Tesis para obtener el Título de Magíster en Bioinformática

Programa de Desarrollo de las Ciencias Básicas

Universidad de la República

Caracterización del Transcriptoma de Riñón del Ratón Oliváceo Sudamericano *Abrothrix olivacea*

Lic. Facundo Giorello

Tutores

Dr. Enrique Lessa y Dr. Ricardo Fraiman

Tribunal

Dr. Hugo Naya, Dr. Gustavo Guerberoff y Dr. Héctor Romero.

Setiembre 2014, Montevideo, Uruguay.

AGRADECIMIENTOS

En primer lugar agradezco a mis padres; sin su apoyo no hubiese sido posible ni siquiera concebir esta nueva etapa.

Agradezco a los doctores Enrique Lessa y Ricardo Fraiman por haberme aceptado como maestreando y permitirme así avanzar en la carrera.

A la Agencia Nacional de Investigación e Innovación por su apoyo económico; hay indicios de vida después de la ciencia.

A todos aquellos que su ayuda, colaboración y trabajo fue imprescindible para que este trabajo empezara.

Por innumerables razones, muchas gracias Matías. Fuiste un excelente compañero para ir aprendiendo sobre la marcha, para compartir y repartir responsabilidades, por entender y tolerar las vueltas en círculo que me caracterizan. En definitiva, fuiste fundamental para darle forma a este trabajo y para cristalizarlo.

A la densa (tan así que a veces era necesario abrir un poco la ventana) pero jocosa atmósfera de la oficina, fruto de la convivencia junto a tres (y ahora cuatro con la llegada de "el Parada") *Homo sapiens* machos adultos: Alejandro, Matías y "el Calvello". Obviamente, gracias a Ciro por sus continuas visitas y contribuir también con ella.

Gracias nuevamente Enrique, por la confianza y libertad otorgada.

Gracias Daniel, por tus consejos, experiencia y aportes.

Muchas gracias a mi queridísima, por saber soportarme desanimado y más aún entusiasmado.

CONTENIDO

RESUMEN

El ratón oliváceo, *Abrothrix olivacea* (Waterhouse 1837), habita un gran variedad de ambientes en la región austral de América del Sur y cuenta a lo largo de su gran distribución con numerosas subespecies. Dada su capacidad de establecer y habitar ambientes notoriamente contrastantes, el ratón oliváceo se presenta como una especie interesante para estudiar la variación geográfica en respuesta a la variación ambiental. Dada la gran diferencia en precipitación que existe entre algunas regiones en las que habita, en esta tesis se caracteriza el transcriptoma de riñón de *A. olivacea*. Con el propósito de establecer un transcriptoma renal de referencia, trabajamos con 13 individuos de cuatro puntos geográficos distintos, cubriendo así en gran medida la distribución del ratón oliváceo y evaluamos la capacidad de ensamblado de varias estrategias. Por otro lado, caracterizamos los genes de más alta expresión entre los 13 individuos y los comparamos con los descriptos en el ratón doméstico.

Entre las distintas estrategias de ensamblado, constatamos que aquellas que involucran un algoritmo de normalización, TrinityNorm y DigiNorm, a efectos de eliminar las lecturas (*reads*) redundantes, produjeron ensamblados mucho más fragmentados pero con un número mayor de genes que la estrategia Multireads. Entre TrinityNorm y DigiNorm, el primero mostró un mejor balance en cuanto a genes hallados, nivel de reconstrucción de las secuencias codificantes y tiempo de ensamblado. En general, obtuvimos secuencias codificantes pertenecientes a más de 15.000 genes de los cuales 10.000, aproximadamente, tienen al menos una secuencia codificante reconstruida en un 90% o más. De los 283 genes que se encontraban en el 5% más expresado entre los 13 individuos de *A. olivacea*, encontramos 17 en común con los descriptos en la parte distal del túbulo renal de ratón.

El número de secuencias codificantes de *A. olivacea* reconstruidas fue muy similar a los reportados para otras especies de mamíferos, lo que indica la calidad de nuestros ensamblados. Respecto a las estrategias para establecer un transcriptoma de referencia, la normalización debe ser evitada si se necesita reconstruir toda o casi toda la secuencia de los transcriptos. Para los casos donde el número de lecturas sea tal que impida seguir una estrategia Multireads, nuestros resultados sugieren que la estrategia TrinityNorm es más conveniente que DigiNorm. Dado que este es el primer transcriptoma de un miembro de la subfamilia Sigmodontinae la cual contiene más de 400 especies, creemos que este trabajo ayudará a encaminar futuros estudios ecológicos y evolutivos en *A. olivacea* y en otras especies.

INTRODUCCIÓN

CARACTERÍSTICAS E IMPORTANCIA DE ABROTHRIX OLIVACEA COMO MODELO DE ESTUDIO.

Uno de los principales objetivos de la biología evolutiva es tratar de entender cómo se da la diversificación de los organismos, los procesos implicados y las causas. La diversificación alopátrica y el rol posterior de la deriva génica han sido una de las explicaciones más favorecidas históricamente. Más recientemente, fenómenos como la plasticidad fenotípica han cobrado importancia como desencadenantes de la diversificación y especiación de los organismos presentes en distintas condiciones ambientales (Pfennig et al. 2010). En particular, se ha estado discutiendo su relevancia en la especiación ecológica (Thibert-Plante & Hendry 2011; Nosil 2012). Varios de estos procesos y fenómenos, y su posible papel en la especiación pueden ser abordados adecuadamente utilizando a *Abrothix olivacea* como especie modelo y haciendo uso de las nuevas tecnologías de secuenciación.

El ratón oliváceo, *Abrothrix olivacea* (Waterhouse 1837) es un roedor de la subfamilia Sigmodontinae, una de las más grandes dentro de los mamíferos (86 géneros, y unas 400 especies). Se encuentra ampliamente distribuido en la porción austral de América del Sur, extendiéndose por casi toda la Patagonia Argentina y Chilena hasta Tierra del Fuego (Mann 1978), y habita una gran variedad de ambientes, desde desiertos costeros al norte, bosques Valdivianos y Magallánicos al sur y a lo largo de Chile, hasta la estepa Patagónica al este y sur de la Argentina. Al menos 6 subespecies se han identificado a lo largo de su distribución (Rodríguez-Serrano et al. 2006) y trabajos genético-poblacionales han señalado la existencia de 3 filogrupos: en Chile central, Patagonia Continental y Tierra del Fuego (Lessa et al. 2010; Abud 2011; Smith et al. 2001).

El potencial de *A. olivacea* como especie modelo para estudiar el proceso de colonización y asentamiento en distintos ambientes y los fenómenos biológicos asociados es claro; habita regiones notoriamente contrastantes en términos de temperatura, precipitación y altitud las cuales representan oportunidades de divergencia para las poblaciones del ratón oliváceo (Bozinovic et al. 2011; Abud 2011). Sería interesante analizar el rol de la plasticidad fenotípica y/o de los cambios adaptativos asociados al ambiente, así como evaluar, dado el importante número de subespecies que presenta, si estamos frente a un fenómeno de especiación ecológica incipiente. A su vez, se podría estudiar si algunos de estos fenómenos se han estado desarrollando de manera paralela entre los filogrupos Patagónico y Fueguino, ya que éstos se disponen de manera latitudinal en contraste a la orientación preferentemente longitudinal de los macroambientes de estepa y bosque.

Por otro lado, el ratón oliváceo cuenta con una historia demográfica que ha propiciado un interesante escenario para estudiar también el papel de la divergencia asociado al aislamiento geográfico. La actual distribución y diversidad de *A. olivacea* parece ser producto de expansiones demográficas, así como de diferenciación *in situ* en Tierra del Fuego (Abud 2011; Lessa et al. 2010).

En la actualidad, gracias a las nuevos métodos de secuenciación, podemos contar con la secuencia y expresión de miles de genes y marcadores moleculares de "especies no modelo" para analizar diversos problemas biológicos. En lo que refiere particularmente a *A. olivacea*, estas nuevas herramientas no solo se muestran útiles para estudiar fenómenos tales como la plasticidad fenotípica, sino también para precisar las características de su historia demográfica y definir las subespecies y filogrupos que posee, que en definitiva conforman las hipótesis de trabajo para abordar dichos fenómenos.

LAS NUEVAS TECNOLOGÍAS DE SECUENCIACIÓN MASIVA Y LAS OPORTUNIDADES QUE OFRECE LA TRANSCRIPTÓMICA PARA EL ESTUDIO DE LAS POBLACIONES NATURALES

La principal diferencia de las nuevas tecnologías de secuenciación masiva con respecto al método tradicional de Sanger (Sanger et al. 1977) es su capacidad de secuenciar millones de fragmentos de ADN en paralelo. Esta capacidad ha permitido que estas nuevas tecnologías produzcan mucho más información por cada corrida respecto al método tradicional, y ha provocado una gran reducción en el costo de la secuenciación (Metzker 2010). Entre las plataformas más destacadas se encuentran: GS FLX Titanium (Roche), Genome Analyzer/HiSeq 2000 (Illumina) y SOLiD/ Ion Torrent PGM (Life Technologies). Éstas se pueden clasificar según el método de secuenciación que utilizan o en base a cómo preparan la muestra (Metzker 2010), pero todas tienen la capacidad de generar masivamente lecturas, de manera cíclica y ordenada, de los fragmentos de cDNA que se están secuenciando. Sus principales diferencias radican en el largo de las lecturas generadas, la cantidad de datos generados por corrida y los costos asociados.

Actualmente, se están desarrollando la tercera generación de secuenciadores representados por GridION (Oxford Nanopore Technologies) y PacBio RS II (Pacific Biosciences). A diferencia de los métodos anteriores, éstas tecnologías paralelizan la secuenciación de una sola molécula de ADN, no requieren amplificación por PCR y las señales emitidas por el proceso enzimático son monitoreadas en tiempo real (Liu et al. 2012). GridION, hace uso de un nanoporo para ver las diferencias en voltaje de cada nucleótido de la cadena y así identificarlos (Schadt et al. 2010), mientras que la plataforma desarrollada por Pacific Bioscience captura la fluorescencia que emite cada nucleótido marcado a medida que se va

sintetizando una de las cadenas del ADN (Schadt et al. 2010). Ambas tecnologías aventajan a sus predecesoras porque requieren menos tiempo para preparar las muestras, la "corrida" es más rápida, y fundamentalmente porque producen lecturas más largas, que en promedio se extienden más allá de las 1000pb (Liu et al. 2012) y que pueden alcanzar más de 10kb (Schadt et al. 2010). Cabe destacar que la secuenciación mediante nanoporos aún se encuentra en desarrollo, mientras que la plataforma de Pacific Bioscience ya ha sido utilizada para ensamblar varios genomas (Koren et al. 2012; Bashir et al. 2012), pero sólo recientemente se han podido ensamblar y *cerrar* genomas microbianos únicamente con ésta tecnología (Chin et al. 2013; Ferrarini et al. 2013). En general, ésta tecnología se ha utilizado en conjunto con otras plataformas (Bashir et al. 2012; Koren et al. 2012; Ribeiro et al. 2012) o aplicando determinadas estrategias (Travers et al. 2010), para lidiar con su elevada tasa de error (Koren et al. 2012).

