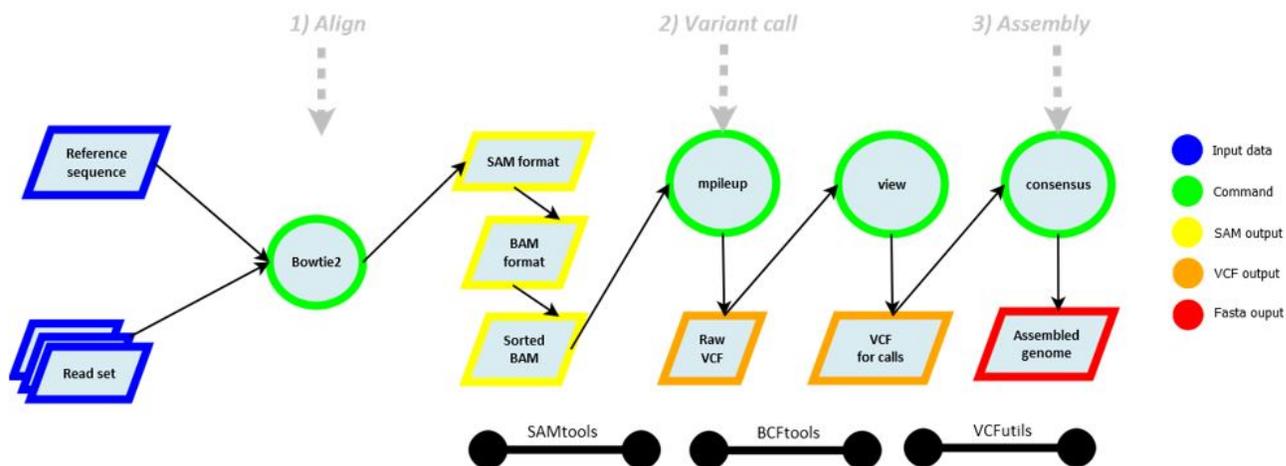


Figura 10.2: Ensamble por referencia.

Los archivos de lectura son alineados sobre el genoma de referencia. Se usa samtools para manipular el archivo de alineamiento y en conjunto con bcftools y vcfutils para generar la secuencia consenso.

Los registros del alineamiento fueron ordenados e indexados con samtools v0.1.19 (Li et al., 2009) con los módulos sort e index. Posteriormente, se utilizó en una única línea de comando una combinación de los softwares participantes en la creación de la secuencia consenso; estos son bcftools v1.4.1-16-g034870e y vcfutils.pl (Li and Barrett, 2011) en su versión de perl (Ver Figura 10.2). La finalidad de usar estos tres programas en conjunto fue la de encontrar todas las posibles variantes en posiciones específicas del genoma; estas variantes son dadas por la cobertura de las lecturas sobre el genoma de referencia, por lo que es posible observarlas como SNP's "apilados" entre si gracias al módulo mpileup de samtools. El programa bcftools con su módulo view y su opción -c, se encargan de contar las variables presentes en una posición específica del genoma de referencia gracias a las lecturas, con el fin de determinar el nucleótido con mayor cantidad de repeticiones (Danecek et al., 2011). El programa vcfutils.pl es el encargado de traducir estas variantes en una secuencia consenso con el módulo vcf2fq, el cual como su nombre lo indica, sustituye las bases donde existen variantes por aquellas con mayor cantidad de repeticiones y transforma la salida en un archivo con propiedades .fastq. Por último, se recurrió al módulo seqret de EMBOSS v 6.6.0.0 para transformar el archivo de extensión .fastq por .fasta. (Ver Anexo 3)

10.3. Ensamble de transcritos

El ensamble de transcritos fue realizado con el software TRINITY v2.8.5, el cual es un ensamblador *de novo* popularmente utilizado en el ensamble de transcritos mediante lecturas de RNA-Seq. Este programa se divide en tres fases: en el primer paso Trinity construye contigs usando los reads como si se trataran de K-mers, extendiéndolos el mayor largo posible, luego estos contigs son clusterizados; debido a que se trata de un ensamblador *de novo*, Trinity hace uso del grafo De Bruijn, en donde por cada uno de los clusters generados, son extraídas las secuencias más probables de ser una isoforma al gen secuenciado [Grabherr et al., 2011, Haas et al., 2013]. El ensamble fue generado usando 150G de los 252G disponibles en Exxact y 64 de las 128 CPU's disponibles (Ver Anexo 1).

Para verificar la integridad de los ensambles se usó el software BUSCO v3.0.1 con cuatro distintos linajes a los cuales pertenece el organismo *A. sagei* para observar la integridad de las secuencias ensambladas por Trinity. Los cuatro linajes elegidos fueron: Eucariota, Vertebrado, Mamífero y Tetrápodo.

10.4. Anotación de genomas

Para la anotación de genomas se utilizó una batería de softwares provenientes del protocolo de anotación MAKER v2.31.10 [Cantarel et al., 2008]. Durante su ejecución, MAKER da uso de múltiples softwares para cada uno de los procesos señalados en la Figura 1.5; para ajustar los parámetros de cada uno de estos programas, el comando `maker -CTL` genera tres archivos de parámetros necesarios para su ejecución:

- `maker_exe.ctl`: En este archivo se especifica la ruta de cada uno de los programas participantes en la anotación de genomas. Al ejecutar el comando generador de archivos de parámetros, maker rellena de manera automática los campos de los softwares cuya ruta se encuentre en el perfil `.bash` de Linux. Los programas se encuentran separados según la función que desempeñan en la anotación y muchos de ellos no son obligatorios para su ejecución; otros, por el contrario, son indispensables para ello, como es el caso de BLAST v 2.8.1, RepeatMasker v4.1.0, Augustus v2.5.5 y snap v2006-07-28 según el tipo de anotación que se quiera realizar. (Ver Anexo 4)
- `maker_bopts.ctl`: En este archivo se indican los parámetros de los programas BLAST y Exonerate; para el primero indicando umbrales de porcentaje de cobertura y porcentaje de identidad, y puntaje de corte de e-value y bit; todo esto para los

módulos blastn, blastx (para alineamiento de proteína-genoma y enmascaramiento de elemento transponible) y tblastx. En el caso de Exonerate se especifican los umbrales de puntaje en términos porcentajes con el fin de obtener la porción de alineamiento que se encuentren sobre el porcentaje indicado del total de matches tanto de proteínas como de nucleótidos. (Ver Anexo 5)

- `maker_opts`: Este archivo es el más extenso de los tres. En él se deben indicar la ruta del genoma para anotar también de los archivos de que contienen la evidencia para el alineamiento, los parámetros de uso del resto de los programas de predicción y enmascaramiento, opciones de re-anotación si este fuera el caso y las opciones del comportamiento de maker en el registro de genes en el archivo gff. (Ver Anexo 6)

Previo a la ejecución de la anotación, existe la opción de entrenar el modelo matemático usado por los predictores de genes *ab initio*, en este caso, Augustus y SNAP; ambos dan uso de modelos de Markov para su predicción, sin embargo, el primero es utilizado en conjunto con la evidencia para aumentar la precisión del resultado [Korf, 2004, Stanke and Waack, 2003]. Los modelos matemáticos son entrenados usando una muestra del transcriptoma de la referencia, debido a que es el organismo más cercano a *A. sagei* y en el cual fue basado el ensamble.

Para el proceso de anotación del genoma ensamblado por referencia de *A. sagei*, se usaron los transcritos ensamblados anteriormente por el software Trinity para la ejecución de los tres tipos de modelos de anotación. En primera instancia se ocupó RepeatMasker [Smit et al., 2013] para realizar soft-masking al genoma, el soft-masking reemplaza todas las secuencias repetidas, elementos transponibles (obtenidos de la base de datos RepBase) y regiones de baja complejidad por nucleótidos en tipografía minúscula para diferenciarla de regiones codificantes, las cuales irán en mayúscula. Posteriormente fueron alineados los transcriptomas ensamblados de novo y secuencias de proteínas del organismo de referencia mediante BLAST [Altschul et al., 1990, Camacho et al., 2009] y Exonerate [Slater and Birney, 2005]. Por último, son ejecutados los predictores de genes SNAP y Augustus, con el fin de encontrar aquellos genes no identificados por el alineamiento de transcritos y proteínas (Ver Figura 10.3). Esto es repetido automáticamente por cada uno de los scaffolds y contigs presentes en el archivo contenedor del genoma, obteniendo de esta manera una anotación por cada secuencia encontrada en el fasta [Campbell et al., 2014, Holt and Yandell, 2011].

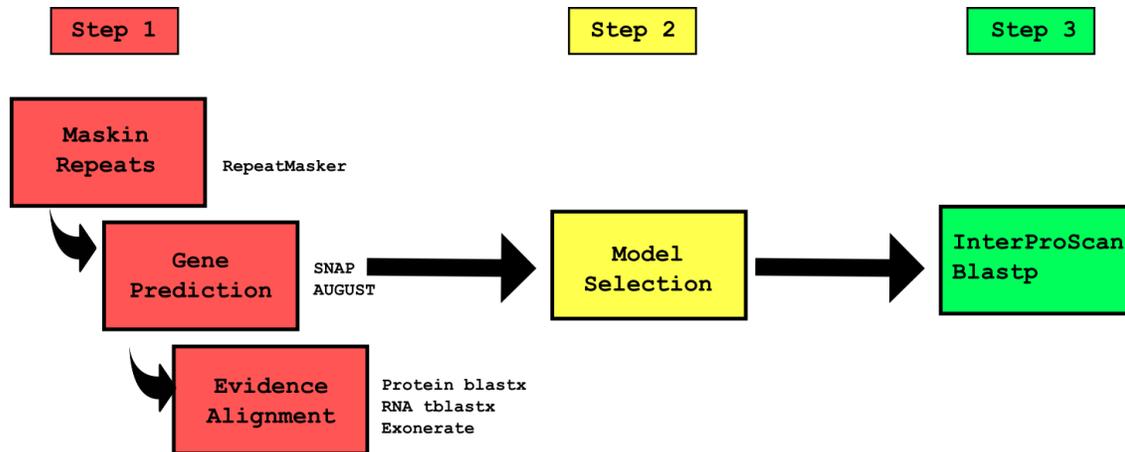


Figura 10.3: Pipeline de anotación MAKER

En el paso 1 se identifican y enmascaran las secuencias repetidas, se predicen modelos de genes y se alinea la evidencia de RNA y proteína. En el paso 2 se selecciona el modelo de predicción a ser anotado y se completa la anotación estructural. En el paso 3 los modelos se someten a un alineamiento local con InterProScan y blastp para la anotación funcional.