A pesar de las facilidades que ofrece la tercera generación de secuenciadores, secuenciar y ensamblar genomas de mamíferos sigue siendo una tarea ardua, especialmente para las especies no modelos para las cuales no se cuenta con un genoma de referencia. En estos casos, secuenciar y ensamblar su transcriptoma resulta en una estrategia más efectiva y económica (Perry et al. 2012). Mediante la secuenciación del ARN, es posible acceder a la secuencia de un gran número de genes, estudiar la estructura de los transcriptos (producto del empalme alternativo), la información alélica, y la expresión (Wang et al. 2009). En efecto, haciendo uso de la información que ofrece el transcriptoma, varios trabajos han abordado, por ejemplo, el fenómeno de la especiación de varias especies no modelos analizando las secuencias codificantes, SNPs (del inglés, *Single Nucleotide Polimorphisms*) así como la expresión génica (Wolf et al. 2010; Renaut et al. 2010; Gagnaire et al. 2012; Pavey et al. 2010). Otros trabajos, han utilizado el transcriptoma para inferir la historia demográfica de especies no modelos (McCoy et al. 2014) mediante la identificación cuidadosa de SNPs (Gayral et al. 2013) y haciendo uso de nuevos modelos-métodos poblacionales (Gutenkunst et al. 2009; Excoffier et al. 2013).

CARACTERIZACIÓN DEL TRANSCRIPTOMA, MÉTODOS Y ANÁLISIS ASOCIADOS

Para las especies no modelo, caracterizar el transcriptoma, es decir, describir qué genes se encuentran expresados en un determinado tejido, etapa de la vida o condición fisiológica de un organismo, es un paso previo e ineludible al análisis de la expresión génica y de la identificación de SNPs en exones (Ekblom & Galindo 2011). Para estas especies que en general no cuentan con un genoma de referencia, el ensamblado *de novo* es la única alternativa para construir las secuencias de los transcriptos. Varios algoritmos existen para ensamblar las lecturas *y* reconstruir cóntigos (adaptado del inglés, *contigs*) [eg. Oases (Schulz et al. 2012), Trans-ABySS (Robertson et al. 2010)], que idealmente representan los distintos transcriptos. Una vez obtenidos los cóntigos, éstos se anotan utilizando genomas de otras especies o base de datos de secuencias [e.g Ensembl (Flicek et al. 2014)] como referencia. Posteriormente, se puede proceder a la caracterización funcional de éstos

mediante recursos como el DAVID (Dennis et al. 2003), que haciendo uso del vocabulario estandarizado del proyecto de Ontología Génica, describe los atributos de los cóntigos después de obtener la presunta homología con genes conocidos. En particular, a cada cóntigo se le asigna un "proceso biológico" al cual contribuye, una "función molecular" que refiere a la actividad bioquímica que realiza el gen y un "componente celular" que describe la localización subcelular donde el producto del gen es activo (Ashburner et al. 2000). A pesar que la mayoría de los estudios que caracterizan el transcriptoma son principalmente descriptivos, establecen un importante punto de partida y brindan valiosos recursos para análisis posteriores (Ellegren 2008).

ENSAMBLADO *DE NOVO* DEL TRANSCRIPTOMA

En el sentido más amplio, ensamblar implica reconstruir una secuencia dada a partir de subfragmentos de ésta, obtenidos como partes no identificadas de un gran conjunto de lecturas. Para resolver este tipo de problemas la ciencia de la computación utiliza grafos, un constructo que contiene nodos y aristas, y todas las cadenas (*strings*) posibles de un largo arbitrario k (k-meros) de los subfragmentos. Una vez obtenido el grafo, se buscan ciclos Hamiltonianos, es decir recorrer el grafo de modo tal que cada vértice se recorra una sola vez, o ciclos Eulerianos, el cual implica recorrer el grafo visitando cada arista una sola vez que, en caso de existir, aseguran la reconstrucción de la secuencia original (Compeau et al. 2011). Actualmente, la gran mayoría de los ensambladores, tanto para la reconstrucción del genoma como del transcriptoma, buscan ciclos Eulerianos (o caminos si la secuencia original a ensamblar es lineal) (Compeau et al. 2011) principalmente por dos motivos: i) encontrar ciclos Hamiltonianos es mucho más difícil y ii) la construcción de los grafos para encontrar ciclos Hamiltonianos implica hacer millones de alineamientos pareados entre los k-meros, lo que vuelve a ésta aproximación aún más costosa computacionalmente. Para encontrar caminos o ciclos Eulerianos se construyen los grafos conocidos como de Bruijn, donde los nodos, que son únicos, son los prefijos o sufijos de longitud k-1 de cada k-mero y las aristas conectan dos nodos si hay un k-mero que contenga esos prefijos y/o sufijos.

Mientras que el ensamblado de una secuencia a partir de sus subfragmentos sin errores es una tarea sencilla utilizando los grafos de Bruijn y buscando caminos o ciclos Eulerianos, no lo es para ensamblar secuencias a escala genómica. Primero, porque dependiendo del valor de k, puede no ser posible obtener todos los k-meros del genoma, y segundo porque los k-meros van a tener una tasa de error. Estos errores sumado a regiones repetitivas en el genoma dan lugar a burbujas, caminos divergentes, espolones (*spurs*) y caminos que convergen y después divergen (Miller et al. 2010).

A diferencia del ensamblado *de novo* de un genoma, para el cual se espera pocos grafos reconstruidos, pero muy conectados y grandes representando la conectividad de las lecturas de un cromosoma entero, el ensamblado del transcriptoma implica reconstruir muchos grafos donde cada uno representaría las diferentes alternativas de empalme de un gen. Uno de los primeros programas específicamente diseñados para el ensamblado de transcriptos y que ha mostrado ser uno de las mejores programas al respecto (Zhao et al. 2011) es Trinity (Grabherr et al. 2011) (ver Recuadro 1).

RECUADRO 1. ALGORITMO DEL TRINITY

Esquemáticamente el ensamblado por medio de Trinity consiste en 3 etapas y, al igual que otros, utiliza los grafos de Bruijn para el ensamblado de los transcriptos. La primera etapa, *Inchworm*, parte de un k-mero y lo extiende utilizando otro, siempre y cuando: i) presente una frecuencia "similar" y ii) presente un solapamiento de k-1 con el anterior. Más precisamente, parte del k-mero más abundante y busca el siguiente en abundancia con dicho solapamiento. En caso que no existe tal k-mero, el proceso se reinicia e itera hasta que no haya ningún k-mero más por extender. Cómo producto, se obtiene los transcriptos más expresados y las secuencias únicas (exones) de aquellas isoformas menos expresados. La siguiente etapa, *Chrysalis*, agrupa los cóntigos-transcriptos, completos y parciales producto de *Inchworm*, que surgen de un determinado loci o gen parálogo y construye por cada agrupación un grafo de Bruijn ponderado. En primera instancia la sub-etapa *GraphFromFasta*, agrupa cóntigos si poseen un solapamiento perfecto de k-1 mero entre ellos y si el número de lecturas que presentan un alineamiento idéntico de (k-1)/2 bases de cada lado de la unión es mayor a un número dado. Los cóntigos finalmente agrupados, aparecen al final del ensamblado bajo un mismo "componente". La sub-etapa *FastaToDeBruijn,* construye el grafo de Bruijn por cada componente donde los nodos tiene un tamaño de k-1 y los vértices k. *ReadsToTranscripts* mapea las lecturas que tiene más k-meros en común con los cóntigos agrupados y luego *QuantifyGraph* pondera las aristas del grafo, de largo k, en función del número de lecturas que tiene tal k-mero. Finalmente, la última etapa *Butterfly*, simplifica el grafo, poda los bordes del grafo que representan desviaciones, probablemente causadas por errores en la secuenciación, y utilizando programación dinámica, recorre los diferentes caminos de éste buscando mantener el mayor número de lecturas y sus pares correspondientes.

ANÁLISIS DE LA EXPRESIÓN GÉNICA

El procedimiento bioinformático para llevar a cabo el análisis de expresión, pretende "convertir" las millones de lecturas, que se obtuvieron como resultado de la secuenciación masiva, en una medida de expresión. Este procedimiento se puede esquematizar en cuatro etapas (Oshlack et al. 2010). La primera, consiste en mapear (*mapping*) o alinear las lecturas a un transcriptoma o genoma. La idea de esta etapa es encontrar para cada lectura, su ubicación única (en caso de que corresponda) en la referencia. La segunda, consiste en agregar las lecturas en una unidad con sentido biológico, como exones, transcriptos o genes. Luego, se procede a estandarizar (o normalizar) la expresión de los genes para volverlos comparables. Este procedimiento difiere según qué comparaciones se pretenden hacer; se puede comparar tanto los perfiles de expresión de genes entre muestras como dentro de cada muestra. Por ejemplo, para el análisis dentro de cada muestra es necesario normalizar por la longitud del transcripto, mientras que esto no es necesario para el análisis entre muestras. Finalmente, el análisis de la expresión génica propiamente dicho, trata de establecer si las diferencias en la cantidad de lecturas que fueron agregadas en una determinada unidad biológica, son estadísticamente significativas. Aquí, los diferentes programas difieren, entre otras cosas, en las distribuciones (Binomial negativa, Poisson, etc) que asumen para modelar el conteo de las lecturas.

Dado que un gen puede expresar transcriptos distintos que comparten exones, una fracción de las lecturas termina alineándose múltiples veces en las secuencias ensambladas. En general, la estrategia tradicional para calcular la expresión génica consiste, simplemente, en eliminar tales lecturas. Actualmente, existen programas que al lidiar con la multimapabilidad de las lecturas, incorporan la expresión de las isoformas y por lo tanto producen estimaciones menos sesgadas de la expresión génica. Uno de ellos es RSEM (Li & Dewey 2011), el cual mediante un modelo generativo y un algoritmo EM (del inglés, *Expectation Maximization*), que se utiliza para maximizar la verosimilitud de que un fragmento dado derive de un transcripto particular, calcula los valores de expresión de los genes e isoformas (Li & Dewey 2011; Li et al. 2010). La intuición detrás de este abordaje estadístico se puede esquematizar en un algoritmo de 3 pasos: i) se estima la abundancia de cada transcripto en base a las lecturas que alinean únicamente, ii) por cada lectura degenerada, dividirla en cada transcripto en función de su abundancia estimada en el punto uno y iii) recalcular las estimaciones de abundancia utilizando los nuevos conteos de cada transcripto y vuelta al paso i). Éstos pasos se repiten hasta que las estimaciones de abundancia no cambien más allá de un umbral. Una iteración del algoritmo EM correspondería aproximadamente al algoritmo de "rescate" de Mortazavi et al. 2008 (Li et al. 2010)

NORMALIZACIÓN DIGITAL

A pesar de los avances en las tecnologías de secuenciación, los errores en la secuenciación y los sesgos en el muestreo del genoma aún dificultan el ensamblado completo de genomas. Con el objetivo de muestrear todo el genoma se ha apuntado a secuenciaciones cada vez más profundas, lo que limita la velocidad y efectividad de los algoritmos de ensamblado. Lo mismo sucede con los transcriptomas, donde el transcripto menos expresado es quién determina la cobertura mínima para muestrear todo el transcriptoma. Con el fin de sistematizar y equilibrar la cobertura y volver el ensamblado más eficiente, apareció la normalización digital. Básicamente, consiste en estimar la cobertura de las regiones genómicas o transcriptómicas sin utilizar un genoma de referencia, y eliminar aquellas lecturas provenientes de regiones que ya cuenten con la cobertura suficiente. Para la estimación de la cobertura, utiliza la media de las abundancias de los k-meros de una lectura, ya que ésta está fuertemente correlacionada con la cobertura calculada a partir de los alineamientos de las lecturas. Por otro lado, dada la tasa de error de la secuenciación que se ubica en un 1% para Illumina, al descartar, por ejemplo, una lectura de 100 pb con una base errónea, se eliminan k k-meros erróneos. De esta manera, este algoritmo también ayuda a mitigar los efectos que producen los errores de secuenciación en los algoritmos de ensamblado, especialmente en la construcción del grafo.

Actualmente, existen dos algoritmos de normalización digital, el desarrollado por (Brown et al. 2012) y el presentado recientemente por (Haas et al. 2013), más específico para análisis subsiguientes con el programa Trinity.

JUSTIFICACIÓN Y OBJETIVOS

El ratón oliváceo es una especie interesante para evaluar el rol de la plasticidad, de la adaptación y de la expresión génica en la colonización y persistencia en ambientes contrastantes. A su vez, un estudio más detallado de su historia demográfica contribuiría a entender la diversidad y disposición actual de la biota Patagónica y Fueguina. Varios estudios ya se han aproximado a problemas similares utilizando datos exclusivamente transcriptómicos (Wolf et al. 2010; Renaut et al. 2010; Gagnaire et al. 2012). Es por esto que en el presente trabajo, procedimos a caracterizar el transcriptoma de riñón de *A. olivacea* utilizando 13 ejemplares, pertenecientes a 4 puntos geográficos distintos y dos ambientes, estepa y bosque, con el fin de obtener un transcriptoma renal de referencia para la especie. De esta manera, pretendemos encaminar y facilitar el estudio de diversos fenómenos evolutivos y ecológicos como los mencionados anteriormente y contribuir a determinar las características (magnitud y tiempo) de su expansión demográfica. Elegimos particularmente caracterizar tejido renal por la enorme diferencia que existe en precipitación media anual entre la estepa y bosque, principalmente a la altura de la latitud 42ºS. La información generada en el presente trabajo, principalmente la relativa a las secuencias génicas, también serán de gran utilidad para las otras especies de Sigmodontinae.

OBJETIVOS GENERALES:

(a) Caracterizar el transcriptoma de riñón de *A. olivacea.*

(b) Evaluar distintos métodos para obtener un transcriptoma de referencia.

OBJETIVOS ESPECÍFICOS:

(a) Ensamblar el transcriptoma de riñón de *A. olivacea.*

(b) Anotación de los cóntigos ensamblados.

(c) Determinar y caracterizar los genes más expresados.

(d) Comparar los genes más expresados con los descritos en la bibliografía.

(e) Comparar los métodos de normalización digital y analizar su rendimiento con otras estrategias.

CAPÍTULO 1

CARACTERIZACIÓN DEL TRANSCRIPTOMA DE RIÑÓN DEL RATÓN OLIVÁCEO SUDAMERICANO *ABROTHRIX OLIVACEA*

Dada la importancia de *A. olivacea* como especie modelo para estudiar varios problemas biológicos y los escasos recursos genéticos con que cuenta la especie para abordarlos, decidimos caracterizar el transcriptoma de riñón del ratón oliváceo. Elegimos el riñón porque existen poblaciones que habitan regiones notoriamente contrastantes en cuanto a precipitación. Con el propósito de establecer un transcriptoma renal de referencia trabajamos con 13 individuos de cuatro puntos geográficos distintos, cubriendo así en gran medida la distribución del ratón oliváceo, y evaluamos la capacidad de ensamblado de varias estrategias. Por otro lado, caracterizamos los genes de más alta expresión entre los 13 individuos y los comparamos con los descriptos en ratón. Entre las distintas estrategias de ensamblado, constatamos que aquellas que involucran un algoritmo de normalización, TrinityNorm y DigiNorm, a efectos de eliminar las lecturas redundantes, produjeron ensamblados mucho más fragmentados pero con un número mayor de genes que la estrategia Multireads. Entre TrinityNorm y DigiNorm, el primero mostró un mejor balance en cuanto a genes hallados, nivel de reconstrucción de las secuencias codificantes y tiempo de ensamblado. En general, obtuvimos secuencias codificantes pertenecientes a más de 15.000 genes de los cuales 10.000, aproximadamente, tienen al menos una secuencia codificante reconstruida en un 90% o más. De los 283 genes que se encontraban en el 5% más expresado entre los 13 individuos de *A. olivacea*, encontramos 17 en común con los descriptos en la parte distal del túbulo renal de ratón. Respecto a las estrategias para establecer un transcriptoma de referencia, hay que tener en cuenta que la normalización debe ser evitada si se necesita reconstruir toda o casi toda la secuencia de los transcriptos. Para los casos donde el número de lecturas sea tal que impida el ensamblado siguiendo una estrategia Multireads, nuestros resultados sugieren que la estrategia TrinityNorm sería más conveniente que DigiNorm.

BMC Genomics

RESEARCH ARTICLE Open Access

Characterization of the kidney transcriptome of the South American olive mouse *Abrothrix olivacea*

Facundo M Giorello[1*], Matias Feijoo[1], Guillermo D'Elía[2], Lourdes Valdez[2], Juan C Opazo[2], Valeria Varas[2], Daniel E Naya[1] and Enrique P Lessa[1]

Abstract

Background: The olive mouse *Abrothrix olivacea* is a cricetid rodent of the subfamily Sigmodontinae that inhabits a wide range of contrasting environments in southern South America, from aridlands to temperate rainforests. Along its distribution, it presents different geographic forms that make the olive mouse a good focal case for the study of geographical variation in response to environmental variation. We chose to characterize the kidney transcriptome because this organ has been shown to be associated with multiple physiological processes, including water reabsorption.

Results: Transcriptomes of thirteen kidneys from individuals from Argentina and Chile were sequenced using Illumina technology in order to obtain a kidney reference transcriptome. After combining the reads produced for each sample, we explored three assembly strategies to obtain the best reconstruction of transcripts, TrinityNorm and DigiNorm, which include its own normalization algorithms for redundant reads removal, and Multireads, which simply consist on the assembly of the joined reads. We found that Multireads strategy produces a less fragmented assembly than normalization algorithms but recovers fewer number of genes. In general, about 15000 genes were annotated, of which almost half had at least one coding sequence reconstructed at 99% of its length. We also built a list of highly expressed genes, of which several are involved in water conservation under laboratory conditions using mouse models.

Conclusion: Based on our assembly results, Trinity's *in silico* normalization is the best algorithm in terms of cost-benefit returns; however, our results also indicate that normalization should be avoided if complete or nearly complete coding sequences of genes are desired. Given that this work is the first to characterize the transcriptome of any member of Sigmodontinae, a subfamily of cricetid rodents with about 400 living species, it will provide valuable resources for future ecological and evolutionary genomic analyses.

Keywords: *Abrothrix olivacea*, Abrotrichini, Cricetidae, Sigmodontinae, Muroidea, RNA-Seq, Gene expression, *De novo* assembly, Normalization methods

Background

The olive mouse *Abrothrix olivacea* [1] is a cricetid rodent of the subfamily Sigmodontinae, one of the largest mammalian subfamilies with about 400 species and 86 living genera [2,3]. The olive mouse is distributed along Chile and Argentinean Patagonia, from 18°S to 55°S latitude [4], extending for over 1000 km latitudinally, and encompassing a great variety of environments: coastal deserts in the north, Mediterranean scrubs in central Chile, Valdivian and Magallanic forests through the south of Chile and Argentina and Patagonian steppe towards the Atlantic coast. *A. olivacea* must withstand the arid Chilean north and the Patagonia steppe, as well as the Valdivian rain forest with 2700 mm or more of annual rainfall [5]. Given the striking biotic and abiotic differences among these environments, differences in thermoregulation and osmoregulation, among other physiological traits, are expected to occur. Higher tolerance to water shortage in populations from xeric habitat has already been demonstrated [6]. On the basis of variation in morphology, coloration patterns, and more recently DNA sequence data [4,7], many *A. olivacea* subspecies have been described and at

* Correspondence: fagire@gmail.com
[1]Departamento de Ecología y Evolución, Facultad de Ciencias, Universidad de la República, Montevideo, Uruguay
Full list of author information is available at the end of the article

least two phylogeographic breaks have been found along its distribution [8]. All these characteristics make *A. olivacea* a good focal case for the study of geographical variation in response to environmental variation.