10.5. Post-procesamiento de la anotación

Como resultado de la ejecución de maker se obtienen múltiples carpetas, donde por cada una de estas existe una anotación para un único scaffold o contig. Para combinar estas anotaciones en un único archivo maker provee una serie de accesorios para su ejecución, entre ellos se encuentra `gff3_merge` y `fasta_merge`, los cuales al ser ejecutados recolectan y unen la información para crear archivos únicos de anotación (`gff3`) y secuencias (`fasta` de proteínas y transcritos para el organismo *A. sagei*). Muchos de estos registros, sin embargo, no son identificados más allá de las coordinadas en las que se encuentran. Para solucionar esto se realizó una anotación funcional.

La anotación funcional es el método por el cual todos los genes identificados dentro del genoma reciben descripciones basados en la función del producto al cual codifican (Nombre del gen, producto proteico y otros detalles). Se usó `blastp` e `InterProScan v5.44-79.0` [Jones et al., 2014] contra la base de datos `SwissProt/UniProt (D506-D515, 2019)` y la colección `Pfam`, lo que nos permitió obtener la terminología `GO (Gene Ontology)` [Ashburner et al., 2000, Carbon et al., 2019] y toda la información necesaria para la descripción los genes anotados. Esta información fue mapeada y reemplazada en el archivo de anotación en bruto y en los archivos de secuencia de transcritos y proteínas mediante el uso de los módulos `map`, `maker_functional` e `iprscan` de software `MAKER` (Ver anexo 7), obteniendo así la anotación final para el organismo *A. sagei*.

11. Resultados

A continuación, se presentan las características de las lecturas de secuenciación del organismo *A. sagei*, generadas mediante tecnología Illumina HiSeq 2500.

Muestra	<i>Aconaemys sagei</i>
Tamaños de inserto (bp)	350
	550
	5,000
	8,000
	12,000
Tipo de librería	R1, R2 (paired-end)
	R0 (single-end)

Muestra	Tamaño de inserto (bp)	N. ^o Total de Reads
<i>Aconaemys sagei</i>	350	410,173,564
	550	495,295,125
	5,000	229,952,122
	8,000	220,996,627
	12,000	226,188,656

Tabla 11.1: Detalle de librerías de secuenciación Illumina HiSeq 2500 en muestras de *A.sagei*.

Los resultados de calidad del set de lecturas de DNA fueron recopilados por cada uno de los tamaños de inserto y expresados en un solo gráfico de calidad por base para cada una de las librerías paired-end (R1 y R2) y single-end (R0).



Figura 11.1: Resultados de calidad posterior al trimming.

En el eje x se encuentra la posición en el read (1 - 150), en el eje y se encuentra el puntaje promedio de las calidades por posición específica para cada uno de los reads.

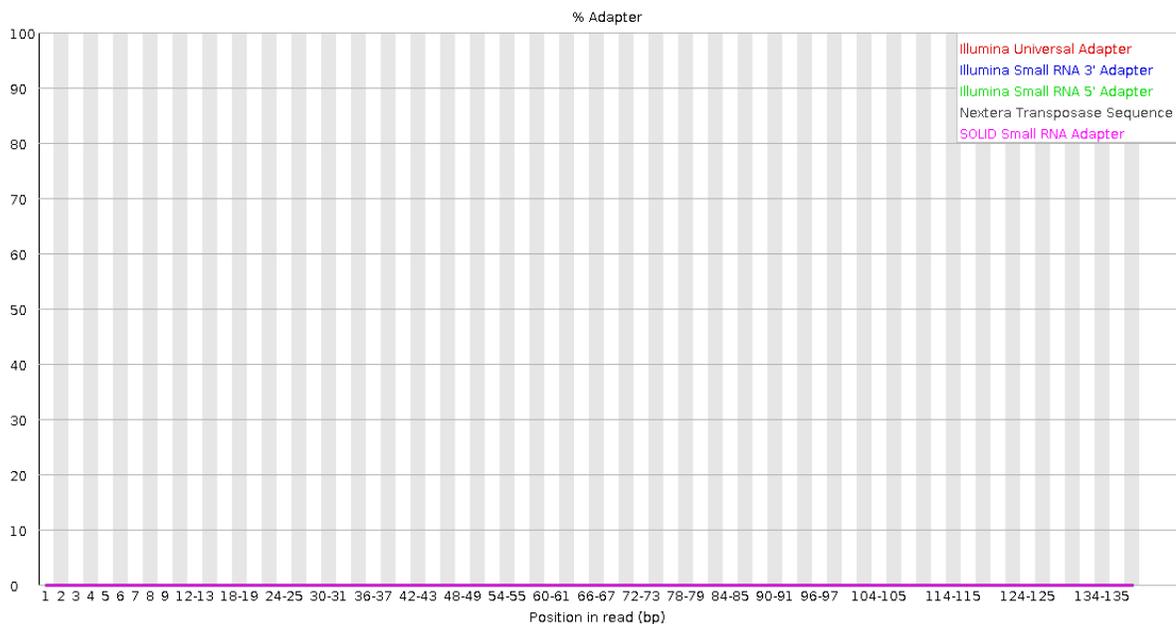


Figura 11.2: Ausencia de adaptadores después del trimming.

En el eje x se encuentra la posición en las lecturas, en el eje y el porcentaje de adaptadores encontrados

A continuación, se presentan los resultados del alineamiento de las lecturas de secuenciación de *A. sagei* sobre los genomas de las tres distintas especies escogidas como candidato a modelo para el ensamble por referencia.

Organism	Alignment rate
<i>Chinchilla lanígera</i>	43.55 %
<i>Octodon degus</i>	84.83 %
<i>Cavia porcellus</i>	16.70 %

Tabla 11.2: Resultados de alineamiento en candidatos a modelo de ensamble por referencia.

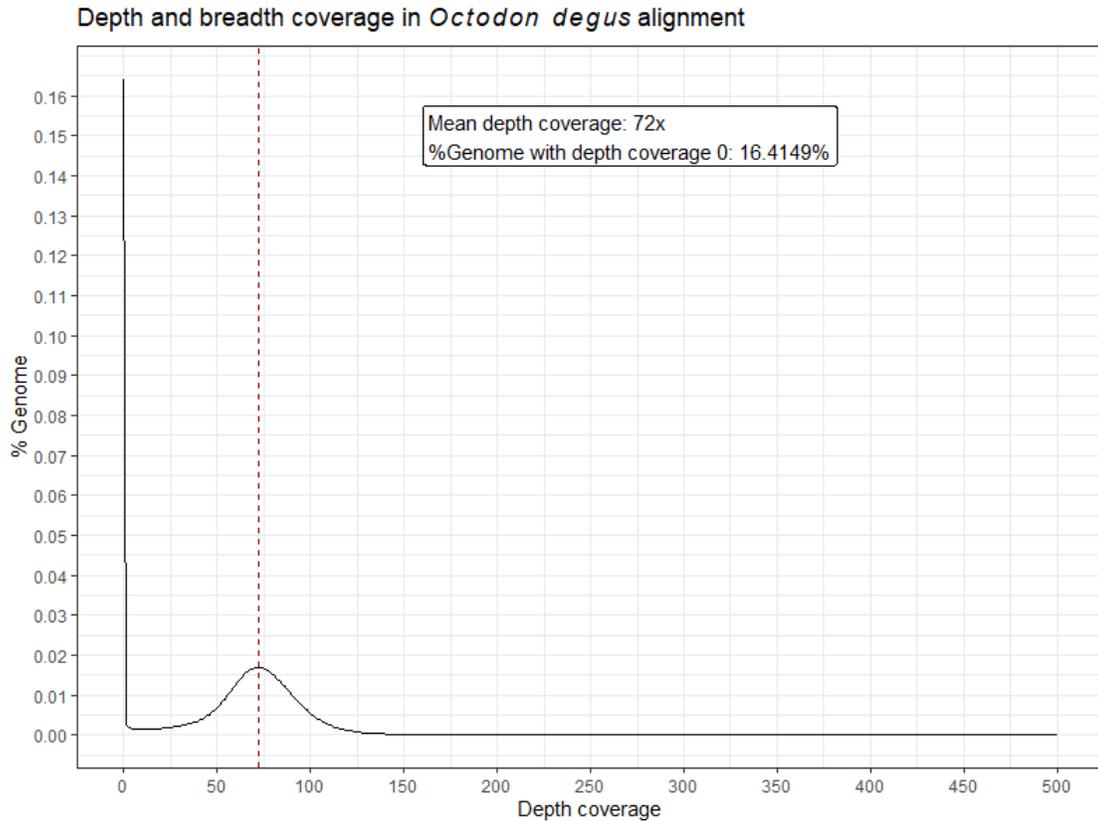


Figura 11.3: Análisis de cobertura.

En el eje x se muestra la distribución de cobertura luego del alineamiento, en el eje y se muestra la fracción del genoma con x cobertura. En la anotación se muestra el peak de cobertura promedio y el % del genoma con cobertura 0.

La calidad del ensamblaje por referencia puede verse graficado por sus estadísticas principales como lo son el N50, promedio de largos y número de scaffolds presentes en el archivo .fasta generado (Figura 11.4 y 11.5).

El software BUSCO tiene su propio script para generar gráficos en R a partir del archivo resumen de los resultados obtenidos. A partir de ello se puede expresar de manera gráfica la porción de BUSCO's por cada uno de los linajes analizados. En base a esto, se realizó el ensamblaje de transcriptoma por Trinity, el cual duró 2 horas con 17 minutos, siendo posteriormente analizado en BUSCO (Figura 11.5).