High-throughput sequencing (HTS) [9] has a wide range of applications, from clinical [10] to functional studies in genomics [11], molecular ecology [12], and microbial diversity [13]. Recently, HTS has also been used to characterize transcriptomes of a growing number of non-model species (e.g. [14-17]). RNA-seq is a cost-effective way to obtain large amounts of coding sequences and information about gene expression levels [18]. The goal of covering entire genome or transcriptomes, along with the reduction of the HTS costs [9], has motivated digital normalization strategies [19] to systematize the increasing but uneven coverage in shotgun sequencing datasets. Normalization methods estimate the read abundance, regardless of a reference, using the k-mer median abundance of that read and then decides whether to reject or accept it based on the chosen coverage value [19,20]. In this manner, normalization algorithms remove redundant reads but also greatly reduce the total number of k-mers by discarding the majority of the erroneous ones. For example, with a sequencing base error rate of 1bp per 100 bp sequenced [9], k erroneous k-mers will be produced, being k equal to k-mers size. This data and error reduction notably decreases the computational requirements for *de novo* assembly.

In this study, we adopted paired-end Illumina sequencing to characterize the kidney transcriptome of *A. olivacea*. We chose kidney because of its association with multiple physiological processes, including water conservation [21] and nutrition [22]. This transcriptome will serve as a reference for comparative studies of geographical variation within this species, as well as for other studies on the diverse sigmodontine rodents. More than 800 million (M) reads were generated for 13 kidney transcriptomes of individuals sampled across Chile and Argentina. We explored various normalization strategies in order to obtain the best transcripts reconstruction and identify the most expressed genes. This is the first report of a sigmodontine transcriptome.

Results

Transcriptome sequencing and assembly

Transcriptome sequencing of 13 libraries using Illumina yielded a total of ~ 87 Gb of data, formed by ~430 M of paired reads with an average length of 101 bp (Additional file 1: Table S1). Trimming of low quality bases from the 3' end, prior to Trinity [23] *de novo* assembly, reduced average read length to 83 bp. The number of reconstructed contigs per library ranged from 62,499 to 120,209; with average length ranging from 972 to 1174 bp and median from 488 to 585 bp (Table 1). Detailed results for each library are shown in Additional file 1: Table S2.

To obtain a good reference transcriptome, we also explored three strategies: (i) combining reads of all libraries (Multireads), (ii) Trinity's *in silico* normalization (TrinityNorm) [20], and (iii) digital normalization (DigiNorm) [19]. The last two strategies involve, in order to improve assembly efficiency from high coverage sequencing datasets, the deletion of redundant reads, ideally without harming the quality of the final reconstructed genes. Of these two, TrinityNorm was more severe than DigiNorm in reducing the total number of paired-ends reads from ~430 M to ~22 M vs. ~50 M (Table 1). Meanwhile, digital normalization was faster than *in silico* Trinity normalization: 9 hours vs. 14 hours.

As expected, the Multireads strategy led to a far more time consuming and computationally demanding assembly than either of the normalization methods, being five and over nine times slower than the assembly from DigiNorm and Trinity, respectively (Table 1). Also, the average and median lengths of reconstructed contigs from the Multireads data set were smaller than the assembled contigs from normalized reads, with 1,060 and 443 bp for mulitreads, 1,210 and 575 bp for TrinityNorm, and 1,269 and 696 bp for DigiNorm. These results are consistent with the distribution of the contigs, where almost half (46%) of the reconstructed contigs from the Multireads strategy were between 200 and 400 bp (Additional file 1: Table S3). On the other hand, the Multireads strategy reconstructed the longest contigs (Additional file 1: Table S3) with 4,212 above 6,400 bp. TrinityNorm and Diginorm reconstructed only 3,073 and 2,726 of contigs above this length, respectively.

The two normalization strategies produced similar assembly results in terms of average and median length of contigs, with a small advantage for DigiNorm values, but they significantly differed in the number of contigs assembled, DigiNorm assembled 85,902 more contigs than TrinityNorm and 87,013 more than the Multireads strategy (Table 1).

Gene annotation and evaluation of reconstructed coding sequences

Annotation was based on BLASTX searches against: (i) OMA browser mouse protein database, which contains the protein isoforms of *Mus musculus* genes [24] and (ii) NCBI non-redundant vertebrate protein database. For the two databases the same e-value threshold of 1e-10 was set. For the Multireads, TrinityNorm and DigiNorm strategies, each assembled transcript was also analyzed through the Pfam database [25] using HMMER [26,27] for proteins domain identification. A file summarizing the Pfam and BLASTX results for each of the three strategies is available as Additional file 2.

Table 1 Main assembly metrics for the three assembly strategies and individual libraries

	Range (of individual libraries)[a]		Multireads	TrinityNorm	DigiNorm
	min	max			
Reads	27041064	42477318	430525978	21757448	50557782
Total contigs	62499	120209	275903	277014	362916
Max contig length	9942	15496	20648	19625	15961
Min contig length	201	201	201	201	201
Average length	972	1174	1060	1210	1269
Median length	488	585	443	575	696
Running time (hours)	n/a	n/a	94 (12 threads)	10 (12 threads)	19 (12 threads)
Normalization time (hours)	n/a	n/a	n/a	14 (1 thread)	9 (1 thread)

[a]"Range (of individual libraries)" shows for each row the maximum and minimum value found among the 13 individual libraries of kidney transcriptome of the olive mouse *Abrothrix olivacea*.

The maximum number of mouse genes annotated within a particular library was 12,988 from the significant hits of 55,332 contigs of the 120,209 assembled (Table 2 and specimen PPA 443 library in Additional file 1: Table S2 and Table S4a). The union of the 13 individual BLASTX runs only added 1,630 significant hits (14,618 in total), indicating the high level of redundant information across libraries. On the other hand, when using the extensive non-redundant vertebrate database as reference, the maximum number of contigs annotated within a single library was 58,404, 3072 contigs more than with the OMA database (Additional file 1: Table S4b). Detailed results for each library are shown in Additional file 1: Table S4. Hereafter we present the results based only on mouse proteins from OMA. This database allow us to count the number of genes and their corresponding reconstructed coding sequences (CDS) and obtain an upper bound estimation of genes orthologous with mouse.

Of the 14,618 mouse genes annotated through the union of all libraries, almost one half (7,060) had at least one putative CDS reconstructed at > 99%, 9,290 at > 90%, and 10,104 at > 80%, of the total expected length. More importantly, in total, 9,434 distinct mouse isoforms of the 24,338 (~39%) available at OMA browser were almost fully reconstructed (>90%) for *A. olivacea* (Table 2).

Among the three strategies carried out to obtain a reference transcriptome, Multireads reached the lowest number of mouse genes, 14,788; meanwhile TriniNorm and DigiNorm, reached 15,077 and 15,095 respectively. Despite having found the lowest number of genes, the Multireads strategy performed best at reconstructing coding sequences, obtaining similar values to those gathered through the union of the single libraries, with 7053 distinct mouse coding sequences reconstructed at > 99%, 9,543 at > 90%, and 10,480 at > 80% (Table 2). With regard to genes, of the 14,788 annotated by the Multireads alternative, 47% had at least one CDS fully (>99%) reconstructed, clearly surpassing the 39% and 34% of TrinityNorm and DigiNorm respectively (Figure 1). Between the two normalization strategies, TrinityNorm

Table 2 Gene annotation and coding sequences reconstruction using BLASTX to OMA browser mouse protein database

	Minimum % of CDS reconstructed	Range (of individual libraries)		DigiNorm	Multireads	TrinityNorm	Library union
		min	max				
Genes	99	3090 (3102)	5021 (5037)	5211 (5245)	7017 (7053)	5895 (5926)	7060 (7131)
Contigs		4371	7763	15815	14881	13671	n/a
Genes	90	4534 (4557)	6882 (6916)	7354 (7436)	9467 (9543)	8252 (8325)	9290 (9434)
Contigs		6861	11361	26904	22137	21825	n/a
Genes	80	5706 (5745)	8636 (8636)	8530 (8665)	10377 (10480)	9347 (9467)	10104 (10367)
Contigs		5422	14129	36814	26745	28270	n/a
Genes	50	8025 (8138)	9905 (10089)	12261 (12610)	12587 (12818)	12498 (12769)	11874 (12544)
Contigs		14168	25674	76156	40607	51140	n/a
Total genes		11564 (11941)	12988 (13434)	15095 (15717)	14788(15204)	15077 (15605)	14618 (15772)
Total contigs		32934	55332	157390	70380	100786	n/a

The first value indicates the number of mouse genes found (for which at least one coding sequence was reconstructed). Values in parenthesis are the number of distinct coding sequences reconstructed at each level. The row corresponding to "contigs" indicates the number of contigs that reconstructed coding sequences (CDS) at each level.

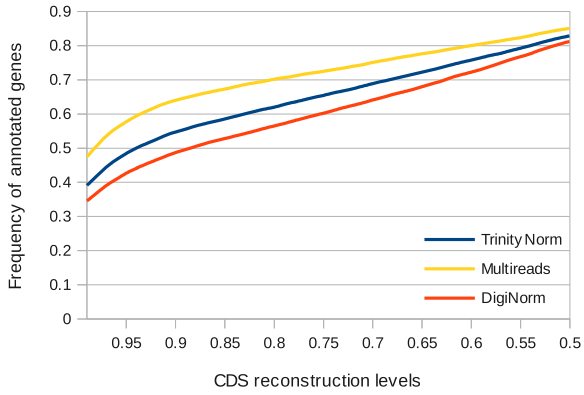

Figure 1 Frequency of genes annotated for *Abrothrix olivacea* at corresponding CDS reconstruction levels. Frequency of genes annotated per total number of genes were calculated for each strategy (see text) and for the corresponding mouse coding sequences (CDS) reconstruction level. The CDS reconstruction intervals are of 0.01 percent. Only the largest reconstructed CDS for each gene are taken into account.

outperformed DigiNorm at each reconstruction level in terms of numbers of mouse genes found and percentage of coding sequence reconstructed (Table 2 and Figure 1). On the other hand, the assembly from DigiNorm had more contigs at each level of reconstruction and also more contigs with at least one Pfam domain, followed by TriniNorm and the Multireads strategies (Table 2 and Additional file 1: Table S5). This is expected if those contigs represent distinct fragments of the same coding sequence or if they are isotigs (overlapping contigs) representing (ideally) distinct isoforms. However, when the number of potential isoforms from Trinity assembly were inferred and counted (see methods), the average number of alternative reconstructions per contig was 4.9 for DigiNorm and only 2.9 for the Multireads strategy (data not shown). Thus, those contigs are alternative reconstruction (isotigs) representing (possibly) isoforms and not subfragments of a given reference.

Functional annotation

For the functional annotation of the transcriptome, we selected the genes found by TrinityNorm. This strategy was the one with the best tradeoff between CDS reconstructed, genes found, and computational speed. To this end, the Database for Annotation, Visualization and Integrated Discovery (DAVID) [28], was used to classify them with Gene Ontology (GO) terms.

For the 15,077 genes found by TrinityNorm strategy, 9,793 GO terms were categorized in Biological Processes, 9,486 in Molecular function and 8,978 in Cellular Components. Most genes at Biological Processes belong either to "Regulation of transcription" (1,726), "Transcription" (1,441) and to "Regulation of RNA metabolic process" (1,093) (Figure 2). Likewise, the Molecular Function category subdivided annotated sequences into "ion binding" (3,234), "cation binding" (3,201), and "metal ion binding" (3,172) as the most represented (Figure 2). Within the category Cellular Component, the three principal groups were: "intrinsic to membrane" (3,667), "integral to membrane" (3,506) and "plasma membrane" (2,167) (Figure 2).

The most expressed genes

To determine the most expressed genes in the *A. olivacea* kidney transcriptome, TPM (*Transcripts Per Million*) expression values were calculated for each single library with RSEM software [29]. For this purpose, a set with 5% of most expressed genes (~600 genes) for each of the 13 transcriptomes was identified; these were cross searched to identify those genes common to all libraries. Two hundred eighty-three genes resulted to be present in all libraries (Additional file 3: Table S5). The average TPM values ranged from 333 to 17,798 (Additional file 3: Table S5). Five genes that showed the highest average TPM values were: predicted gene 4076 (possibly a NADH-ubiquinone oxidoreductase) (ENSMUSG00000096449), glutathione peroxidase 3 (ENSMUSG00000018339), ferritin heavy chain 1 (ENSMUSG00000024661), hemoglobin beta adult major chain (ENSMUSG00000052305), and phosphoenolpyruvate carboxykinase 1 cytosolic (ENSMUSG000000027513) (Additional file 3: Table S5).

GO terms for the 283 genes obtained from DAVID showed that the most enriched terms among the three domains using the TrinityNorm genes list as background were: "hydrogen ion transporting ATP synthase activity, rotational mechanism" (17.4 Fold Enrichment, domain: molecular function), "proton-transporting ATP synthase complex" (25.4 Fold Enrichment, cellular component) and "mitochondrial ATP synthesis coupled electron transport" (25.3 Fold Enrichment, biological process) (Additional file 3: Table S6).

Subsequently, these 283 genes were cross-checked with a list obtained from Pradervand *et al.* [30], who enumerated the most expressed genes in the distal part of the mouse renal tubule using microarrays. Seventeen genes resulted to be in common (Additional file 3: Table S5 in bold): two transcription factors, one small GTPase, eight transporters and channels, and six cytoskeleton-related genes. Among these genes are Aquaporin 1 (*Aqp1*), Ras-related protein Rab-7a (*Rab7*), Sodium/potassium-transporting ATPase gamma chain (*Fxyd2*), Voltage-dependent anion channel 1 (*Vdac1*), and Guanine nucleotide binding protein, alpha stimulating (*Gnas*).

Discussion

The subfamily Sigmodontinae includes about 400 species and 86 living genera [2,3]. Of these, *A. olivacea* inhabits a wide range of contrasting environments and presents

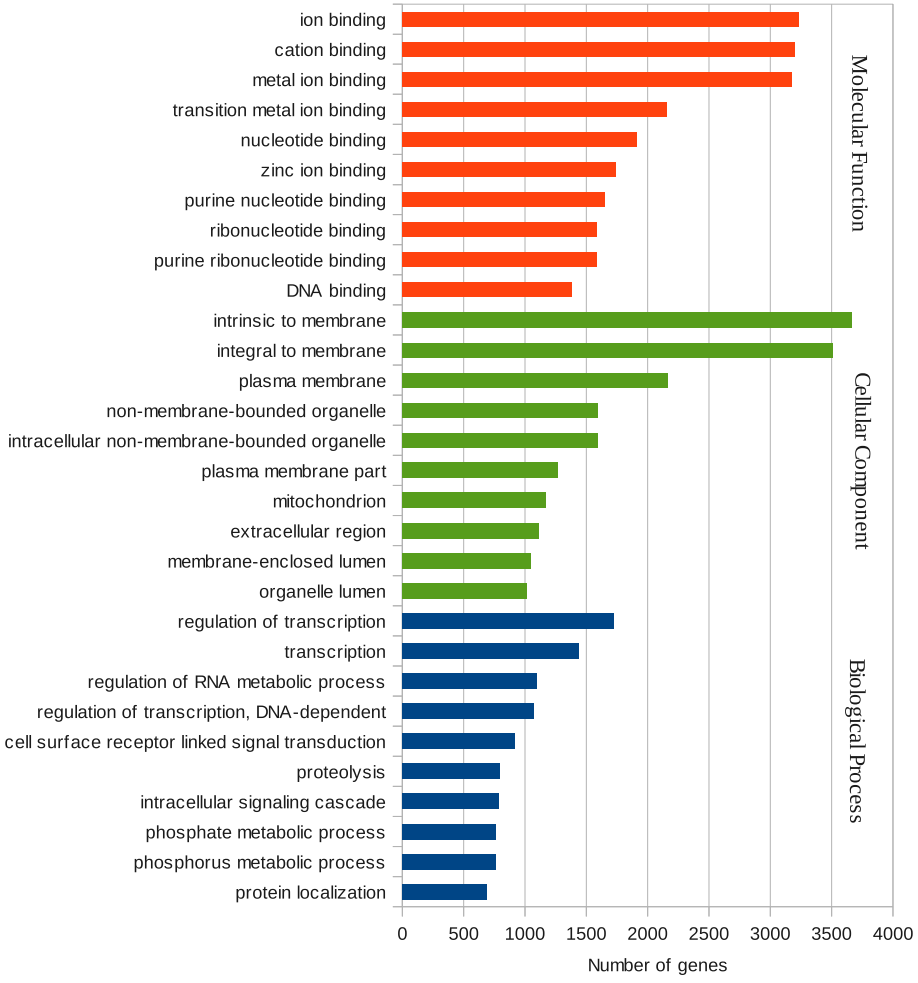

Figure 2 Gene Ontology classification of genes found by TrinityNorm strategy. Results are summarized in three main categories: Biological process, Molecular function, and Celullar component.

different geographic forms. In this work, 13 individuals were sampled in Argentina and Chile, covering both the arid Patagonian steppe and the wet Valdivian and Magellanic forests. More than 800 M reads were generated in what constitutes the first characterization of a sigmodontine transcriptome. In addition, we present a set of highly expressed genes of which some are possible candidates for ecological studies of the response of the species to environmental variation in water and dietary items availability.

Transcriptome sequencing and assembly

As the cost of high-throughput sequencing falls and more cDNA sequences are generated, the importance of appropriate normalization strategies prior to contig assembling increases. In this study, we used two normalization strategies and compared their performance with a non-normalized (Multireads) alternative.

In terms of length and number of contigs assembled regardless of the strategy, our results were similar to those found in previous studies in which Trinity was used [31,32]. Among the three strategies, if only contigs descriptive statistics are taken into account, the assembly after both normalization strategies clearly outperformed the Multireads approach. Normalizations not only showed the largest average and mean contig lengths but also ran considerably faster than the Multireads counterpart. This is a consequence of discarding reads that are considered to be redundant and the concomitant sequencing error removal. If transcriptome coverage is moderate, it is necessary to keep in mind that, as noted by Brown *et al.* [19], the memory requirements will be roughly the same, with or without normalization, due to limited removal of erroneous k-mers.

Gene annotation and evaluation of reconstructed coding sequences

To evaluate the capacity of each strategy to assembly the *A. olivacea* transcriptome, we quantified the number of

contigs that resulted to be homologs of mouse OMA proteins at different reconstruction levels. We found that the Multireads strategy performs the best for the > 99%, > 90 and > 80% of the total mouse CDS length and obtained the most similar values to those found through the union of the individual sets of contigs. Even though the Multireads strategy produced the assembly with smaller median and mean contig sizes, larger contigs were reconstructed (Additional file 1: Table S3), thus explaining the better results obtained for the CDS reconstruction. Therefore, studies that require complete reconstruction of coding sequences should avoid normalization; but obviously, when billions of reads are available for assembly, the Multireads approach becomes prohibitive and normalization is the only way to proceed. These results are consistent with those found by [19]; it seems that normalization generates a more fragmented assembly, at least when Trinity is used. Our results are consistent with the notion that fragmentation is a by-product of normalization, but also that normalization negatively impacts the completeness of the coding sequences reconstruction at all levels. For both normalization methods a k = 25 (k-mer) was set, and it is not clear why such fragmentation is produced.

If assembly capacity and computational requirements are considered, we consider that assembly from TrinityNorm has the best cost-benefit return. This strategy was second after Multireads in number of genes reconstructed for each category while using about 1/10 of the time. Moreover, a high number of putative isoforms were reconstructed, supported by the number of contigs reconstructed per category (Table 2) and the isotigs counted from Trinity assembly (see Results). Even though DigiNorm reached the largest number of isotigs and the highest number of genes, its performance at reconstructing full and almost full CDS was the worst and very similar to the best single library. Also, the assembly from DigiNorm was two times slower than that from TrinityNorm. The latter required more time for normalization but is capable of running multithreading and so outperforms DigiNorm.

Regardless of the strategy, a large number of homologous mouse genes were obtained in our study. According to microarray studies [33], about 7600 genes are expressed in the kidneys of adult humans; meanwhile, recent HTS transcriptome studies found 15,369 for the baboon kidneys [34]. Clearly, in this study we were able to find almost the same number of genes (15,077 through TrinityNorm) that Spradling *et al.* [34] even though we used a very stringent e-value for the BLASTX analysis. Moreover, of those genes, 40 % had at least one full (>99%) CDS reconstructed. Given that, we reconstructed 8,325 distinct mouse isoforms out of 24,338 (~34%) for at least 90% of the total expected length, we established an important set of sequences, likely orthologous to mouse genes, which will be useful for future analyses of molecular evolution, population genomics, and phylogenetics.

List of most expressed genes

The assessment of gene differential expression tends to be problematic for contigs with low counts [35]; therefore, a good strategy is establishing a set of highly expressed genes for directing efforts to study differential expression. In this work, 13 transcriptome libraries were used to identify the most expressed genes of the kidneys of *A. olivacea*. Some of them had already been described for model species, while many of the new ones have a clear relationship with renal function and could serve as potential candidates for future evolutionary and ecological genomic studies.