A partir del registro realizado en la anotación del genoma de *A. sagei* se puede obtener desde su segunda columna, la cantidad de elementos totales anotados luego de la predicción de genes ab initio y el alineamiento de la evidencia en el genoma. La anotación funcional del mismo modo, también incluye una etiqueta en la descripción del elemento, gracias a ello es posible realizar un conteo de los elementos totales anotados y la cantidad de genes

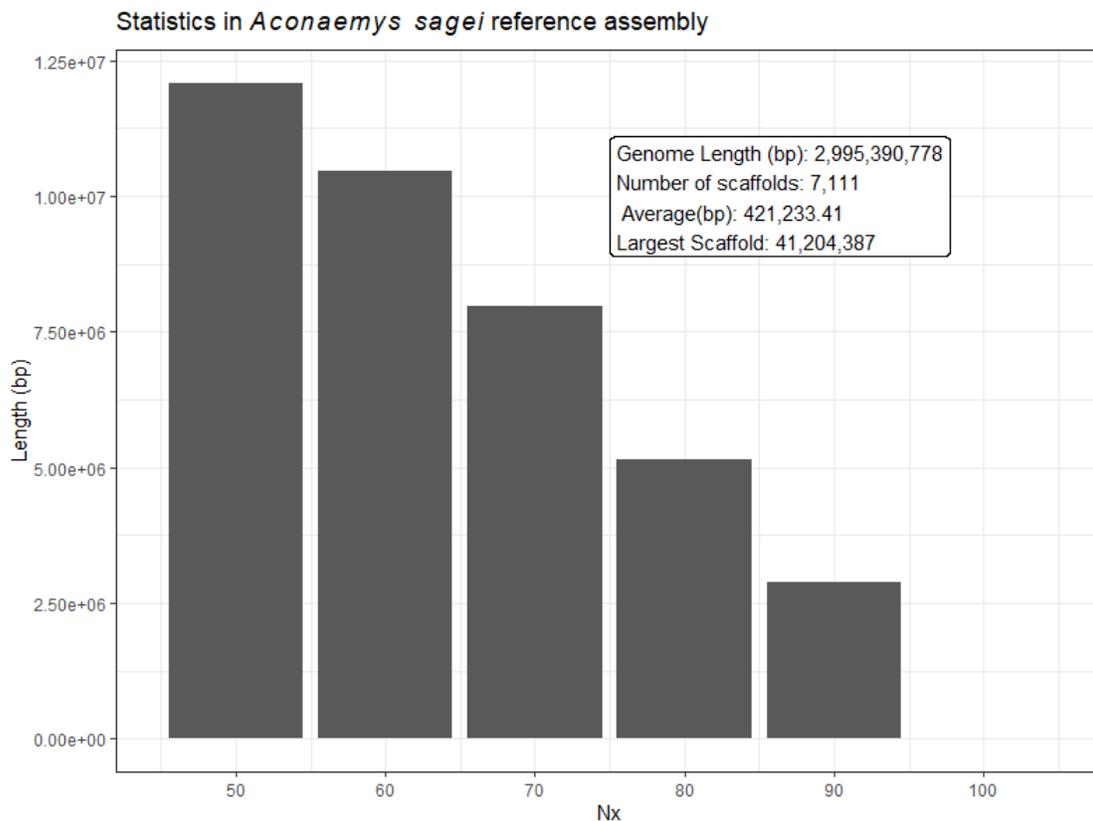


Figura 11.4: Análisis de calidad de ensamble.

En el eje x se muestra la métrica Nx donde x va de 50 a 100. En el eje y se muestra la cantidad de bases correspondientes al Nx. En la esquina superior derecha se muestra un resumen del tamaño total, número de scaffolds, promedio de largos y scaffold más largo.

que recibieron su “GO name” luego de la anotación funcional (Tabla 11.3 y Figura 11.7). La duración total del proceso fue de 5 días con 4 horas para un genoma de casi 3 Gbp, esto fue realizado en un computador con 128 procesadores usando solamente 60 de ellos para la realización de alineamientos locales por scaffold, con una capacidad máxima de memoria de 256G.

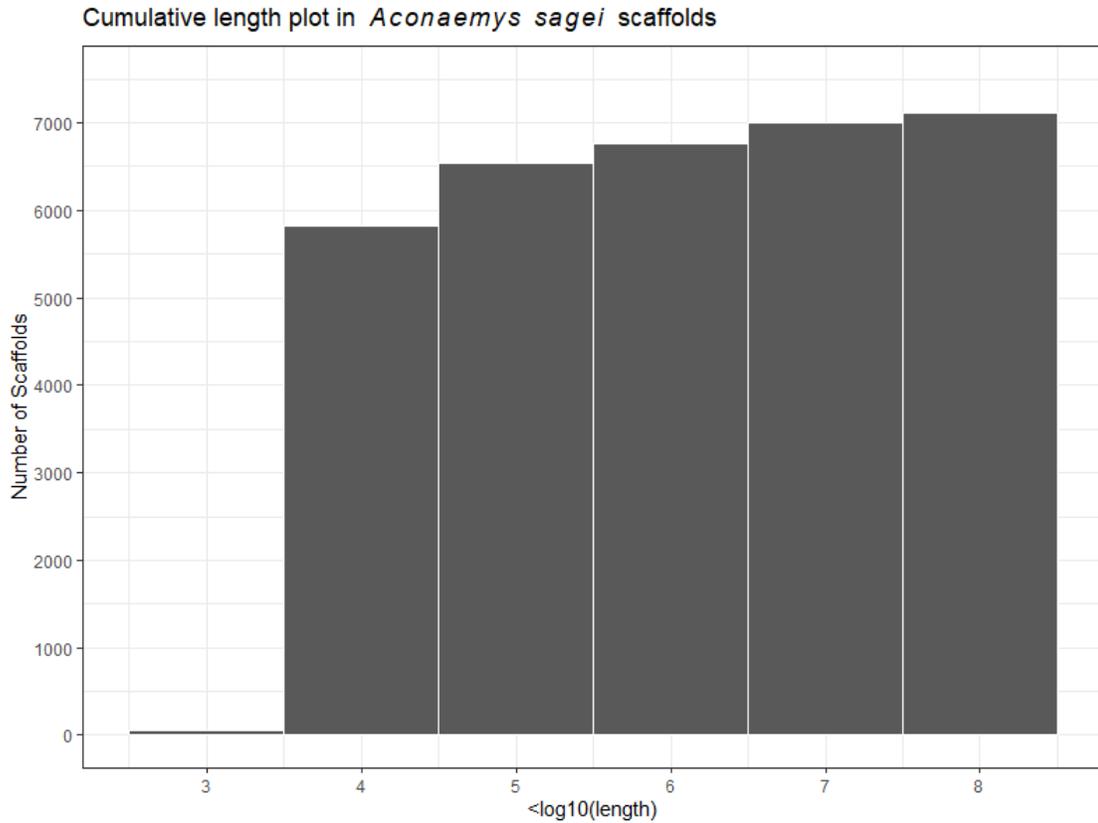


Figura 11.5: Distribución de acumulada de largo de scaffolds en ensamble por referencia.

En el eje x se muestra el umbral de largo con magnitudes de 10^x , en el eje y se muestra la cantidad de scaffolds con largo inferior al umbral.

Por último, es posible analizar la integridad del transcriptoma final en adición a la evidencia proveniente de la referencia y la predicción de genes ab initio mediante el mismo análisis realizado en los resultados de Trinity (Figura 11.8).

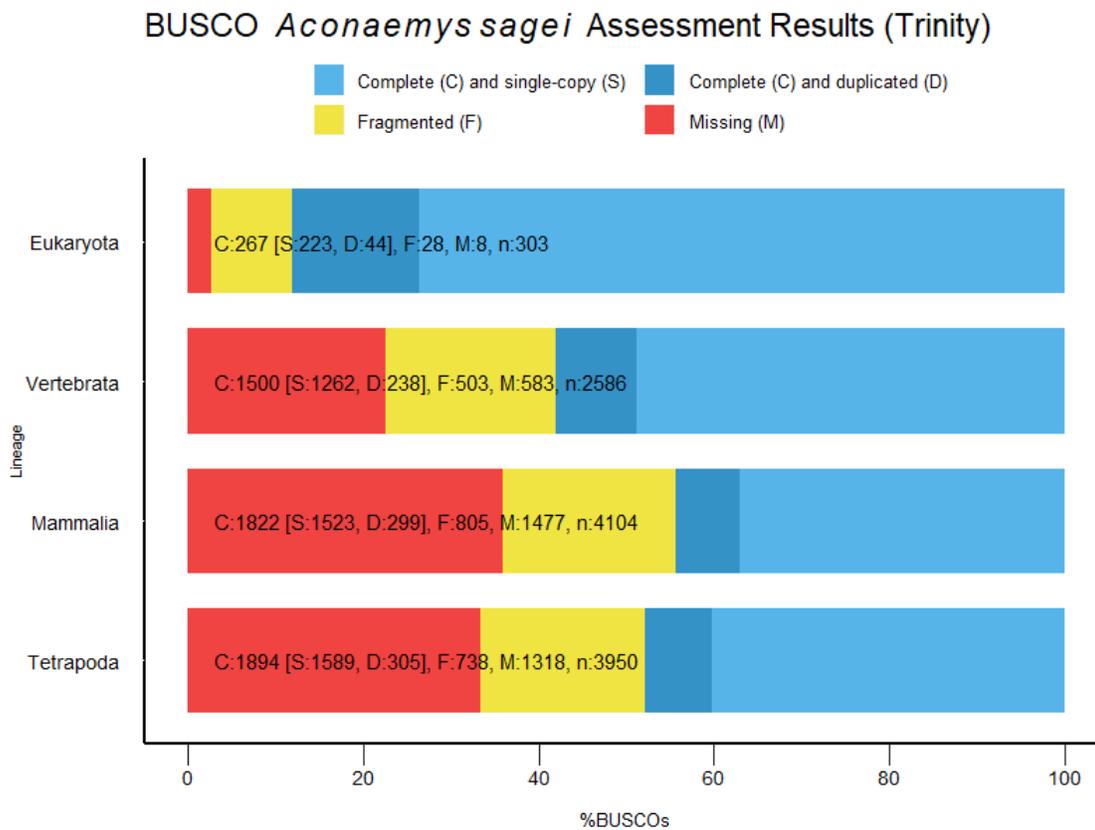


Figura 11.6: Resultados BUSCO del ensamble de transcritos de *A. sagei* en Trinity.

En el eje y se muestran los linajes analizados. Cada linaje posee una cantidad distinta de transcritos. En el eje x se muestra el porcentaje de integridad encontrado en dichas secuencias.

Aconaemys sagei annotation summary	
Gene	31,717
mRNA	31,717
CDS	191,040
Exón	192,894
Five_prime_UTR	5,926
Three_prime_UTR	7,149
Genes With GO name	Genes Without GO name
15,694	16,023

Tabla 11.3: Resumen de anotación del organismo *Aconaemys sagei*.