Seventeen of 283 most expressed genes found in this work were previously singled out by Pradervand *et al.* [30] using microarrays in the distal part of the mouse renal tubule. Although for some of these genes their precise function is not clear, for others knowledge on their function is reasonably good. For example, *Aqp1* is involved in water reabsorption at the apical and basolateral plasma membrane of the proximal tubule [36]; mutations in *Fxyd2* have been associated with renal hypomagnesemia-2 [37]; *Rab7*, as a Rab member, could be implicated in the transport, docking, and fusion of endocytotic vesicles [30], and finally, *Gnas* codifies the alpha subunit of heterotrimeric G proteins, which mediates the vasopressin receptor type 2 signaling after the binding with vasopressin, and ultimately increases water reabsorption in the collecting duct [36]. In our expression analysis, *Fxyd2* and *Aqp1* are among the top 10 and top 50 of the most expressed genes, respectively. The latter represents a good candidate gene for the study of differential responses to variation of environmental water availability.

Among the 266 remaining highly expressed genes, additional putative candidates associated with renal function were found; for example: i) kallikrein (*Klk1)* encodes a proteolytic protein which produces the kinin proteins, which may counteract the hydrosmotic effect of vasopressin [36]; ii) Uromodulin (*Umod*) encodes the most abundant protein in urine [38], and mouse knockouts for this gene have shown urine concentration problems [36]; iii) Glyoxylate reductase (*Grhpr*), an enzyme that catalyzes the reduction of glyoxylate to glycolate, is associated with a disorder that can cause nephrolithiasis (kidney stone), nephrocalcinosis, and renal failure [39,40]; and iv) Sorbitol Dehydrogenase (*Sord*), along with Aldose reductase, are possibly involved in osmoregulation in the kidney [36,41].

Regarding the GO-term classification, no important differences were found between the set of all genes and the 283 most expressed ones, except that an expected excess of mitochondrial related GO-terms was found among the latter (Additional file 3: Table S6). This enrichment is not

surprising as the kidney is an energetically demanding organ [42,43].

Conclusion

In order to obtain the best-reconstructed transcripts from the kidney of the olive mouse *A. olivacea* on the basis of 13 individual libraries, we first explored three alternative assembling strategies. Results indicate that the Trinity's *in silico* normalization is the best algorithm in terms of cost-benefit return. We annotated more than 10,000 genes that were almost fully reconstructed, calculated their expression levels, and identified the most expressed ones. Various genes involved in water conservation in mouse models under laboratory conditions were reconstructed and showed high expression levels in *A. olivacea*, demonstrating the value of RNA-seq technology. Given that this work is the first to characterize the transcriptome of any member of Sigmodontinae, a subfamily of cricetid rodents with about 400 species, it will provide valuable resources for future ecological genomics and evolutionary analyses and will serve as assembly reference for a large number of species. In particular, it will facilitate the study of variation in levels of gene expression in the olive mouse and other sigmodontines that occupy a wide range of environmental conditions—from aridlands to temperate rainforests—in South America.

Methods

Data collection

Individuals were collected with Sherman traps from the following localities: Fundo San Martín, Los Ríos (n =4) and Sector Barrancoso, Aysén (n = 4) in Chile, and Gan Gan, Chubut (n =2) and Río Oro, Santa Cruz (n =3) in Argentina (further details in Additional file 1: Table S1). Kidneys were frozen in liquid nitrogen in the field immediately following euthanization. All steps involving live animals followed the recommendations of Sikes *et al.* [44].

RNA extraction and library construction

For each individual, RNA extraction was conducted in one half of the kidney after a lengthwise cut. To this end, the RNeasy mini kit (Qiagen) was employed following recommendations of the manufacturer. RNA quantity and purity was assessed with NanoDrop 1000 Technologies spectrophotometer. RNA integrity was checked through electrophoresis in Formaldehyde-agarose 1,2% denaturing gels. Libraries were constructed and sequenced at Macrogen (Korea). Poly-A based mRNA enrichment method and paired-ends library preparation were done following the Illumina TruSeqTM RNA sample preparation kit, according to the instructions of the manufacturer. Library sequencing was performed on Illumina HiSeq 2000 platform.

De novo transcriptome assembly

Assembly was carried out using default Trinity settings, after removing low quality reads, filtering adaptors and primers, and trimming the 3' ends of reads with a quality less than 24 (Q < 24) with FASTX-Toolkit (http://hannonlab.cshl.edu/fastx_toolkit/). Quality control was checked by FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). All assemblies were done on the same single node-machine with 256G memory and 4 Intel Xeon CPU E7-8837 (8 core) processors. In order to obtain a more complete set of the genes expressed, we pooled individuals from different points of the species distribution and analyzed three strategies: i) merge the reads of the 13 libraries (Multireads) ii) Trinity *in silico* read normalization (TrinityNorm v2013-08-15), and iii) digital normalization (DigiNorm) with khmer (0.8.2). TrinityNorm and DigiNorm were ran on the same computer (Intel Core i7-3820 processor). Normalization algorithms were designed to systematize the coverage in shotgun sequencing data sets, thereby removing redundant reads. As a consequence, computational requirements are reduced, supposedly without negatively impacting assembly quality. For TrinityNorm the default commands were run with a max coverage (max_cov) of 30. Before running DigiNorm, reads were shuffled and a kmer length of 25 and coverage of 30 were specified. We trimmed off likely erroneous k-mers with the filter-abund.py script. For each assembly we tracked the runtimes.

Gene annotation and GO-terms assignment

BLASTX (e-value cut offs < 1e-10) searches were performed against OMA browser mouse proteins and NCBI non-redundant vertebrate protein databases. Search against OMA browser database is a cost-effective way of gene annotation and allows an upper bound estimation of genes orthologous to those of the mouse among the reconstructed contigs. This database contains most or all exons of a given gene, keeping the number of sequences as low as possible. The longest variant is always retained; shorter variants are also kept if they differ by at least in 10% of their sequence from the longer variants retained.

For the BLASTX analysis we report, i) the number of contigs that overlap the proteins of mouse genes at > 99%, > 90%, > 80% and > 50% of their length; ii) the number of distinct mouse proteins reconstructed by a putatively homologous contig using those cutpoints; and iii) the number of distinct genes found with at least one putatively homologous contig reconstructing the gene CDS at or above those percentages. We also reported, the number of contigs, coding sequences, and genes that had a significant hit (e-value < 1e-10) independently of the alignment proportion. To report the number of genes, contigs annotated as putatively homologs to

mouse OMA proteins entries were grouped into the corresponding mouse genes using the oma-ensembl file at OMA-Browser webpage through in-house-scripts (Additional file 4).

The contigs assembled from the Multireads, TrinityNorm and DigiNorm strategies, were annotated for protein domains through the Pfam database using HMMER. An e-value threshold of 1e-2 was set. Before running this analysis we first predicted the exon/intron structure of each contig using the software Augustus [45] trained with *Homo sapiens*. This software has been used extensively for gene prediction (e.g. [46-48]). The GTF files from Augustus output are available upon request. Only the non-overlapping protein domains found were reported on the summary file of BLASTX and Pfam results. A in-house-script was used for this purpose (Additional file 4).

The average number of potential isoforms reconstructed from the Trinity assembly was calculated averaging the times that a particular "comp_XXX" (as given by the ID of Trinity assembled contig) is repeated.

Gene Ontology analysis was done using the DAVID bioinformatics database, using the Benjamini correction of $p < 0.05$ as criterion for enrichment. First, we classify the most common GO-terms from the genes list obtained from the TrinityNorm assembly. Then, we used that gene list as background for analyzing the ontology of the most expressed genes.

List of most expressed genes

To determine the most expressed genes in the kidneys of *A. olivacea*, we sought for genes that were in common among the 5% most expressed genes in each of the 13 transcriptomes. Despite this being a very conservative approach, it was preferred because it would generate a reliable list of genes.

To this end, we first aligned RNA-Seq reads in a paired end fashion against each reference transcript using the aligner Bowtie [49]. Then, we calculated gene-level TPM values using RSEM (v1.2.4). The results of BLASTX for each transcriptome against mouse OMA browser protein, and the OMA-ensembl corresponding pair were used to specify which transcripts were from the same gene. This program hands reads that map to multiple transcripts avoiding throwing away data and biased estimates without relying on the existence of a reference genome. Finally, bash commands and in-house scripts (Additional file 4) were used to obtain the most expressed genes as described above.

Availability of supporting data

The sequencing data has been deposited to the Sequence Read Archive database (accession number SRP033780).

Animal ethics statements

All methods involving *A. olivacea* were carried out in accordance with a protocol reviewed and approved by the Ethics Committee of the Fondo Nacional de Ciencia y Tecnología (FONDECYT, Chile) and the Ethics Committee of the Universidad Austral de Chile (UACh, Chile), as part of the review process for the Fondecyt Research Grant 1110737.

Additional files

Additional file 1: Sampling localities, assembly metrics, contig distribution and gene annotation. Table S1. Specimen ID numbers, sampling localities, and read lengths before and after trimming. **Table S2.** Descriptive statistics of individual RNA-seq samples and reconstructions. **Table S3.** Distribution of contig sizes and their relative proportions for each assembly protocol. **Table S4a.** Gene annotation and CDS reconstruction using BLASTX to OMA browser mouse protein database. **Table S4b.** Annotation using BLASTX to NCBI non-redundant vertebrate protein database. **Table S5.** Annotation through Pfam database using HMMER.

Additional file 2: Summarizing the BLASTX and Pfam results for Multireads, TrinityNorm and DigiNorm strategies.

Additional file 3: List of the 283 most expressed genes and its GO-terms classification. Table S5. List of 283 genes represented in the 5% most expressed genes of each of the individual samples. **Table S6.** Gene Ontology classification of the 283 most expressed genes.

Additional file 4: Containing the in-house-python scripts.

Abbreviations

GO: Gene ontology; CDS: Coding sequences; TrinityNorm: Trinity *in silico* normalization; DigiNorm: Digital normalization.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

MF, GD, LV, DEN, and EPL participated in the field work and sample preparations. VV did the RNA preparations. FG, MF, and LV carried out the analyses. GD, JCO, DEN, and EPL designed and supervised the project. FG wrote the manuscript. All authors approved the final manuscript.

Acknowledgments

We are grateful to Ulyses Pardiñas, Pablo Teta, Daniel Udrizar-Sauthier, Mauro Tammone, and Ivanna Tomasco for assistance in the field. We thank Matthew MacManes and Joseph Cook for comments on an earlier version of the manuscript, and the Bioinformatics Unit, Institut Pasteur de Montevideo, for computational support. Work was supported by grants from CSIC-Universidad de la República, Agencia Nacional de Investigación e Innovación (ANII), Uruguay, and FONDECYT 1110737, Chile. Facundo Giorello and Matías Feijoo are supported by graduate fellowships from ANII.