A la izquierda el elemento anotado seguido de la cantidad de veces encontrado en el archivo de anotación. En la zona inferior, se encuentra la cantidad de genes con y sin terminología GO luego de la anotación funcional

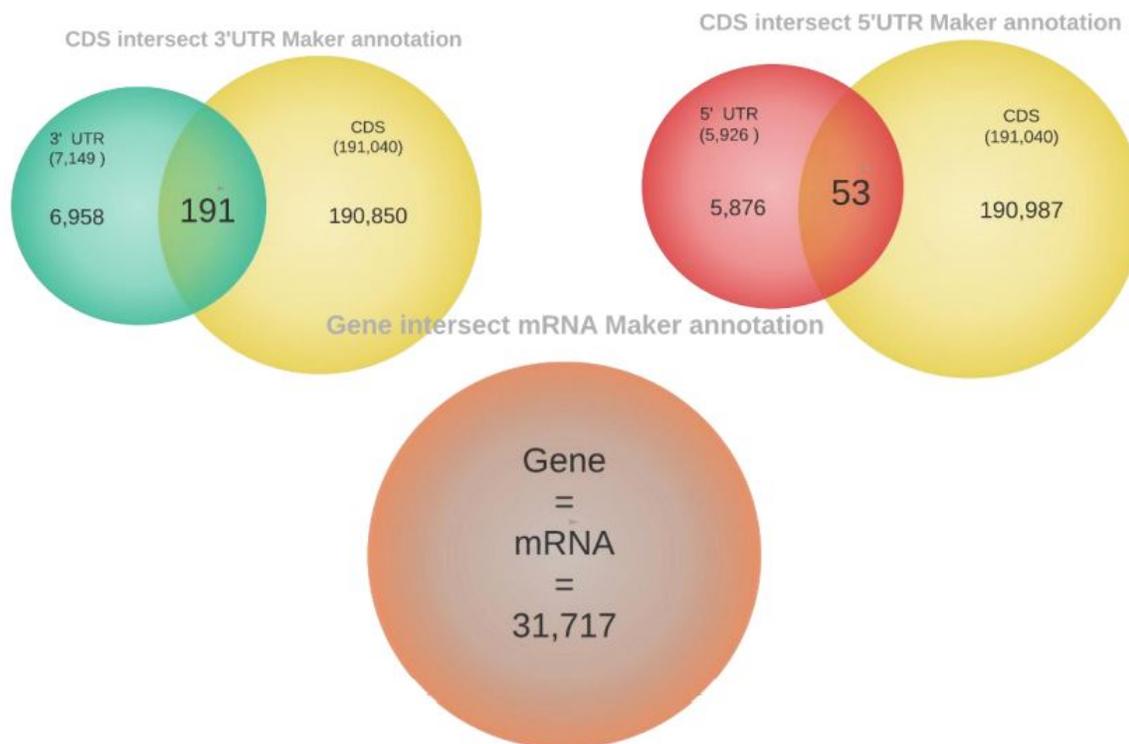


Figura 11.7: Diagramas de Venn en registros de elementos presentes en la anotación.

En la parte superior se encuentran de manera gráfica las intersecciones de CDS con elementos UTR. En la parte inferior se muestra la concordancia Gene-mRNA en la anotación.

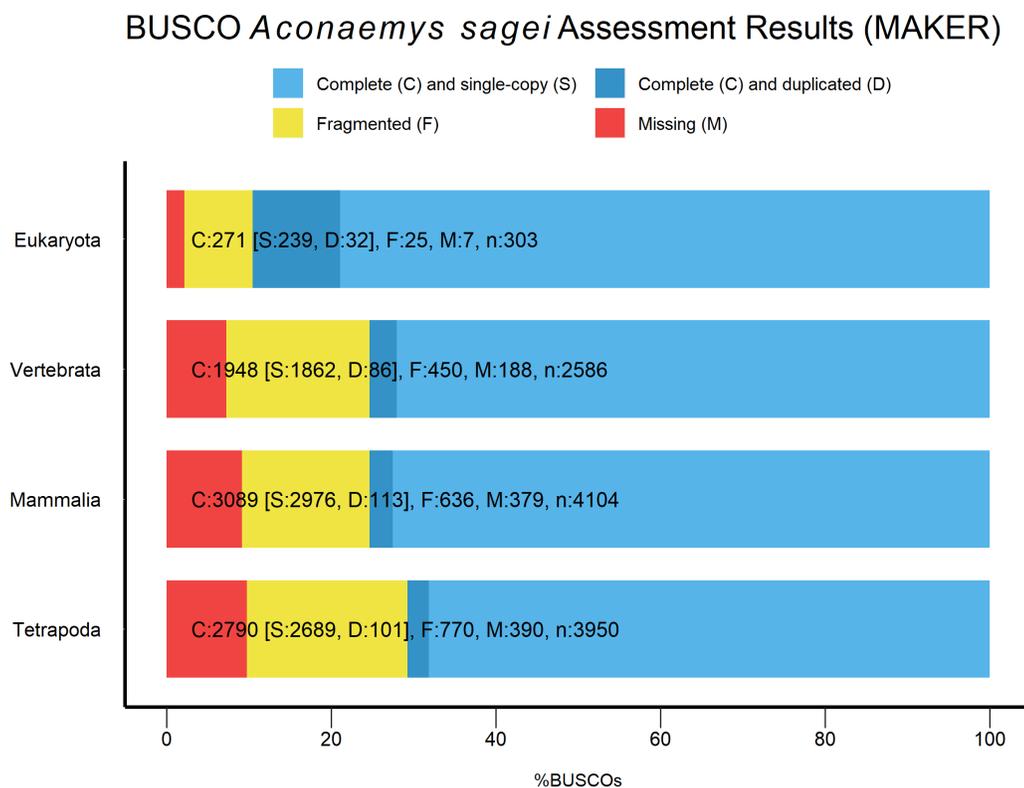


Figura 11.8: Resultados BUSCO del ensamble de transcritos de *Aconaemys sagei* en MAKER.

En el eje y se muestran los linajes analizados. Cada linaje posee una cantidad distinta de transcritos. En el eje x se muestra el porcentaje de integridad encontrado en dichas secuencias.

12. Discusión

El análisis de calidad realizado sobre las lecturas de DNA arrojó un promedio de puntaje superior a los 30 puntos de calidad en el sistema phred33. En el sistema phred33, la probabilidad de error de un nucleótido en el proceso de secuenciación es anotada en base a caracteres del sistema ASCII, de este modo cuando un nucleótido tiene una probabilidad de 1 entre 1,000 de ser secuenciada incorrectamente, y usando la siguiente fórmula:

$$Q = -10\log_{10}P$$

Donde P es la probabilidad de que la base sea incorrecta y Q corresponde al puntaje. De esta forma obtendremos un puntaje de 30, lo cual nos indica que la base tiene un 99,9 % de probabilidad de estar bien secuenciada.

Para todas las librerías de reads podemos observar puntajes sobre 30 de calidad, tanto para secuenciación de DNA y RNA. Sin embargo, existió una pequeña cantidad de reads en las librerías paired-end con presencia de contaminación por adaptador Nextera-Seq (Figura 11.1), proveniente de la técnica de secuenciación en librerías mate-pair. La eliminación de secuencias con Trimmomatic fue realizada de manera efectiva según FASTQC, lo que también mantuvo la calidad de los puntajes (Figura 11.2).

El alineamiento de lecturas de *A. sagei* sobre el genoma de *O. degus* resultó con un 84,3 % del total de reads alineando contra el genoma de referencia (Tabla 11.2). Dado que el tiempo estimado de divergencia entre ambas especies es de 3.69 ± 0.93 millones de años, durante el Plioceno [Opazo, 2005], es esperado que el mayor porcentaje de alineamiento sea dado por este individuo. Sin embargo, un 84 % de similitud entre ambas especies resulta relativamente bajo en comparación a especies con mayor tiempo de divergencia, como lo son el chimpancé y el humano con un tiempo estimado de 6.2 a 8.4 millones de años y con diferencias en el genoma cercana al 4 % [Luke and Verma, 1995]. El alineamiento de

secuencias nos permite de igual manera calcular la cobertura de profundidad real de la secuenciación (Figura 11.3), la cual resulta ser cercana a un 72x, lo que significa que en promedio una misma base fue secuenciada hasta 72 veces.

El resultado del ensamble (Figura 11.4) resulta muy similar a las características del genoma referenciado, obteniendo solamente 24 secuencias menos que este, es de asumir que dichas secuencias no fueron cubiertas por las lecturas de DNA, siendo desechadas por los software `bcftools` y `vcfutils` al momento de construir la secuencia consenso. En cuanto al N50, este mantiene un valor de 12,091,372 bp lo que es idéntico a las estadísticas del genoma de referencia de *O. degus* lo que indica que el ensamble no ha perdido la contigüidad durante la ejecución del ensamble por referencia. Su valor puede traducirse como el tamaño de la secuencia más corta si realizamos la sumatoria de todas las secuencias en un orden de mayor a menor tamaño hasta alcanzar el 50 % del tamaño total del ensamble, es decir 2,995,390,778 bp.

El gráfico de largo acumulado (Figura 11.5) nos indica que una gran parte de las secuencias ensambladas se encuentran entre los 1,000 y 10,000 pb. Se desprende también que existen algunas secuencias con largo entre los 10 y 40 Mbp; organismos roedores a nivel de cromosoma como *Mus musculus* poseen largos de secuencia superiores a los 61 Mbp, por lo que se puede entender que el ensamble generado aún se encuentra a nivel de scaffold, puesto que no se ha logrado concretar e identificar un cromosoma completo.

Los transcritos ensamblados con Trinity no lograron completar el perfil BUSCO de manera satisfactoria (Figura 11.6) para los linajes de tetrápodo, vertebrados y mamíferos; obteniendo porcentajes bajos, cercanos a un 40%, del total de transcritos BUSCO completos por cada uno de los linajes. Además, la proporción de transcritos BUSCO faltantes y fragmentados iban en aumento a medida que se era más específico en el linaje. Para solucionar esto, se consideró alinear los reads de transcriptoma contra el genoma de referencia y eliminar las lecturas no mapeadas del mismo modo de filtrado; para posteriormente reunir los reads integrados en el alineamiento y realizar el ensamble en ellos. A pesar de esto, se obtuvieron exactamente los mismos resultados. Los resultados de la anotación de *A. sagei* (Tabla 11.3) muestran una cantidad idéntica de genes y mRNA. Esto nos indica que por cada uno de los genes identificados existe un producto de mRNA asociado dentro de la anotación; sin embargo, no se indican las distintas variaciones que podría tener un gen a causa del splicing, lo que en consecuencia debería registrar una mayor cantidad de elementos mRNA en la anotación [Brett et al., 2002]. Por otro lado, las diferencias existentes entre CDS y exón son por causa de la identificación de exones. Un CDS se encuentra caracterizado por ser una sección codificante de un gen, es decir, sin

intrones ni regiones terminales; por otro lado, el exón es parte de la región codificante de un gen, sin embargo, este si puede ser parte de las regiones terminales o UTR. Esto quiere decir que tanto CDS, como 5' UTR y 3' UTR se encuentran dentro de las coordenadas de los exones.

Existe un pequeño porcentaje de secuencias cuya anotación pareciera ser inviable (Figura 11.7): se trata de aquellas regiones terminales, tanto 5' y 3', cuyas ubicaciones se encuentran en las mismas coordenadas que los CDS, lo cual no puede ser posible porque los CDS no corresponden a regiones terminales; un posterior análisis con el software bedtools [Quinlan and Hall, 2010](#) permitió identificar la coordenada de estos elementos y verificar que su anotación fue realizada en hebras opuestas.

En comparación con la anotación del genoma de referencia (Anexo 8), los elementos “gene” anotados en *A. sagei* parecen sobrepasar la cantidad del mismo elemento recopilado en la anotación de *O. degus*. Sin embargo, la cantidad de elementos “mRNA” de la referencia es mayor no tan solo a la cantidad del mismo elemento anotados en *A. sagei*, sino que también superior a los elementos “gene” de la misma referencia; esto sugiere que se identificaron las distintas variaciones genéticas, reflejada en la cantidad de transcritos. Por otro lado la referencia no recopila información de ontología genética (GO) en su anotación; sin embargo, al observar su anotación, esta sí presenta descripciones funcionales del gena recopiladas a partir de la base de datos RefSeq.

El análisis BUSCO de las secuencias encontradas con la predicción de genes ab initio, en conjunto con las evidencias alineadas provenientes del ensamblador Trinity y las secuencias de proteínas de *O. degus* (Figura 11.8), mejoraron en gran medida la integridad del set de transcriptomas generado en comparación con el ensamblado anteriormente, aumentando hasta un 70% de BUSCO's completos en casos más específicos. Esto era de esperar, puesto que no sólo se añadieron secuencias de predicción al set de transcritos, sino que también secuencias provenientes del alineamiento por tblastx de proteínas del organismo de referencia, completando de esta manera gran parte de los transcritos perdidos en el análisis del ensamble *de novo* generado por el ensamblador Trinity. Para finalizar, el set de transcritos generados para *A. sagei* resulta ser menos íntegro que el set de transcritos recopilado desde la referencia (Anexo 9) según el perfil BUSCO usado por cada linaje; sin embargo, el porcentaje de genes busco perdidos y fragmentados en conjunto no logra superar el 30% en cada linaje, lo cual pareciera ser un resultado satisfactorio para una primera aproximación al transcriptoma de *A.sagei*.

13. Conclusión

En esta memoria de título se obtuvieron los siguientes resultados:

1. El establecimiento de un modelo de referencia usando como base el tiempo de divergencia entre especies.
2. Una secuencia genómica generada a partir de una especie subterránea perteneciente a la familia *Octodontidae*.
3. Una primera anotación estructural y funcional del genoma de la rata de los pinares menor.

El resultado obtenido es una primera versión tanto del ensamble como de la anotación, la adición de información en base a nuevos datos de secuenciación podría ayudar a dilucidar las secciones desconocidas no presentes en el ensamble o registrar nuevos elementos durante el proceso de anotación. Si bien existen diferencias entre la cantidad de elementos anotados en comparación con la anotación del organismo de referencia, se logró localizar una gran cantidad de genes para su posterior análisis y estudio evolutivo. El campo del ensamble y anotación de genomas crece con el pasar de los años y a medida que surgen nuevas técnicas de secuenciación. En muchos casos más de un grupo por separado se dedica a anotar un mismo genoma, muchos de ellos usando procedimientos distintos e incluso ensamblajes diferentes. La recopilación y unión de toda esta información podría permitirnos esclarecer un panorama que cada vez parece más desconocido como lo es la estructura genómica de los seres vivos.

Bibliografía

- [Abril and Castellano, 2018] Abril, J. F. and Castellano, S. (2018). Genome annotation. In *Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics*, volume 1-3, pages 195–209. Elsevier.
- [Altschul et al., 1990] Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):403–410.
- [Andrews, 2010] Andrews, S. (2010). FASTQC. A quality control tool for high throughput sequence data.
- [Ashburner et al., 2000] Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., and Sherlock, G. (2000). Gene ontology: Tool for the unification of biology.
- [Bolger et al., 2014] Bolger, A. M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics (Oxford, England)*, 30(15):2114–20.
- [Bradnam et al., 2013] Bradnam, K. R., Fass, J. N., Alexandrov, A., Baranay, P., Bechner, M., Birol, I., Boisvert, S., Chapman, J. A., Chapuis, G., Chikhi, R., Chitsaz, H., Chou, W. C., Corbeil, J., Fabbro, C. D., Docking, R. R., Durbin, R., Earl, D., Emrich, S., Fedotov, P., Fonseca, N. A., Ganapathy, G., Gibbs, R. A., Gnerre, S., Godzaridis, É., Goldstein, S., Haimel, M., Hall, G., Haussler, D., Hiatt, J. B., Ho, I. Y., Howard, J., Hunt, M., Jackman, S. D., Jaffe, D. B., Jarvis, E. D., Jiang, H., Kazakov, S., Kersey, P. J., Kitzman, J. O., Knight, J. R., Koren, S., Lam, T. W., Lavenier, D., Laviolette, F., Li, Y., Li, Z., Liu, B., Liu, Y., Luo, R., MacCallum, I., MacManes, M. D., Maillet, N., Melnikov, S., Naquin, D., Ning, Z., Otto, T. D., Paten, B., Paulo, O. S., Phillippy,

- A. M., Pina-Martins, F., Place, M., Przybylski, D., Qin, X., Qu, C., Ribeiro, F. J., Richards, S., Rokhsar, D. S., Ruby, J. G., Scalabrin, S., Schatz, M. C., Schwartz, D. C., Sergushichev, A., Sharpe, T., Shaw, T. I., Shendure, J., Shi, Y., Simpson, J. T., Song, H., Tsarev, F., Vezzi, F., Vicedomini, R., Vieira, B. M., Wang, J., Worley, K. C., Yin, S., Yiu, S. M., Yuan, J., Zhang, G., Zhang, H., Zhou, S., and Korf, I. F. (2013). Assemblathon 2: Evaluating de novo methods of genome assembly in three vertebrate species. *GigaScience*, 2(1):10.
- [Brett et al., 2002] Brett, D., Pospisil, H., Valcárcel, J., Reich, J., and Bork, P. (2002). Alternative splicing and genome complexity. *Nature Genetics*, 30(1):29–30.
- [Bronner et al., 2014] Bronner, I. F., Quail, M. A., Turner, D. J., and Swerdlow, H. (2014). Improved protocols for Illumina sequencing. *Current Protocols in Human Genetics*, (SUPPL.80):18.2.1–18.2.42.
- [Brownlee and Sanger, 1967] Brownlee, G. and Sanger, F. (1967). Nucleotide sequences from the low molecular weight ribosomal RNA of *Escherichia coli*. *Journal of Molecular Biology*, 23(3):337–IN9.
- [Camacho et al., 2009] Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., and Madden, T. L. (2009). BLAST+: Architecture and applications. *BMC Bioinformatics*, 10.
- [Campbell et al., 2014] Campbell, M. S., Holt, C., Moore, B., and Yandell, M. (2014). Genome Annotation and Curation Using MAKER and MAKER-P. *Current Protocols in Bioinformatics*, 2014(1):4.11.1–4.11.39.
- [Cantarel et al., 2008] Cantarel, B. L., Korf, I., Robb, S. M., Parra, G., Ross, E., Moore, B., Holt, C., Alvarado, A. S., and Yandell, M. (2008). MAKER: An easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Research*, 18(1):188–196.
- [Carbon et al., 2019] Carbon, S., Douglass, E., Dunn, N., Good, B., Harris, N. L., Lewis, S. E., Mungall, C. J., Basu, S., Chisholm, R. L., Dodson, R. J., Hartline, E., Fey, P., Thomas, P. D., Albou, L. P., Ebert, D., Kesling, M. J., Mi, H., Muruganujan, A., Huang, X., Poudel, S., Mushayahama, T., Hu, J. C., LaBonte, S. A., Siegele, D. A., Antonazzo, G., Attrill, H., Brown, N. H., Fexova, S., Garapati, P., Jones, T. E., Marygold, S. J., Millburn, G. H., Rey, A. J., Trovisco, V., Dos Santos, G., Emmert, D. B., Falls, K., Zhou, P., Goodman, J. L., Strelets, V. B., Thurmond, J., Courtot, M., Osumi, D. S., Parkinson, H., Roncaglia, P., Acencio, M. L., Kuiper, M., Lreid,