Author details

[1]Departamento de Ecología y Evolución, Facultad de Ciencias, Universidad de la República, Montevideo, Uruguay. [2]Instituto de Ciencias Ambientales y Evolutivas, Universidad Austral de Chile, Valdivia, Chile.

Received: 13 December 2013 Accepted: 27 May 2014
Published: 8 June 2014

References

1. Waterhouse GR: **Characters of new species of the genus Mus, from collection of Mr. Darwin.** *Proc Zool Soc London* 1837, **5**:15–21. 27–32.
2. D'Elía G, Pardiñas UFJ, Teta P, Patton JL: **Definition and diagnosis of a new tribe of sigmodontine rodents (Cricetidae: Sigmodontinae), and a revised classification of the subfamily.** *Gayana* 2007, **71**:187–194.

3. D'Elía G, Pardiñas UFJ: **Subfamily Sigmodontinae Wagner, 1843.** In *Mamm South Am Vol 2 Rodents*. Edited by Patton JL, Pardiñas UFJ, D'Elía G. Chicago: University of Chicago Press; in press.
4. Mann G: **Los pequeños mamíferos de Chile.** *Gayana Zool* 1978, **40:**1–342.
5. Pizarro R, Valdés R, García-chevesich P, Vallejos C, Sangüesa C, Morales C, Balocchi F, Abarza A, Fuentes R: **Latitudinal analysis of rainfall intensity and mean annual precipitation in chile.** *Chil J Agric Res* 2012, **72:**252–261.
6. Bozinovic F, Rojas JM, Gallardo PA, Palma RE, Gianoli E: **Body mass and water economy in the South American olivaceous field mouse along a latitudinal gradient: Implications for climate change.** *J Arid Environ* 2011, **75:**411–415.
7. Rodríguez-Serrano E, Cancino R a, Palma RE: **Molecular phylogeography of Abrothrix olivaceus (Rodentia: Sigmodontinae) in Chile.** *J Mammal* 2006, **87:**971–980.
8. Lessa EP, D'Elía G, Pardiñas UFJ: **Genetic footprints of late Quaternary climate change in the diversity of Patagonian-Fueguian rodents.** *Mol Ecol* 2010, **19:**3031–3037.
9. Metzker ML: **Sequencing technologies - the next generation.** *Nat Rev Genet* 2010, **11:**31–46.
10. Rizzo JM, Buck MJ: **Key principles and clinical applications of "next-generation" DNA sequencing.** *Cancer Prev Res* 2012, **5:**887–900.
11. Morozova O, Marra M a: **Applications of next-generation sequencing technologies in functional genomics.** *Genomics* 2008, **92:**255–264.
12. Ekblom R, Galindo J: **Applications of next generation sequencing in molecular ecology of non-model organisms.** *Heredity (Edinb)* 2011, **107:**1–15.
13. Shokralla S, Spall JL, Gibson JF, Hajibabaei M: **Next-generation sequencing technologies for environmental DNA research.** *Mol Ecol* 2012, **21:**1794–1805.
14. Garg R, Patel RK, Tyagi AK, Jain M: **De novo assembly of chickpea transcriptome using short reads for gene discovery and marker identification.** *DNA Res* 2011, **18:**53–63.
15. Shi C-Y, Yang H, Wei C-L, Yu O, Zhang Z-Z, Jiang C-J, Sun J, Li Y-Y, Chen Q, Xia T, Wan X-C: **Deep sequencing of the Camellia sinensis transcriptome revealed candidate genes for major metabolic pathways of tea-specific compounds.** *BMC Genomics* 2011, **12:**131.
16. Lorizzo M, Senalik DA, Grzebelus D, Bowman M, Cavagnaro PF, Matvienko M, Ashrafi H, Van Deynze A, Simon PW: **De novo assembly and characterization of the carrot transcriptome reveals novel genes, new markers, and genetic diversity.** *BMC Genomics* 2011, **12:**389.
17. Feldmeyer B, Wheat CW, Krezdorn N, Rotter B, Pfenninger M: **Short read Illumina data for the de novo assembly of a non-model snail species transcriptome (Radix balthica, Basommatophora, Pulmonata), and a comparison of assembler performance.** *BMC Genomics* 2011, **12:**317.
18. Wang Z, Gerstein M, Snyder M: **RNA-Seq: a revolutionary tool for transcriptomics.** *Nat Rev Genet* 2009, **10:**57–63.
19. Brown CT, Howe A, Zhang Q, Pyrkosz AB, Brom TH: **A reference-free algorithm for computational normalization of shotgun sequencing data.** 2012, arXiv:1203.4802v2 [q-bio.GN].
20. Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, Couger MB, Eccles D, Li B, Lieber M, Macmanes MD, Ott M, Orvis J, Pochet N, Strozzi F, Weeks N, Westerman R, William T, Dewey CN, Henschel R, Leduc RD, Friedman N, Regev A: **De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis.** *Nat Protoc* 2013, **8:**1494–1512.
21. Mbassa GK: **Mammalian renal modifications in dry environments.** *Vet Res Commun* 1988, **12:**1–18.
22. Shultz PJ, Tolins JP: **Adaptation to increased dietary salt intake in the rat. Role of endogenous nitric oxide.** *J Clin Invest* 1993, **91:**642–650.
23. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, Chen Z, Mauceli E, Hacohen N, Gnirke A, Rhind N, Di Palma F, Birren BW, Nusbaum C, Lindblad-Toh K, Friedman N, Regev A: **Full-length transcriptome assembly from RNA-Seq data without a reference genome.** *Nat Biotechnol* 2011, **29:**644–652.
24. Altenhoff AM, Schneider A, Gonnet GH, Dessimoz C: **OMA 2011: orthology inference among 1000 complete genomes.** *Nucleic Acids Res* 2011, **39:**D289–D294.
25. Finn RD, Mistry J, Tate J, Coggill P, Heger A, Pollington JE, Gavin OL, Gunasekaran P, Ceric G, Forslund K, Holm L, Sonnhammer ELL, Eddy SR, Bateman A: **The Pfam protein families database.** *Nucleic Acids Res* 2010, **38:**D211–D222.
26. Eddy SR: **Profile hidden Markov models.** *Bioinformatics* 1998, **14:**755–763.
27. Eddy SR, Mitchison G, Durbin R: **Maximum discrimination hidden Markov models of sequence consensus.** *J Comput Biol* 1995, **2:**9–23.
28. Dennis G, Sherman BT, Hosack DA, Yang J, Gao W, Lane HC, Lempicki RA: **DAVID: Database for Annotation, Visualization, and Integrated Discovery.** *Genome Biol* 2003, **4:**P3.
29. Li B, Dewey CN: **RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome.** *BMC Bioinforma* 2011, **12:**323.
30. Pradervand S, Zuber Mercier A, Centeno G, Bonny O, Firsov D: **A comprehensive analysis of gene expression profiles in distal parts of the mouse renal tubule.** *Pflügers Arch Eur J Physiol* 2010, **460:**925–952.
31. Hershkovitz V, Sela N, Taha-Salaime L, Liu J, Rafael G, Kessler C, Aly R, Levy M, Wisniewski M, Droby S: **De-novo assembly and characterization of the transcriptome of Metschnikowia fructicola reveals differences in gene expression following interaction with Penicillium digitatum and grapefruit peel.** *BMC Genomics* 2013, **14:**168.
32. Liu T, Zhu S, Tang Q, Chen P, Yu Y, Tang S: **De novo assembly and characterization of transcriptome using Illumina paired-end sequencing and identification of CesA gene in ramie (Boehmeria nivea L. Gaud).** *BMC Genomics* 2013, **14:**125.
33. Yano N, Endoh M, Fadden K, Yamashita H, Kane A, Sakai H, Rifai A: **Comprehensive gene expression profile of the adult human renal cortex: analysis by cDNA array hybridization.** *Kidney Int* 2000, **57:**1452–1459.
34. Spradling KD, Glenn JP, Garcia R, Shade RE, Cox LA: **The baboon kidney transcriptome: analysis of transcript sequence, splice variants, and abundance.** *PLoS One* 2013, **8:**e57563.
35. Bullard JH, Purdom E, Hansen KD, Dudoit S: **Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments.** *BMC Bioinforma* 2010, **11:**94.
36. Fenton RA, Knepper MA: **Mouse models and the urinary concentrating mechanism in the new millennium.** *Physiol Rev* 2007, **87:**1083–1112.
37. Meij IC, Koenderink JB, Van Bokhoven H, Assink KF, Groenestege WT, De Pont JJ, Bindels RJ, Monnens LA, Van Den Heuvel LP, Knoers NV: **Dominant isolated renal magnesium loss is caused by misrouting of the Na+, K + −ATPase gamma-subunit.** *Ann N Y Acad Sci* 2003, **986:**265–266.
38. Bachmann S, Dawnay AB, Bouby N, Bankir L: **Tamm-Horsfall protein excretetion during chronic alterations in urinary concentration and protein intake in rat.** *Ren Physiol Biochem* 1991, **14:**236–245.
39. Cramer SD, Ferree PM, Lin K, Milliner DS, Holmes RP: **The gene encoding hydroxypyruvate reductase (GRHPR) is mutated in patients with primary hyperoxaluria type II.** *Hum Mol Genet* 1999, **8:**2063–2069.
40. Lam CW, Yuen YP, Lai CK, Tong SF, Lau LK, Tong KL, Chan YW: **Novel mutation in the GRHPR gene in a Chinese patient with primary hyperoxaluria type 2 requiring renal transplantation from a living related donor.** *Am J Kidney Dis* 2001, **38:**1307–1310.
41. Steffgen J, Kampfer K, Grupp C, Langenberg C, Müller GA, Grunewald RW: **Osmoregulation of aldose reductase and sorbitol dehydrogenase in cultivated interstitial cells of rat renal inner medulla.** *Nephrol Dial Transplant* 2003, **18:**2255–2261.
42. Al Samri MT, Al Shamsi M, Al-Salam S, Marzouqi F, Al Mansouri A, Al-Hammadi S, Balhaj G, Al Dawaar SKM, Al Hanjeri RSMS, Benedict S, Sudhadevi M, Conca W, Penefsky HS, Souid A-K: **Measurement of oxygen consumption by murine tissues in vitro.** *J Pharmacol Toxicol Methods* 2011, **63:**196–204.
43. Tahara EB, Navarete FDT, Kowaltowski AJ: **Tissue-, substrate-, and site-specific characteristics of mitochondrial reactive oxygen species generation.** *Free Radic Biol Med* 2009, **46:**1283–1297.
44. Sikes RS, Gannon WL: **Guidelines of the American Society of Mammalogists for the use of wild mammals in research.** *J Mammal* 2011, **92:**235–253.
45. Stanke M, Keller O, Gunduz I, Hayes A, Waack S, Morgenstern B: **AUGUSTUS: ab initio prediction of alternative transcripts.** *Nucleic Acids Res* 2006, **34:**W435–W439.
46. Cornman R, Bennett AK, Murray K, Evans JD, Elsik CG, Aronstein K: **Transcriptome analysis of the honey bee fungal pathogen. Ascosphaera apis: implications for host pathogenesis.** *BMC Genomics* 2012, **13:**285.
47. Wang S, Furmanek T, Kryvi H, Krossøy C, Totland GK, Grotmol S, Wargelius A: **Transcriptome sequencing of Atlantic salmon (Salmo salar L.) notochord prior to development of the vertebrae provides clues to regulation of positional fate, chordoblast lineage and mineralisation.** *BMC Genomics* 2014, **15:**141.