- A., Logie, C., Lovering, R. C., Huntley, R. P., Denny, P., Campbell, N. H., Kramarz, B., Acquaah, V., Ahmad, S. H., Chen, H., Rawson, J. H., Chibucos, M. C., Giglio, M., Nadendla, S., Tauber, R., Duesbury, M. J., Del, N. T., Meldal, B. H., Perfetto, L., Porras, P., Orchard, S., Shrivastava, A., Xie, Z., Chang, H. Y., Finn, R. D., Mitchell, A. L., Rawlings, N. D., Richardson, L., Sangrador-Vegas, A., Blake, J. A., Christie, K. R., Dolan, M. E., Drabkin, H. J., Hill, D. P., Ni, L., Sitnikov, D., Harris, M. A., Oliver, S. G., Rutherford, K., Wood, V., Hayles, J., Bahler, J., Lock, A., Bolton, E. R., De Pons, J., Dwinell, M., Hayman, G. T., Laulederkind, S. J., Shimoyama, M., Tutaj, M., Wang, S. J., D'Eustachio, P., Matthews, L., Balhoff, J. P., Aleksander, S. A., Binkley, G., Dunn, B. L., Cherry, J. M., Engel, S. R., Gondwe, F., Karra, K., MacPherson, K. A., Miyasato, S. R., Nash, R. S., Ng, P. C., Sheppard, T. K., Shrivatsav Vp, A., Simison, M., Skrzypek, M. S., Weng, S., Wong, E. D., Feuermann, M., Gaudet, P., Bakker, E., Berardini, T. Z., Reiser, L., Subramaniam, S., Huala, E., Arighi, C., Auchincloss, A., Axelsen, K., Argoud, G. P., Bateman, A., Bely, B., Blatter, M. C., Boutet, E., Breuza, L., Bridge, A., Britto, R., Bye-A-Jee, H., Casals-Casas, C., Coudert, E., Estreicher, A., Famiglietti, L., Garmiri, P., Georghiou, G., Gos, A., Gruaz-Gumowski, N., Hatton-Ellis, E., Hinz, U., Hulo, C., Ignatchenko, A., Jungo, F., Keller, G., Laiho, K., Lemercier, P., Lieberherr, D., Lussi, Y., Mac-Dougall, A., Magrane, M., Martin, M. J., Masson, P., Natale, D. A., Hyka, N. N., Pedruzzi, I., Pichler, K., Poux, S., Rivoire, C., Rodriguez-Lopez, M., Sawford, T., Speretta, E., Shypitsyna, A., Stutz, A., Sundaram, S., Tognolli, M., Tyagi, N., Warner, K., Zaru, R., Wu, C., Chan, J., Cho, J., Gao, S., Grove, C., Harrison, M. C., Howe, K., Lee, R., Mendel, J., Muller, H. M., Raciti, D., Van Auken, K., Berriman, M., Stein, L., Sternberg, P. W., Howe, D., Toro, S., and Westerfield, M. (2019). The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Research*, 47(D1):D330–D338.
- [Danecek et al., 2011] Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., Handsaker, R. E., Lunter, G., Marth, G. T., Sherry, S. T., McVean, G., and Durbin, R. (2011). The variant call format and VCFtools. *Bioinformatics*, 27(15):2156–2158.
- [Elissamburu and Vizcaíno, 2004] Elissamburu, A. and Vizcaíno, S. F. (2004). Limb proportions and adaptations in caviomorph rodents (Rodentia: Caviomorpha). *Journal of Zoology*, 262(2):145–159.
- [Evans et al., 2017] Evans, B. J., Upham, N. S., Golding, G. B., Ojeda, R. A., and Ojeda, A. A. (2017). Evolution of the largest mammalian genome. *Genome Biology and Evolution*, 9(6):1711–1724.

- [Fiers et al., 1976] Fiers, W., Contreras, R., Duerinck, F., Haegeman, G., Iserentant, D., Merregaert, J., Min Jou, W., Molemans, F., Raeymaekers, A., Van den Berghe, A., Volckaert, G., and Ysebaert, M. (1976). Complete nucleotide sequence of bacteriophage MS2 RNA: Primary and secondary structure of the replicase gene. *Nature*, 260(5551):500–507.
- [Grabherr et al., 2011] Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., Chen, Z., Mauceli, E., Hacohen, N., Gnirke, A., Rhind, N., Di Palma, F., Birren, B. W., Nusbaum, C., Lindblad-Toh, K., Friedman, N., and Regev, A. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology*, 29(7):644–652.
- [Haas et al., 2013] Haas, B. J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P. D., Bowden, J., Couger, M. B., Eccles, D., Li, B., Lieber, M., Macmanes, M. D., Ott, M., Orvis, J., Pochet, N., Strozzi, F., Weeks, N., Westerman, R., William, T., Dewey, C. N., Henschel, R., Leduc, R. D., Friedman, N., and Regev, A. (2013). De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nature Protocols*, 8(8):1494–1512.
- [Heather and Chain, 2016] Heather, J. M. and Chain, B. (2016). The sequence of sequencers: The history of sequencing DNA.
- [Holt and Yandell, 2011] Holt, C. and Yandell, M. (2011). MAKER2: An annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics*, 12(1):491.
- [Huttlin et al., 2017] Huttlin, E. L., Bruckner, R. J., Paulo, J. A., Cannon, J. R., Ting, L., Baltier, K., Colby, G., Gebreab, F., Gygi, M. P., Parzen, H., Szpyt, J., Tam, S., Zarraga, G., Pontano-Vaites, L., Swarup, S., White, A. E., Schweppe, D. K., Rad, R., Erickson, B. K., Obar, R. A., Guruharsha, K. G., Li, K., Artavanis-Tsakonas, S., Gygi, S. P., and Wade Harper, J. (2017). Architecture of the human interactome defines protein communities and disease networks. *Nature*, 545(7655):505–509.
- [Jones et al., 2014] Jones, P., Binns, D., Chang, H. Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H., Maslen, J., Mitchell, A., Nuka, G., Pesseat, S., Quinn, A. F., Sangrador-Vegas, A., Scheremetjew, M., Yong, S. Y., Lopez, R., and Hunter, S. (2014). InterProScan 5: Genome-scale protein function classification. *Bioinformatics*, 30(9):1236–1240.

- [Khan et al., 2018] Khan, A. R., Pervez, M. T., Babar, M. E., Naveed, N., and Shoaib, M. (2018). A Comprehensive Study of De Novo Genome Assemblers: Current Challenges and Future Prospective. *Evolutionary Bioinformatics*, 14:1176934318758650.
- [Kitts et al., 2016] Kitts, P. A., Church, D. M., Thibaud-Nissen, F., Choi, J., Hem, V., Sapojnikov, V., Smith, R. G., Tatusova, T., Xiang, C., Zherikov, A., DiCuccio, M., Murphy, T. D., Pruitt, K. D., and Kimchi, A. (2016). Assembly: A resource for assembled genomes at NCBI. *Nucleic Acids Research*, 44(D1):D73–D80.
- [Korf, 2004] Korf, I. (2004). Gene finding in novel genomes. *BMC Bioinformatics*, 5(1):59.
- [Langmead and Salzberg, 2012] Langmead, B. and Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, 9(4):357–359.
- [Lessa et al., 2008] Lessa, E. P., Vassallo, A. I., Verzi, D. H., and Mora, M. S. (2008). Evolution of morphological adaptations for digging in living and extinct ctenomyid and octodontid rodents. *Biological Journal of the Linnean Society*, 95(2):267–283.
- [Li and Barrett, 2011] Li, H. and Barrett, J. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. 27(21):2987–2993.
- [Li et al., 2009] Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16):2078–2079.
- [Lischer and Shimizu, 2017] Lischer, H. E. and Shimizu, K. K. (2017). Reference-guided de novo assembly approach improves genome reconstruction for related species. *BMC Bioinformatics*, 18(1):474.
- [Losos, 2010] Losos, J. B. (2010). Adaptive radiation, ecological opportunity, and evolutionary determinism : American society of naturalists E. O. Wilson award address. *American Naturalist*, 175(6):623–639.
- [Luke and Verma, 1995] Luke, S. and Verma, R. S. (1995). Human (*Homo sapiens*) and chimpanzee (*Pan troglodytes*) share similar ancestral centromeric alpha satellite DNA sequences but other fractions of heterochromatin differ considerably. *American Journal of Physical Anthropology*, 96(1):63–71.
- [Manov et al., 2013] Manov, I., Hirsh, M., Iancu, T. C., Malik, A., Sotnichenko, N., Band, M., Avivi, A., and Shams, I. (2013). Pronounced cancer resistance in a subterranean rodent, the blind mole-rat, *Spalax*: In vivo and in vitro evidence. *BMC Biology*, 11:91.

- [Miller et al., 2010] Miller, J. R., Koren, S., and Sutton, G. (2010). Assembly algorithms for next-generation sequencing data.
- [Morgan and Verzi, 2006] Morgan, C. C. and Verzi, D. H. (2006). MORPHOLOGICAL DIVERSITY OF THE HUMERUS OF THE SOUTH AMERICAN SUBTERRANEAN RODENT CTENOMYS (RODENTIA, CTENOMYIDAE). *Journal of Mammalogy*, 87(6):1252–1260.
- [Morgan and Verzi, 2011] Morgan, C. C. and Verzi, D. H. (2011). Carpal-metacarpal specializations for burrowing in South American octodontoid rodents. *Journal of Anatomy*, 219(2):167–175.
- [Mullis and Faloona, 1987] Mullis, K. B. and Faloona, F. A. (1987). Specific Synthesis of DNA in Vitro via a Polymerase-Catalyzed Chain Reaction. *Methods in Enzymology*, 155(C):335–350.
- [Nevo, 1979] Nevo, E. (1979). Adaptive Convergence and Divergence of Subterranean Mammals. *Annual Review of Ecology and Systematics*, 10(1):269–308.
- [Novikov et al., 2015] Novikov, E., Kondratyuk, E., Petrovski, D., Titova, T., Zadubrovskaya, I., Zadubrovskiy, P., and Moshkin, M. (2015). Reproduction, aging and mortality rate in social subterranean mole voles (*Ellobius talpinus* Pall.). *Biogerontology*, 16(6):723–732.
- [Olofsson et al., 2019] Olofsson, J. K., Cantera, I., Van de Paer, C., Hong-Wa, C., Zedane, L., Dunning, L. T., Alberti, A., Christin, P. A., and Besnard, G. (2019). Phylogenomics using low-depth whole genome sequencing: A case study with the olive tribe. *Molecular Ecology Resources*, 19(4):877–892.
- [Opazo, 2005] Opazo, J. C. (2005). A molecular timescale for caviomorph rodents (Mammalia, Hystricognathi). *Molecular Phylogenetics and Evolution*, 37(3):932–937.
- [Parededa and Novello, 2012] Parededa, M. G. and Novello, A. (2012). Chromosome mosaicism: Extreme karyotype variation in the genus *ctenomys* from Uruguay (Rodentia: Ctenomyidae). *Caryologia*, 65(4):251–257.
- [Pevzner et al., 2001] Pevzner, P. A., Tang, H., and Waterman, M. S. (2001). An Eulerian path approach to DNA fragment assembly. *Proceedings of the National Academy of Sciences of the United States of America*, 98(17):9748–9753.
- [Quinlan and Hall, 2010] Quinlan, A. R. and Hall, I. M. (2010). BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–842.

- [Ruiz, 2011] Ruiz, M. (2011). *Análisis morfológico del aparato excavador en roedores subterráneos del género Ctenomys Blainville, 1826 (Rodentia: Ctenomyidae)*. PhD thesis.
- [Rust et al., 2002] Rust, A. G., Mongin, E., and Birney, E. (2002). Genome annotation techniques: New approaches and challenges.
- [Sayers et al., 2020] Sayers, E. W., Beck, J., Brister, J. R., Bolton, E. E., Canese, K., Comeau, D. C., Funk, K., Ketter, A., Kim, S., Kimchi, A., Kitts, P. A., Kuznetsov, A., Lathrop, S., Lu, Z., McGarvey, K., Madden, T. L., Murphy, T. D., O’Leary, N., Phan, L., Schneider, V. A., Thibaud-Nissen, F., Trawick, B. W., Pruitt, K. D., and Ostell, J. (2020). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research*, 48(D1):D9–D16.
- [Scalzitti et al., 2020] Scalzitti, N., Jeannin-Girardon, A., Collet, P., Poch, O., and Thompson, J. D. (2020). A benchmark study of ab initio gene prediction methods in diverse eukaryotic organisms. *BMC Genomics*, 21(1).
- [Seppey et al., 2019] Seppey, M., Manni, M., and Zdobnov, E. M. (2019). BUSCO: Assessing genome assembly and annotation completeness. In *Methods in Molecular Biology*, volume 1962, pages 227–245. Humana Press Inc.
- [Slater and Birney, 2005] Slater, G. S. C. and Birney, E. (2005). Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics*, 6.
- [Smit et al., 2013] Smit, A., Hubley, R., and Green, P. (2013). RepeatMasker Open-4.0.
- [Sohn and Nam, 2018] Sohn, J. I. and Nam, J. W. (2018). The present and future of de novo whole-genome assembly. *Briefings in Bioinformatics*, 19(1):23–40.
- [Stanke and Waack, 2003] Stanke, M. and Waack, S. (2003). Gene prediction with a hidden Markov model and a new intron submodel. In *Bioinformatics*, volume 19.
- [Tomasco and Lessa, 2010] Tomasco, M. I. H. and Lessa, O. E. P. (2010). Adaptaciones a la hipoxia e hipercapnia del nicho subterráneo en roedores octodontoideos .
- [Watson and Crick, 1953] Watson, J. D. and Crick, F. H. (1953). Molecular structure of nucleic acids: A structure for deoxyribose nucleic acid. *Nature*, 171(4356):737–738.
- [Yandell and Ence, 2012] Yandell, M. and Ence, D. (2012). A beginner’s guide to eukaryotic genome annotation.

- [Yoshinaga et al., 2018] Yoshinaga, Y., Daum, C., He, G., and O'Malley, R. (2018). Genome sequencing. In *Methods in Molecular Biology*, volume 1775, pages 37–52. Humana Press Inc.
- [Zhang et al., 2011] Zhang, J., Chiodini, R., Badr, A., and Zhang, G. (2011). The impact of next-generation sequencing on genomics.

14. Anexos

14.1. Anexo 1.- Comandos de ejecución de programas

14.1.1. FASTQC

```
$ /opt/apps/fastqc/fastqc paired_end_file_R1.fastq # comando de uso FASTQC
```

```
$ nohup /opt/apps/fastqc/fastqc \*.fastq \&
```

```
# Permite ejecutar múltiples análisis fastqc en segundo plano
```

14.1.2. TRINITY

```
$ source conda activate
```

```
#Activamos el entorno virtual conda que posea el módulo salmon
```

```
$ Trinity --seqType fq --left ../rename_reads1.fq --right ../rename_reads2.fq \  
--CPU 64 --max_memory 150
```

```
# --seq_type indica el tipo de archivo, --left/right indican  
la ruta de los archivos de lectura paired-end,  
--CPU para indicar la cantidad de CPU's y {max_memory  
para delimitar la cantidad máxima de memoria RAM para su ejecución
```

14.2. Anexo 2.- Script de ejecución de Trimmomatic

```
#!/bin/bash

pwd=$PWD

#for LIB in 'L004' 'L006' 'L007';

#for LIB in 'L007';

for LIB in 'L004' 'L006';

do

    case $LIB in

        L004)
            ID='-FML779_161005_L004_'
            ;;

        L006)
            ID='-FML779_160930_L006_'
            ;;

        L007)
            ID='-FML779_161005_L007_'
            ;;

        *)
            echo 'Files not found'
            ;;

    esac

    for insert in '5kb' '8kb' '12kb';
do
```

```

for type in 'R0' 'R1';
do

    if [ $type = 'R0' ];then

        java -jar /home/apps/trimmomatic-0.36/trimmomatic-0.36.jar SE -phred33
        "/home/cfierro/genoma1/2_"${insert}${ID}${type}".fastq"
        "2_"${insert}${ID}${type}"_trimmed.fastq"
        ILLUMINACLIP:/home/apps/trimmomatic-0.36/adapters/NexteraPE-PE.fa:2:30:10

    else

        java -jar /home/apps/trimmomatic-0.36/trimmomatic-0.36.jar PE -phred33
        "/home/cfierro/genoma1/2_"${insert}${ID}${type}".fastq"
        "/home/cfierro/genoma1/2_"${insert}${ID}${type}"R2.fastq"
        "2_"${insert}${ID}${type}"_trimmed_paired.fastq"
        "2_"${insert}${ID}${type}"_trimmed_unpaired.fastq"
        "2_"${insert}${ID}${type}"R2_trimmed_paired.fastq"
        "2_"${insert}${ID}${type}"R2_trimmed_unpaired.fastq"
        ILLUMINACLIP:/home/apps/trimmomatic-0.36/adapters/NexteraPE-PE.fa:2:30:10

    fi

done

done

done

```

14.3. Anexo 3.- Comandos de ensamble por referencia

```
$ bowtie2-build genoma_octodon.fasta Octodon
```

```
$ bowtie2 -x Octodon \  
-1 2_350bp-FML779_160930_L006_R1.fastq,\ \  
2_8kb-FML779_161005_L007_R1.fastq,    \  
2_550bp-FML779_161005_L004_R1.fastq,   \  
2_12kb-FML779_160930_L006_R1.fastq,    \  
2_5kb-FML779_161005_L007_R1.fastq,     \  
2_350bp-FML779_161005_L004_R1.fastq,   \  
2_550bp-FML779_161005_L007_R1.fastq,   \  
2_12kb-FML779_161005_L004_R1.fastq,    \  
2_8kb-FML779_160930_L006_R1.fastq,     \  
2_350bp-FML779_161005_L007_R1.fastq,   \  
2_5kb-FML779_160930_L006_R1.fastq,     \  
2_12kb-FML779_161005_L007_R1.fastq,    \  
2_8kb-FML779_161005_L004_R1.fastq,     \  
2_550bp-FML779_160930_L006_R1.fastq    \  
-2 2_350bp-FML779_160930_L006_R2.fastq,\ \  
2_8kb-FML779_161005_L007_R2.fastq,    \  
2_550bp-FML779_161005_L004_R2.fastq,   \  
2_12kb-FML779_160930_L006_R2.fastq,    \  
2_5kb-FML779_161005_L007_R2.fastq,     \  
2_350bp-FML779_161005_L004_R2.fastq,   \  
2_550bp-FML779_161005_L007_R2.fastq,   \  
2_12kb-FML779_161005_L004_R2.fastq,    \  
2_8kb-FML779_160930_L006_R2.fastq,     \  
2_350bp-FML779_161005_L007_R2.fastq,   \  
2_5kb-FML779_160930_L006_R2.fastq,     \  
2_12kb-FML779_161005_L007_R2.fastq,    \  
2_8kb-FML779_161005_L004_R2.fastq,     \  
2_550bp-FML779_160930_L006_R2.fastq    \  
-S Octodon_align.sam  
  
$ samtools view -bS Octodon_align.sam -o Octodon_align.bam  
  
$ samtools sort Octodon_align.bam -o Octodon_align.sort.bam  
  
$ samtools index Octodon_align.sort.bam
```

```
$ samtools mpileup -uf genoma_octodon.fa Octodon_align.sort.bam | \
bcftools view -cg - | \
vcfutils.pl vcf2fq > asagei_consensus.fq
```

```
$ seqret -osformat fasta asagei_consensus.fq -out2 Genoma_aconaemys.fasta
```

14.4. Anexo 4.- Archivo de configuración maker_exe.ctl

```
#-----Location of Executables Used by MAKER/EVALUATOR

makeblastdb=/usr/bin/makeblastdb #location of NCBI+ makeblastdb executable

blastn=/usr/bin/blastn #location of NCBI+ blastn executable

blastx=/usr/bin/blastx #location of NCBI+ blastx executable

tblastx=/usr/bin/tblastx #location of NCBI+ tblastx executable

formatdb=/usr/bin/formatdb #location of NCBI formatdb executable

blastall=/usr/bin/blastall #location of NCBI blastall executable

xdformat= #location of WUBLAST xdformat executable

blasta= #location of WUBLAST blasta executable

RepeatMasker=/home/cfierro/src/RepeatMasker/RepeatMasker

#location of RepeatMasker executable

exonerate=/usr/bin/exonerate #location of exonerate executable

#-----Ab-initio Gene Prediction Algorithms
```

```
snap=/home/cfierro/src/snap/Zoe/snap #location of snap executable
```

```
gmhmm3= #location of eukaryotic genemark executable
```

```
gmhmmmp= #location of prokaryotic genemark executable
```

```
augustus=/home/cfierro/src/augustus.2.5.5/bin/augustus
```

```
#location of augustus executable
```

```
fgenesh= #location of fgenesh executable
```

```
tRNAscan-SE= #location of trnascan executable
```

```
snoscan= #location of snoscan executable
```

```
#-----Other Algorithms
```

```
probuild= #location of probuild executable (required for genemark)
```