48. Eckalbar WL, Hutchins ED, Markov GJ, Allen AN, Corneveaux JJ, Lindblad-Toh K, Di Palma F, Alföldi J, Huentelman MJ, Kusumi K: **Genome reannotation of the lizard Anolis carolinensis based on 14 adult and embryonic deep transcriptomes.** *BMC Genomics* 2013, **14:**49.
49. Langmead B, Trapnell C, Pop M, Salzberg SL: **Ultrafast and memory-efficient alignment of short DNA sequences to the human genome.** *Genome Biol* 2009, **10:**R25.

doi:10.1186/1471-2164-15-446
Cite this article as: Giorello *et al.*: Characterization of the kidney transcriptome of the South American olive mouse *Abrothrix olivacea*. *BMC Genomics* 2014 **15**:446.

PERSPECTIVAS

Para el caso particular de las estrategias TrinityNorm, DigiNorm y Multireads, sería conveniente intentar reducir el número de cóntigos que alinean con alguna secuencia codificante. Aunque es esperable que muchos de ellos representen efectivamente isoformas, este número puede estar sobrestimado ya que al juntar lecturas de distintas muestras es esperable que una misma isoforma se ensamble más de una vez en cóntigos distintos a causa de la variación existente entre los individuos. Para mitigar este posible efecto, se podrían colapsar los cóntigos presuntamente redundantes utilizando programas como el CD-HIT (Fu et al. 2012). De esta manera obtendríamos un transcriptoma de referencia más conciso y depurado para análisis posteriores.

Aunque nuestra aproximación para determinar los genes más expresados entre los 13 individuos establece un grupo de genes confiables para trabajos posteriores, explorar otros métodos menos conservadores podría ser útil para encontrar más genes vinculados a la conservación del agua. A su vez, sería conveniente para futuros análisis remover o enmascarar los genes mitocondriales de la lista de los genes más expresados para encontrar nuevos términos ontológicos que nos ayuden a delimitar grupos de genes potencialmente vinculados a las principales funciones renales. Naturalmente, para aquellos genes de particular interés, como la Aqp 1, su validación utilizando técnicas experimentales como el PCR en tiempo real sería beneficiosa y necesaria para determinar su posible rol en la regulación hídrica.

Finalmente, para mejorar la anotación génica sería conveniente exigir que los cóntigos presenten los mismos dominios proteicos que las secuencias de referencia y un porcentaje de identidad de secuencia mínimo entre ellos.

BIBLIOGRAFÍA

Abud, C., 2011. Variación genética y estructura filogeográfica de Abrothrix olivaceus en la Patagonia argentina y el sur chile. Tesis de maestría. Universidad de la República. Uruguay.

Ashburner, M. et al., 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature genetics*, 25, pp.25–29.

Bashir, A. et al., 2012. A hybrid approach for the automated finishing of bacterial genomes. *Nature Biotechnology*, 30, pp.701–707.

Bozinovic, F. et al., 2011. Body mass and water economy in the South American olivaceous field mouse along a latitudinal gradient: Implications for climate change. *Journal of Arid Environments*, 75, pp.411–415.

Brown, C.T. et al., 2012. A Reference-Free Algorithm for Computational Normalization of Shotgun Sequencing Data. , pp.1–18.

Chin, C.-S. et al., 2013. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nature methods*, 10, pp.563–569.

Compeau, P.E.C., Pevzner, P.A. & Tesler, G., 2011. How to apply de Bruijn graphs to genome assembly. *Nature Biotechnology*, 29, pp.987–991.

Dennis, G. et al., 2003. DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biology*, 4, p.P3.

Ekblom, R. & Galindo, J., 2011. Applications of next generation sequencing in molecular ecology of non-model organisms. *Heredity*, 107, pp.1–15.

Ellegren, H., 2008. Comparative genomics and the study of evolution by natural selection. *Molecular ecology*, 17, pp.4586–4596.

Excoffier, L. et al., 2013. Robust Demographic Inference from Genomic and SNP Data. *PLoS Genetics*, 9.

Ferrarini, M. et al., 2013. An evaluation of the PacBio RS platform for sequencing and de novo assembly of a chloroplast genome. *BMC genomics*, 14, p.670.

Flicek, P. et al., 2014. Ensembl 2014. *Nucleic acids research*, 42, pp.D749–D755.

Fu, L. et al., 2012. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics (Oxford, England)*, 28, pp.3150–3152.

Gagnaire, P.-A., Normandeau, E. & Bernatchez, L., 2012. Comparative genomics reveals adaptive protein evolution and a possible cytonuclear incompatibility between European and American Eels. *Molecular biology and evolution*, 29, pp.2909–2919.

Gayral, P. et al., 2013. Reference-Free Population Genomics from Next-Generation Transcriptome Data and the Vertebrate-Invertebrate Gap. *PLoS Genetics*, 9.

Grabherr, M.G. et al., 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology*, 29, pp.644–652.

Gutenkunst, R.N. et al., 2009. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genetics*, 5.

Haas, B.J. et al., 2013. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nature Protocols*, 8, pp.1494–512.

Koren, S. et al., 2012. Hybrid error correction and de novo assembly of single-molecule sequencing reads. *Nature Biotechnology*, 30, pp.693–700.

Lessa, E.P., D'Elía, G. & Pardiñas, U.F.J., 2010. Genetic footprints of late Quaternary climate change in the diversity of Patagonian-Fueguian rodents. *Molecular Ecology*, 19, pp.3031–3037.

Li, B. et al., 2010. RNA-Seq gene expression estimation with read mapping uncertainty. *Bioinformatics (Oxford, England)*, 26, pp.493–500.

Li, B. & Dewey, C.N., 2011. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*, 12, p.323.

Liu, L. et al., 2012. Comparison of next-generation sequencing systems. *Journal of Biomedicine and Biotechnology*, 2012,pp.1-11.

Mann, G., 1978. Los pequeños mamíferos de Chile. *Gayana, Zoologia*, 40, pp.1–342.

McCoy, R.C. et al., 2014. Genomic inference accurately predicts the timing and severity of a recent bottleneck in a nonmodel insect population. *Molecular ecology*, 23, pp.136–150.

Metzker, M.L., 2010. Sequencing technologies - the next generation. *Nature Reviews Genetics*, 11, pp.31–46.

Miller, J.R., Koren, S. & Sutton, G., 2010. Assembly algorithms for next-generation sequencing data. *Genomics*, 95, pp.315–327.

Mortazavi, A. et al., 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. H. Kitagawa et al., eds. *Nature Methods*, 5, pp.621–628.

Nosil, P., 2012. *Ecological Speciation* P. Harvey et al., eds., London: Oxford University Press.

Oshlack, A., Robinson, M.D. & Young, M.D., 2010. From RNA-seq reads to differential expression results. *Genome biology*, 11, p.220.

Pavey, S.A. et al., 2010. The role of gene expression in ecological speciation. *Annals of the New York Academy of Sciences*, 1206, pp.110–129.

Perry, G.H. et al., 2012. Comparative RNA sequencing reveals substantial genetic variation in endangered primates. *Genome Research*, 22, pp.602–610.

Pfennig, D.W. et al., 2010. Phenotypic plasticity's impacts on diversification and speciation. *Trends in Ecology and Evolution*, 25, pp.459–467.

Renaut, S., Nolte, A.W. & Bernatchez, L., 2010. Mining transcriptome sequences towards identifying adaptive single nucleotide polymorphisms in lake whitefish species pairs (Coregonus spp. Salmonidae). *Molecular ecology*, 19, pp.115–131.

Ribeiro, F.J. et al., 2012. Finished bacterial genomes from shotgun sequence data. *Genome Research*, 22, pp.2270–2277.

Robertson, G. et al., 2010. De novo assembly and analysis of RNA-seq data. *Nature methods*, 7, pp.909–912.

Rodríguez-Serrano, E., Cancino, R. a. & Palma, R.E., 2006. Molecular phylogeography of Abrothrix olivaceus (Rodentia: Sigmodontinae) in Chile. *Journal of Mammalogy*, 87, pp.971–980.

Sanger, F., Nicklen, S. & Coulson, A.R., 1977. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America*, 74, pp.5463–5467.

Schadt, E.E., Turner, S. & Kasarskis, A., 2010. A window into third-generation sequencing. *Human molecular genetics*, 19, pp.R227–R240.

Schulz, M.H. et al., 2012. Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels. *Bioinformatics (Oxford, England)*, 28, pp.1086–1092.

Smith, M.F., Kelt, D.A. & Patton, J.L., 2001. Testing models of diversification in mice in the Abrothrix olivaceus/xanthorhinus complex in Chile and Argentina. *Molecular Ecology*, 10, pp.397–405.

Thibert-Plante, X. & Hendry, A.P., 2011. The consequences of phenotypic plasticity for ecological speciation. *Journal of evolutionary biology*, 24, pp.326–342.

Travers, K.J. et al., 2010. A flexible and efficient template format for circular consensus sequencing and SNP detection. *Nucleic acids research*, 38, p.e159-167.

Wang, Z., Gerstein, M. & Snyder, M., 2009. RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, 10, pp.57–63.

Waterhouse, G.R., 1837. Characters of new species of the genus Mus, from collection of Mr. Darwin. *Proceedings of the Zoological Society of London*, 5, pp.15–21,27–32.

Wolf, J.B.W. et al., 2010. Nucleotide divergence vs. gene expression differentiation: comparative transcriptome sequencing in natural isolates from the carrion crow and its hybrid zone with the hooded crow. *Molecular ecology*, 19, pp.162–175.

Zhao, Q.-Y. et al., 2011. Optimizing de novo transcriptome assembly from short-read RNA-Seq data: a comparative study. *BMC Bioinformatics*, 12, p.S2.