14.5. Anexo 5.- Archivo de configuración maker_boptsctl

```
#-----BLAST and Exonerate Statistics Thresholds
```

```
blast_type=ncbi+ #set to 'ncbi+', 'ncbi' or 'wublast'
```

```
pcov_blastn=0.8 #Blastn Percent Coverage Threshold EST-Genome Alignments
```

```
pid_blastn=0.85 #Blastn Percent Identity Threshold EST-Genome Aligments
```

```
eval_blastn=1e-10 #Blastn eval cutoff
```

```
bit_blastn=40 #Blastn bit cutoff
```

```
depth_blastn=0 #Blastn depth cutoff (0 to disable cutoff)

pcov_blastx=0.5 #Blastx Percent Coverage Threshold Protein-Genome Alignments

pid_blastx=0.4 #Blastx Percent Identity Threshold Protein-Genome Aligments

eval_blastx=1e-06 #Blastx eval cutoff

bit_blastx=30 #Blastx bit cutoff

depth_blastx=0 #Blastx depth cutoff (0 to disable cutoff)

pcov_tblastx=0.8 #tBlastx Percent Coverage Threshold alt-EST-Genome Alignments

pid_tblastx=0.85 #tBlastx Percent Identity Threshold alt-EST-Genome Aligments

eval_tblastx=1e-10 #tBlastx eval cutoff

bit_tblastx=40 #tBlastx bit cutoff

depth_tblastx=0 #tBlastx depth cutoff (0 to disable cutoff)

pcov_rm_blastx=0.5
#Blastx Percent Coverage Threshold For Transposable Element Masking

pid_rm_blastx=0.4
#Blastx Percent Identity Threshold For Transposbale Element Masking

eval_rm_blastx=1e-06 #Blastx eval cutoff for transposable element masking

bit_rm_blastx=30 #Blastx bit cutoff for transposable element masking

ep_score_limit=20 #Exonerate protein percent of maximal score threshold

en_score_limit=20 #Exonerate nucleotide percent of maximal score threshold
```

14.6. Anexo 6.- Archivo de configuración maker_opts.ctl

```
#-----Genome (these are always required)

genome=/home/cfierro/assembly/Reference/Octodon_refAssembly/Genoma_Aconaemys.fasta
#genome sequence (fasta file or fasta embeded in GFF3 file)

organism_type=eukaryotic #eukaryotic or prokaryotic. Default is eukaryotic

#-----Re-annotation Using MAKER Derived GFF3

maker_gff=/home/cfierro/annotation_real/01.Genoma_Aconaemys.maker.output
#MAKER derived GFF3 file

est_pass=0 #use ESTs in maker_gff: 1 = yes, 0 = no

altest_pass=0 #use alternate organism ESTs in maker_gff: 1 = yes, 0 = no

protein_pass=0 #use protein alignments in maker_gff: 1 = yes, 0 = no

rm_pass=0 #use repeats in maker_gff: 1 = yes, 0 = no

model_pass=0 #use gene models in maker_gff: 1 = yes, 0 = no

pred_pass=0 #use ab-initio predictions in maker_gff: 1 = yes, 0 = no

other_pass=0 #passthrough anything else in maker_gff: 1 = yes, 0 = no

#-----EST Evidence (for best results provide a file for at least one)

est=/home/cfierro/transcriptoma/assembly/trinity_out_dir/Trinity.fasta
#set of ESTs or assembled mRNA-seq in fasta format

altest= #EST/cDNA sequence file in fasta format from an alternate organism

est_gff= #aligned ESTs or mRNA-seq from an external GFF3 file
```

```
altest_gff= #aligned ESTs from a closely related species in GFF3 format

#-----Protein Homology Evidence (for best results provide a file for at least one)

protein=/home/cfierro/test_maker/o_degus/protein.fa
#protein sequence file in fasta format (i.e. from multiple organisms)

protein_gff= #aligned protein homology evidence from an external GFF3 file

#-----Repeat Masking (leave values blank to skip repeat masking)

model_org=all #select a model organism for RepBase masking in RepeatMasker

rmlib=
#provide an organism specific repeat library in fasta format for RepeatMasker

repeat_protein=/home/cfierro/src/maker/data/te_proteins.fasta #provide a fasta file of

rm_gff= #pre-identified repeat elements from an external GFF3 file

prok_rm=0
#forces MAKER to repeatmask prokaryotes
#(no reason to change this), 1 = yes, 0 = no

softmask=1
#use soft-masking rather than hard-masking in BLAST
#(i.e. seg and dust filtering)

#-----Gene Prediction

snaphmm=/home/cfierro/test_maker/o_degus/ode10.hmm #SNAP HMM file

gmhmm= #GeneMark HMM file
```

```
augustus_species=human #Augustus gene prediction species model

fgenesh_par_file= #FGENESH parameter file

pred_gff=/home/cfierro/test_maker/o_degus/ref_OctDeg1.0_gnomon_top_level.gff3
#ab-initio predictions from an external GFF3 file

model_gff=
#annotated gene models from an external GFF3 file
#(annotation pass-through)

est2genome=0 #infer gene predictions directly from ESTs, 1 = yes, 0 = no

protein2genome=0 #infer predictions from protein homology, 1 = yes, 0 = no

trna=0 #find tRNAs with tRNAscan, 1 = yes, 0 = no

snoscan_rrna= #rRNA file to have Snoscan find snoRNAs

unmask=0
#also run ab-initio prediction programs
#on unmasked sequence, 1 = yes, 0 = no

#-----Other Annotation Feature Types (features MAKER doesn't recognize)

other_gff= #extra features to pass-through to final MAKER generated GFF3 file

#-----External Application Behavior Options

alt_peptide=C
#amino acid used to replace non-standard amino acids in BLAST databases

cpus=10
#max number of cpus to use in BLAST and RepeatMasker
#(not for MPI, leave 1 when using MPI)
```

```
#-----MAKER Behavior Options

max_dna_len=1000000
#length for dividing up contigs into chunks
#(increases/decreases memory usage)

min_contig=1 #skip genome contigs below this length (under 10kb are often useless)

pred_flank=200 #flank for extending evidence clusters sent to gene predictors

pred_stats=0 #report AED and QI statistics for all predictions as well as models

AED_threshold=1 #Maximum Annotation Edit Distance allowed (bound by 0 and 1)

min_protein=0 #require at least this many amino acids in predicted proteins

alt_splice=0
#Take extra steps to try and find alternative splicing, 1 = yes, 0 = no

always_complete=0 #extra steps to force start and stop codons, 1 = yes, 0 = no

map_forward=0
#map names and attributes forward from old
#GFF3 genes, 1 = yes, 0 = no

keep_preds=0
#Concordance threshold to add unsupported
#gene prediction (bound by 0 and 1)

split_hit=10000
#length for the splitting of hits
#(expected max intron size for evidence alignments)

single_exon=0
#consider single exon EST evidence when
#generating annotations, 1 = yes, 0 = no
```

```

single_length=250
#min length required for single exon ESTs if 'single_exon is enabled'

correct_est_fusion=0 #limits use of ESTs in annotation to avoid fusion genes

tries=6 #number of times to try a contig if there is a failure for some reason

clean_try=1
#remove all data from previous run before retrying, 1 = yes, 0 = no

clean_up=1
#removes theVoid directory with individual analysis files, 1 = yes, 0 = no

TMP=
#specify a directory other than the system default
#temporary directory for temporary files

```

14.7. Anexo 7.- Comandos anotación funcional

```

$ /home/cfierro/src/maker/bin/gff3_merge -l -n -g -d \
Genoma_Aconaemys_master_datastore_index.log -o \
Genoma_aconaemys.all.maker.gff

$ /home/cfierro/src/maker/bin/fasta_merge -d \
Genoma_Aconaemys_master_datastore_index.log \
-o Genoma_aconaemys

$ ab-blastp -query Genoma_aconaemys.all.maker.proteins.fasta \
-mformat=2 -db uniprot_sprot -evalue 1e-6 hspmax 1 -cpus 64

$ interproscan.sh -appl pfam -dp -f TSV -goterms -iprlookup \
-pa -t p -i Genoma_aconaemys.all.maker.proteins.fasta -o \
Genoma_aconaemys.all.maker.proteins.fasta.iprscan

```

```
$ maker_map_ids --prefix GMOD_ --justify 8 Genoma_aconaemys.all.gff > map
```

```
$ map_gff_ids map Genoma_aconaemys.all.gff
```

```
$ map_fasta_ids map Genoma_aconaemys.all.maker.proteins.fasta
```

```
$ map_fasta_ids map Genoma_aconaemys.all.maker.transcripts.fasta
```

```
$ map_data_ids map Genoma_aconaemys.all.maker.proteins.fasta.blastp
```

```
$ map_data_ids map Genoma_aconaemys.all.maker.proteins.fasta.iprscan
```

```
$ ipr_update_gff Genoma_aconaemys.all.gff          \  
Genoma_aconaemys.all.maker.proteins.fasta.iprscan > \  
Genoma_aconaemys.all.2.gff
```

```
$ iprscan2gff3 Genoma_aconaemys.all.maker.proteins.fasta.iprscan \  
Genoma_aconaemys.all.gff >>                                     \  
Genoma_aconaemys.all.2.gff
```

```
$ mv Genoma_aconaemys.all.2.gff Genoma_aconaemys.all.gff
```

```
$ maker_functional_gff uniprot_sprot.fasta          \  
Genoma_aconaemys.all.maker.proteins.fasta.blastp \  
Genoma_aconaemys.all.gff > Genoma_aconaemys.all.2.gff
```

```
$ maker_functional_fasta uniprot_sprot.fasta        \  
Genoma_aconaemys.all.maker.proteins.fasta.blastp \  
Genoma_aconaemys.all.maker.proteins.fasta >       \  
Genoma_aconaemys.all.maker.proteins.2.fasta
```

```
$ maker_functional_fasta uniprot_sprot.fasta        \  
Genoma_aconaemys.all.maker.proteins.fasta.blastp \  
Genoma_aconaemys.all.maker.transcripts.fasta >    \  
Genoma_aconaemys.all.maker.proteins.2.fasta
```

```
$ mv Genoma_aconaemys.all.2.gff Genoma_aconaemys.all.gff
```

```
$ mv Genoma_aconaemys.all.maker.proteins.2.fasta \  
Genoma_aconaemys.all.maker.proteins.fasta
```

```
$ mv Genoma_aconaemys.all.maker.proteins.fasta \  
Genoma_aconaemys.all.maker.proteins.2.fasta
```

14.8. Anexo 8.- Información preliminar de *Octodon degus*

<i>Octodon degus</i> annotation summary	
Gene	25,217
mRNA	42,410
CDS	492,458
Exón	560,210
transcript	5,926
pseudogene	7,149

14.9. Anexo 9.- Análisis BUSCO en *Octodon degus*